REVUE DE STATISTIQUE APPLIQUÉE

PATRICIA POTTIER

Utilisations de l'analyse en composantes principales pour la prévision statistique en météorologie

Revue de statistique appliquée, tome 39, n° 1 (1991), p. 37-49

<http://www.numdam.org/item?id=RSA_1991__39_1_37_0>

Rev. Statistique Appliquée, 1991, **XXXIX** (1), 37-49

UTILISATIONS DE L'ANALYSE EN COMPOSANTES PRINCIPALES POUR LA PRÉVISION STATISTIQUE EN MÉTÉOROLOGIE

Patricia POTTIER*

Septembre 1990

Ce travail a été réalisé sous la direction de MM. Guy Der Mégréditchian et Jean-Pierre Javelle (Météorologie Nationale/CRMD/STAT-MATH) .

1. Introduction

L'outil de base de la prévision météorologique est un modèle physique déterministe de l'atmosphère. Mais les outils statistiques interviennent en amont et en aval. Tout d'abord l'analyse objective permet d'utiliser au mieux les différentes données mesurées des paramètres météorologiques pour initialiser les champs utilisés par le modèle. Les outils statistiques sont ensuite employés lors de l'investigation de la qualité d'un modèle dynamique. Cette investigation peut être intrinsèque à travers le calcul d'indices globaux de qualité permettant de mesurer l'adéquation entre les champs prévus et analysés. D'autre part, sur le plan utilitaire, on cherche à adapter la prévision physique pour parvenir à une prévision locale des variables et phénomènes météorologiques.

La démarche présentée ici se situe dans cette dernière approche, dans le cadre de la prévision statistique : extraire - à l'aide d'outils statistiques - des champs de paramètres physiques issus du modèle déterministe des prévisions de paramètres ou de phénomènes locaux.

Plus précisément, on s'intéressera ici à la prévision, en une centaine de villes françaises, du cycle diurne de température (températures toutes les trois heures et températures minimales et maximales observées au cours d'une journée).

Lors de cette étude, a été mise en évidence la pertinence de l'outil Analyse en Composantes Principales (ACP) pour réaliser la prévision de températures demandée.

* Météorologie Nationale, CRMD, STAT-MATH, 2, avenue Rapp, 75340 Paris Cedex 07.

2. L'Analyse en Composantes Principales

2.1. Les données de l'étude

Au cours des dernières années, la modèlisation de l'atmosphère a progressé grâce notamment à une diminuation de la maille et donc de l'échelle des phénomènes traités (maille de 38 kilomètres pour le modèle PERIDOT [réf.1]).

Malgré cela, les valeurs de température prévues par PERIDOT aux points de grille sont parfois peu représentatives des températures observées aux villes voisines de ces points , surtout dans les zones à relief prononcé. Une adaptation (statistique en ce qui nous concerne) est nécessaire.

La base de données disponible a permis de constituer un fichier d'apprentissage d'une année et un fichier test indépendant de neuf mois. Pour chaque jour de ce fichier, 29 champs météorologiques (dont le champ de température) analysés ou prévus à différentes échéances par le modèle physique PERIDOT de la Météorologie Nationale étaient connus sur la France par leurs valeurs sur une grille de 990 points [fig. 1].

On présentera ici essentiellement les résultats concernant le champ de température à 2 mètres analysé ou prévu à échéance de 24 heures.

