Prévision Quantitative des Précipitations Journalières par une méthode Statistico-Dynamique de Recherche d'Analogues - Application à des Bassins du Pourtour Méditerranéen

Sophie Guilbaud

▶ **To cite this version:**

Sophie Guilbaud. Prévision Quantitative des Précipitations Journalières par une méthode Statistico-Dynamique de Recherche d'Analogues - Application à des Bassins du Pourtour Méditerranéen. Météorologie. INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE, 1997. Français. NNT : . tel-01090981

HAL Id: tel-01090981

https://theses.hal.science/tel-01090981

Submitted on 4 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE

présentée par

Sophie GUILBAUD
Ingénieur ENSHMG-INPG

Pour obtenir le grade de DOCTEUR de

l'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

(Arrêté ministériel du 30 mars 1992)

Spécialité : Mécanique des Milieux Géophysiques et Environnement

Prévision Quantitative des Précipitations Journalières par une méthode Statistico-Dynamique de Recherche d'Analogues

Application à des Bassins du Pourtour Méditerranéen

Date de soutenance : 30 Octobre 1997

Composition du jury:

Jeanny HERAULT	Président
Philippe BOUGEAULT	Rapporteur
Daniel DUBAND	Rapporteur
Stefano BOVO	Examinateur
Carmen LLASAT	Examinateur
Charles OBLED	Examinateur

Thèse préparée au **Laboratoire d'étude des Transferts en Hydrologie et Environnement**
UMR 5564 (CNRS - INPG - ORSTOM - UJF)

REMERCIEMENTS

***M. J. Hérault**, professeur à l'université J. Fourier à Grenoble, a accepté de présider mon jury de thèse. Qu'il en soit remercié.*

***M. D. Duband**, ancien chef de Service à EDF et initiateur de ces méthodes dans les années 70, m'a encouragée à les reprendre et à les développer avec les moyens d'aujourd'hui. Je l'en remercie vivement.*

*Je souhaite aussi exprimer toute ma gratitude à **M. Ph. Bougeault**, chef d'unité à Météo-France, qui a accepté d'être rapporteur de cette thèse. Ses remarques constructives ont vraiment fait progresser le manuscrit.*

***Mme C. Llasat**, professeur à l'Université de Barcelone, et **Dr. S. Bovo**, de la Regione Piemonte à Turin, ont accepté, malgré leur éloignement, d'être examinateurs de cette thèse, je les en remercie vivement.*

*Enfin, **M. Ch. Obled** a été mon directeur de thèse. C'est-à-dire qu'il a suivi et motivé mon travail, toujours présent et jamais à court d'idées. De plus, il m'a ouvert la voie de l'enseignement. Pour tout cela, je souhaite lui exprimer ma gratitude.*

Ce travail de thèse s'est effectué conjointement dans l'équipe Hydro-météorologie du Laboratoire d'étude des Transferts en Hydrologie et Environnement et au sein du service Ressources en Eau de EDF-DTG à Grenoble.

Je tiens donc à remercier M. M. Vauclin, directeur du LTHE, et M. J.D. Creutin, responsable de l'équipe de m'avoir accueillie.

Les responsables du Service Ressources en Eau de EDF-DTG (M. D. Duband puis son successeur M. J. Miquel) m'ont donnée accès à des moyens de calcul ainsi qu'à leurs fichiers de données inestimables. Sans leur accord, cette thèse n'aurait pu se réaliser et je leur en suis très reconnaissante.

Mes remerciements vont aussi à tout le personnel du service Ressources en Eau et en particulier à l'équipe HYdrologie qui m'a accueillie (Rémi, Cécile, Muriel, Annie, Christian...) et à l'équipe Etudes / Prévi (Dominique, Marie-France...) qui suivait avec une attention particulière mes travaux. Je n'oublierais pas non plus J.P. Loffredo, l'informaticien toujours présent, au service des pannes...et du manque de mémoire !

Je pense aussi à Yves Rodriguez qui m'a accueillie à EDF pour un projet de fin d'études....qui ne se termine qu'aujourd'hui. Merci pour m'avoir initiée aux « analogues ».

Je tiens aussi à remercier, pour leur soutien financier, le Programme National « Risques Naturels » de l'INSU, le Ministère des Affaires Etrangères par l'intermédiaire du Programme d'Actions Intégrées PICASSO et celui de l'Education Nationale, de la Recherche et de la Technologie qui m'a accordé une bourse de thèse MESR.

A tous les membres du labo, merci d'avoir été là. Aux secrétaires, nos bonnes fées, toujours souriantes, toujours présentes. Aux anciens thésards, GM, Sophie, Robert toujours présents pour le rire et les coups de blues. A la nouvelle génération de thésards, aux DEA qui passent, pour la vie de tous les jours. Aux permanents, à ceux que j'oublie, pour la vie du labo, faite d'instants partagés autour d'un PC ou d'un café, de discussions....

Enfin, une mention spéciale aux habitués du 38 rivoire de la dame, sans qui la vie serait moins gai. Et à Vincent, pour son mois d'août passé à faire de la PTAO !!!!!
A vous tous et à Jean-Michel, merci pour avoir rendu ces 3 années inoubliables.

SOMMAIRE

SOMMAIRE.....5

INTRODUCTION GENERALE.....13

CHAPITRE I: La prévision par une technique d'analogue - méthode actuelle et bibliographie

I.1 Objectifs de la méthode.....21

I.1.1 Les besoins en prévision de précipitation

I.1.2 Les outils de la prévision

I.2 Historique de la méthode à EDF et état actuel.....23

I.2.1 Historique de la méthode

I.2.2 Etat actuel de la méthode opérationnelle

a) Les données disponibles

b) La sélection des analogues

c) La prévision des pluies

d) Fonctionnement du modèle opérationnel

e) Exemples et performances

I.3 Approches comparables.....34

I.3.1 La prévision à court terme

a) Prévision de température

b) Prévision de pluie

c) Prévisions diverses à court terme

I.3.2 Prévision à long terme

I.3.3 