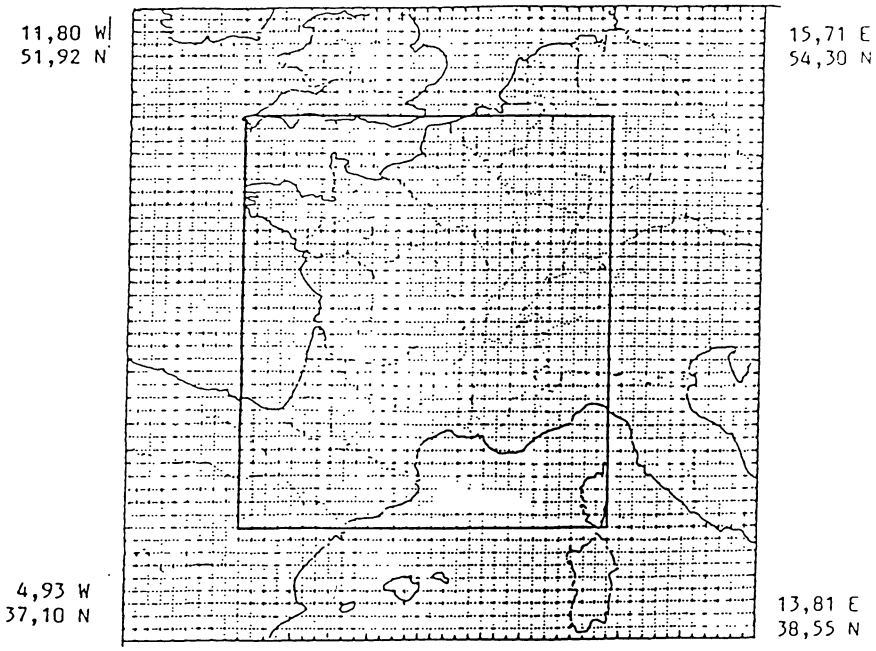

FIGURE 1
La sous-grille PERIDOT utilisée pour notre étude.

2.2. Mise en oeuvre de l'Analyse en Composantes Principales

Cette Analyse en Composantes Principales [réf.2] a été réalisée par diagonalisation de la matrice de variance-covariance V_{XX} avec les procédures de la bibliothèque STATMET développée à la Météorologie Nationale. Précisons les notations employées.

On note $X(t,x)$ la valeur du champ de température prévu par PERIDOT le jour t au lieu (point de grille) x.

Lors de la décomposition on écrit $X(t,x)$:

$$X(t,x) = \sum_j a_j(t) C_j(x),$$

où $C_j(x)$ est le $j^{ème}$ Vecteur Propre, et $a_j(t)$ la $j^{ème}$ Composante Principale (CP) au temps t. On notera aussi λ_j la $j^{ème}$ valeur propre et P_j le pourcentage de variance expliquée par le $j^{ème}$ vecteur propre défini par :

$$P_j = \frac{\lambda_j}{\sum_i \lambda_i}.$$

Remarquons que lors de cette ACP, le vecteur X avait une taille de 990 alors qu'on ne disposait que de 300 de ses observations. Au lieu de diagonaliser la matrice V_{XX} (de taille 990 × 990), on aurait aussi pu réaliser l'ACP duale. Le pourcentage très élevé (90%) de variance expliquée par la première CP s'explique aisément par la forte variabilité saisonnière du paramètre température. Les 15 premières composantes principales expliquent 98% de la variance totale du champ mais ne permettent cependant pas une reconstitution parfaite du champ puisque l'erreur quadratique moyenne entre les champs initiaux et les champs reconstitués (avec 15 CP) reste de l'ordre de 1 degré Celsius.

3. Mise en évidence de la structure spatio-temporelle du champ de température

L'ACP permet de décrire la structure spatio-temporelle du champ de température en mettant en évidence séparément les caractéristiques spatiales et temporelles du champ à travers, d'une part la pondération des vecteurs propres et, d'autre part, l'évolution temporelle des composantes principales.

3.1. Structure spatiale

On s'est intéressé à la corrélation spatiale $r_{Xa}(j,x)$ entre les séries temporelles suivantes :

- série des températures au point x : $X(t,x)$ et

- série des $j^{ème}$ composantes principales : $a_j(t)$.

De plus, $r^2_{Xa}(j,x)$ s'interprète comme le pourcentage de variance du champ X au point x associé au $j^{ème}$ vecteur propre et on a la relation suivante [réf.3] :

$$r^2_{Xa}(j,x) = \frac{\lambda_j C_j^2(x)}{\sum_i \lambda_i C_i^2(x)}$$

Sur les cartes de $r_{Xa}(j,x)$ présentées ici pour quelques vecteurs propres, on retrouve des résultats classiquement obtenus lors de l'ACP des champs météorologiques, perturbés cependant par les effets du relief et de l'inhomogénéité entre la température sur terre et sur mer.

Le premier vecteur propre [fig. 2] apparaît comme une simple pondération continue du champ avec un maximum peu marqué au centre du domaine. Pour le deuxième vecteur propre [fig. 3], la structure simple (une opposition nord-sud) est perturbée par l'effet terre/mer. La même constatation se dégage de la carte concernant le troisième vecteur propre [fig. 4]. Ensuite la structure devient de plus en plus complexe et on note notamment un resserrement des isolignes sur les montagnes [fig. 5].

3.2. Structure temporelle

Etudions maintenant la corrélation temporelle $r_{XC}(j,t)$ entre les séries spatiales suivantes (il ne s'agit pas vraiment d'une corrélation mais plutôt d'un cosinus, les données étant centrées par rapport aux stations et non par rapport au temps).

- carte des températures au jour t : $X(t,x)$ et
- pondérations du $j^{ème}$ vecteur propre : $C_j(x)$, pour x variant de 1 à 990.

De plus, $r^2_{XC}(j,t)$ s'interprète comme le pourcentage de variance du champ X le jour t associé au $j^{ème}$ vecteur propre et on a la relation suivante [réf.3] :

$$r^2_{XC}(j,t) = \frac{a_j^2(t)}{\sum_i a_i^2(t)}$$

Sur les courbes de $r_{XC}(j,t)$ présentées, on décrit l'évolution temporelle des composantes principales normalisées. Pour la première composante [fig. 6], le cycle saisonnier annuel est nettement mis en évidence. Pour les composantes suivantes, l'élément cyclique contenu dans le champ météorologique initial disparaît tandis que l'amplitude des oscillations autour de l'horizontale diminue [fig. 7].

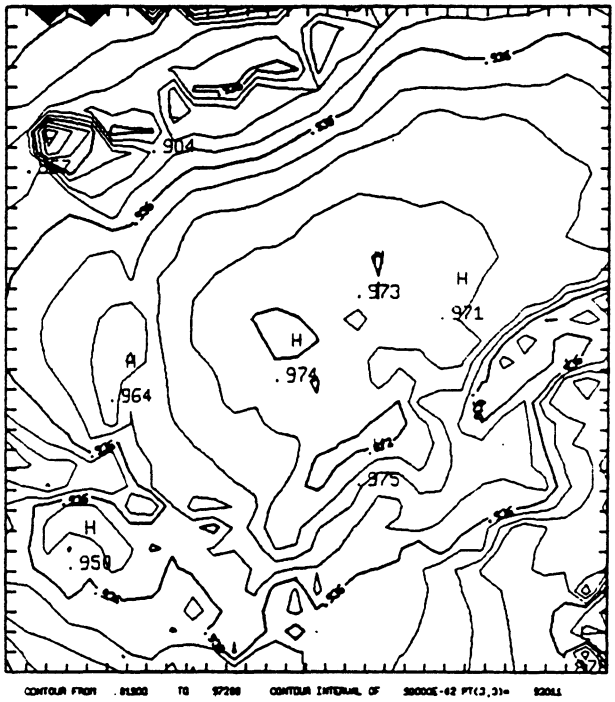

FIGURE 2
Carte de $r_{Xa}(1, x)$.

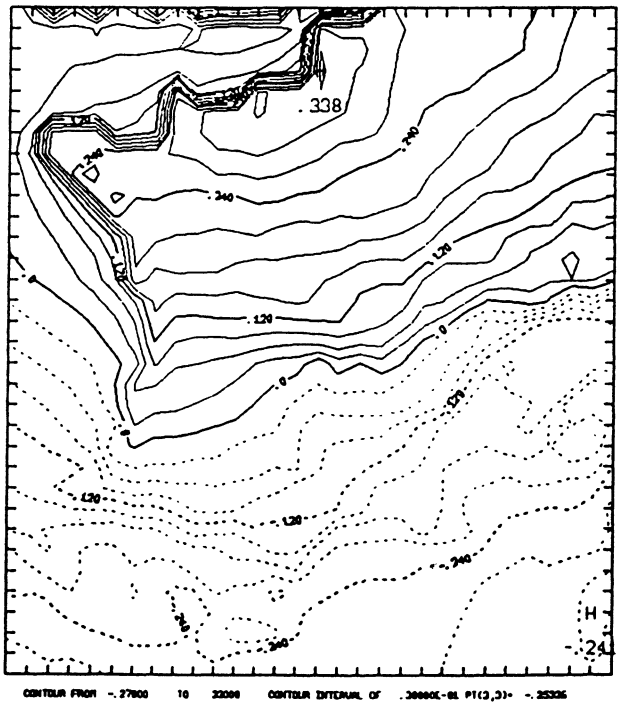

FIGURE 3
Carte de $r_{Xa}(2, x)$.

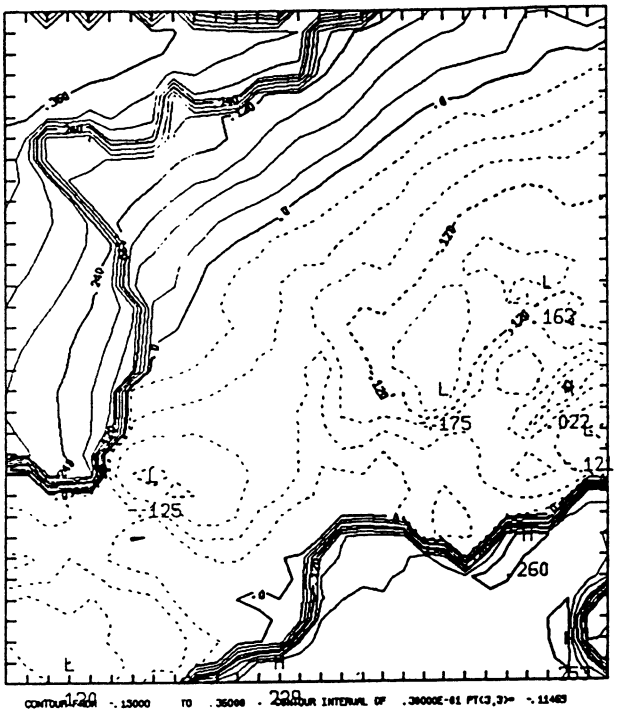

FIGURE 4
Carte de $r_{Xa}(3, x)$.

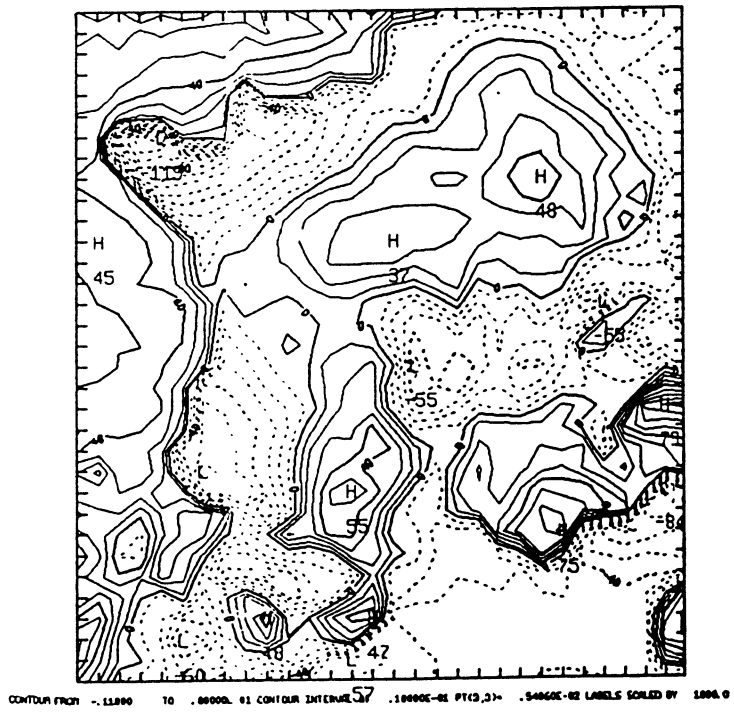

FIGURE 5
Carte de $r_{Xa}(15, x)$.

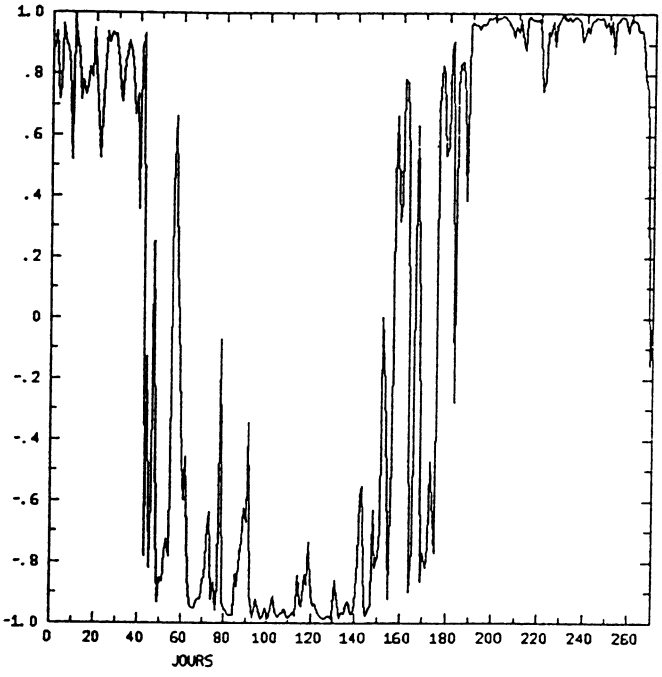

FIGURE 6
Courbe $r_{XC}(1,t)$.

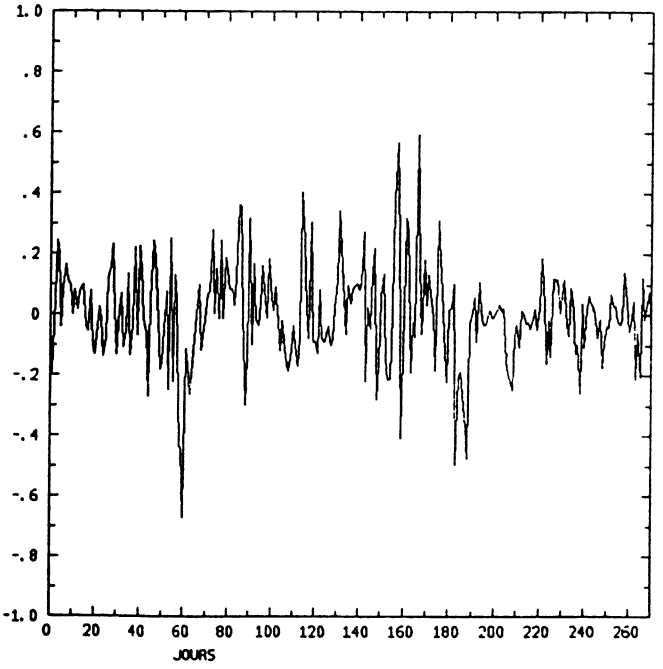

FIGURE 7
Courbe $r_{XC}(4,t)$.

4. Représentativité des Composantes Principales

4.1. Représentativité globale

L'autocorrélation spatiale du champ de température décrit par 990 points de grille est importante. Aussi, peut-on essayer de le reconstituer en ne prenant en compte que les premières Composantes Principales. On obtient ainsi un champ filtré :

$$X_{filtré}(t,x) = \sum_{j}^{n} a_j(t) C_j(x),$$

avec $n \ll N = 990$. Quand le nombre n de Composantes Principales utilisées (prises dans l'ordre des valeurs propres décroissantes) augmente, le champ reconstitué se "rapproche" du champ initial.

Cette utilisation de l'Analyse en Composantes Principales permet de décrire les grandes structures du champ avec un nombre réduit de variables. On peut aussi l'envisager dans le but de lisser un champ trop bruité, ce qui fut notre approche, le champ de température prévu par le modèle PERIDOT en chacun des points de grille étant trop bruité. Cependant, ce champ, même lissé, ne permet pas de bien prévoir la température localement : c'est à dire en chaque ville. Pour cela une autre approche des Composantes Principales a été employée.

4.2. Représentativité locale

Pour prévoir la température en une ville donnée, la méthode la plus performante [réf. 4] s'est révélée être le calcul d'une équation de régression linéaire multiple entre le prédictand (la température observée à la ville) et 15 prédicteurs : les 15 premières Composantes Principales du champ de température prévu par le modèle PERIDOT.

Comme lors de la reconstitution du champ, on utilise les 15 premières Composantes Principales, mais le poids relatif (valeur absolue du coefficient de régression normalisé) affecté à chaque Composante Principale n'est plus alors celui des valeurs propres décroissantes.

Par exemple, dans l'équation de prévision de la température à Marseille, les Composantes Principales ayant les plus forts poids sont : la première, puis la seconde, puis la dixième et la septième. On constate parallèlement sur les cartes de corrélations spatiales (au carré, le signe ici importe peu) du dixième ou du septième vecteur propre [fig. 8] de fortes variations de ces corrélations autour de Marseille.

Plus généralement, il est possible de mettre en parallèle

- les cartes de corrélations spatiales des vecteurs propres,
- les cartes donnant le "rang" de chaque Composante Principale dans les régressions permettant de prévoir la température en différentes villes françaises ("rang" lors d'un classement des Composantes Principales selon

les coefficients de régression normalisés décroissants -en valeur absolue- dans l'équation de régression pour la ville considérée).

Pour le septième vecteur propre par exemple, on constate qu'à des noyaux de fortes variations des corrélations spatiales [fig. 8] correspondent des zones où la septième Composante Principale apparaît en troisième, quatrième, cinquième ou sixième rang [fig. 9].

La méthode de prévision de température en une ville donnée consiste donc en un calcul des 15 premières CP du champ de température prévu par le modèle PERIDOT puis en une régression utilisant ces 15 CP.

Ainsi après la régression sur ces Composantes Principales calculées à partir des 990 valeurs du champ sur la grille initiale, on peut revenir à cette grille et on obtient des coefficients de régression sur les variables initiales.

Ces coefficients n'étaient certes pas calculables directement, compte-tenu des très fortes corrélations existant entre ces 990 valeurs. Cependant nous pouvons, a posteriori, nous intéresser à la pondération (coefficients de régression) des 990 points de grille ainsi obtenue, pour chaque ville, à l'issue de l'ACP et de la régression. On constate alors que l'équation de régression calculée pour une ville, donne un poids important à des points se situant dans un voisinage plus ou moins grand de la ville considérée, selon que les valeurs issues du modèle PERIDOT sont peu ou très représentatives de la température de la ville considérée.

Deux figures illustrent ces deux comportements :

- la figure 10 présente la carte des coefficients de régression obtenus pour la ville de Langres (localisée par $*$) ; on ne constate pas de maximum marqué autour de cette ville ;
- la figure 11, pour la ville de Bordeaux, présente au contraire une zone de maximum centré autour de cette ville.

L'étude de l'ordre des CP dans les régressions ainsi que l'étude des pondérations finales, met en évidence deux comportements :

- les régions où la complexité orographique n'est pas bien prise en compte par le relief du modèle physique : l'ordre des CP est assez différent de l'ordre des valeurs propres décroissantes et les cartes de coefficients de régression associées aux villes de ces régions sont très bruitées ;
- les régions où les points de grille sont représentatifs des zones avoisinantes : l'ordre des CP est peu changé et les cartes de coefficients de régression présentent un maximum marqué autour de la ville considérée.

Parallèlement, la méthode de prévision utilisant l'ACP présente un apport plus important par rapport à des méthodes plus simples (interpolation entre les valeurs aux points de grille entourant une ville par exemple) dans les zones où l'ordre des CP est plus différent de l'ordre des valeurs propres décroissantes.

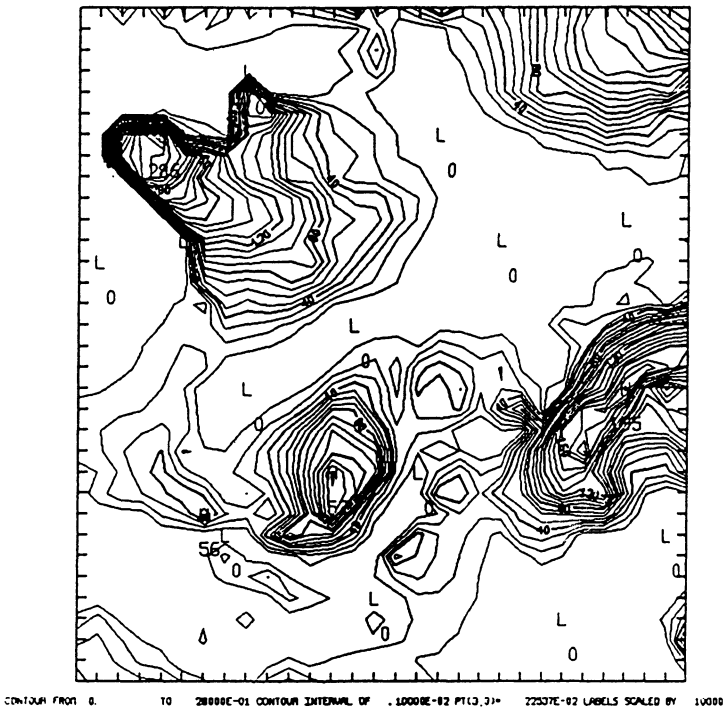

FIGURE 8
Carte de $r^2_{Xa}(7,x)$.

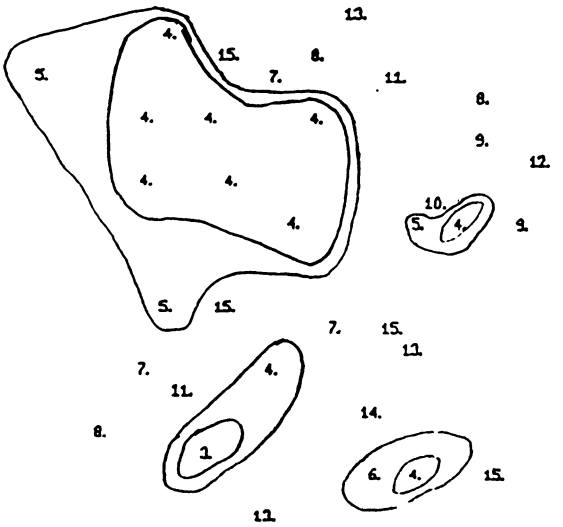

FIGURE 9
Pour chaque ville, on donne le rang de l'importance de la $7^{ème}$ *CP dans l'équation de régression établie pour la ville considérée.*

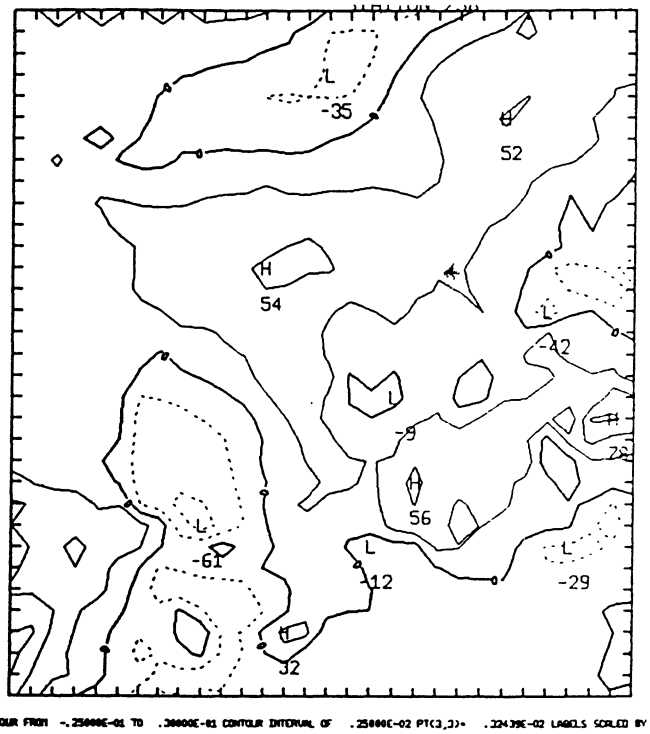

FIGURE 10
Pour la station de Langres, coefficients de régression obtenus après régression sur les 15 CP et retour sur la grille initiale.

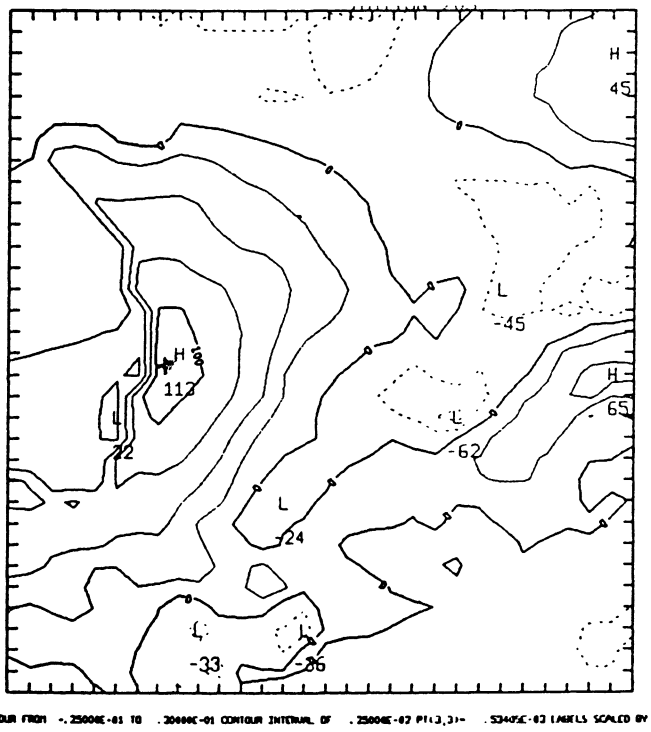

FIGURE 11
Pour la station de Bordeaux, coefficients de régression obtenus après régression sur les 15 CP et retour sur la grille initiale.

5. Conclusions

L'Analyse en Composantes Principales a tout d'abord présenté deux intérêts pour notre étude :

- en premier lieu, un *intérêt descriptif* , par la mise en évidence, de façon distincte, des structures spatiales et temporelles du champ de température prévu par le modèle physique PERIDOT;
- un *intérêt pour la prévision* ensuite, puisque, associée à une régression linéaire multiple, l'ACP permet de corriger localement les prévisions du modèle physique difficilement utilisables directement.

De plus, il est important de noter que l'ACP permet d'extraire du champ de température prévu par le modèle physique, l'information utile à la prévision locale de température :

- les nombreux essais effectués [réf.5, 6 et 7] en utilisant comme prédicteurs de la régression les *valeurs en point de grille* des paramètres météorologiques (et non plus les CP) prévus par le modèle physique, directement ou après différents lissages ou regroupements canoniques, n'ont pas permis de réaliser une *prévision locale* de *qualité équivalente* sans la *prise en compte* d'*autres champs météorologiques* (température en altitude, vent, etc...).

Enfin, on peut signaler qu'une prévision de température basée sur la méthode présentée ici est actuellement opérationnelle à la Météorologie Nationale et fournit quotidiennement, pour environ cent cinquante villes françaises, des prévisions de température toutes les 3 heures pour des échéances de 6 à 60 heures [réf. 8].

6. Bibliographie

[réf.1] : supplément de la revue "La Recherche", Juillet-Août 1988, numéro 201.

[réf.2] : Guy Der Mégréditchian, "Le traitement statistique des données multidimensionnelles", Volumes 1 et 2, Ecole Nationale de la Météorologie, Juin 1988.

[réf.3] : R.W. Preisendorfer, "Principal Component Analysis in Meteorology and Oceanography", Developments in Atmospheric Science, 17, Elseiver 1988.

[réf.4] : Patricia Baptistan-Pottier, "Application de l'Analyse en Composantes Principales pour la prévision du cycle diurne de température par adaptation statistique des sorties du modèle PERIDOT", note de travail du Centre de Recherche en Météorologie Dynamique, n^o 6, mai 1989.

[réf.5] : Patricia Baptistan-Pottier, "Essai d'adaptation statistique du modèle dynamique PERIDOT", mémoire de DEA, Université de Paris VI, juin 1988.

[réf.6] : Patricia Baptistan-Pottier, Guy. Der Mégréditchian, Jean-Pierre Javelle, "Application of statistical methods to operational forecast in France", 11th Conference on Probablity and Statistics, American Meteorological Society, October 1-5, 1989, Monterey.

[réf.7] : Patricia Pottier, "Prévision de température par adaptation statistique du modèle PERIDOT', revue 'La Météorologie", n^o 32, avril 1990.

[réf.8] : Patricia Pottier, "Extension de la prévision de températures réalisée par adaptation statistique du modèle PERIDOT : méthode opérationnelle", note de travail du Centre de Recherche en Météorologie Dynamique, n^o 12, avril 1990.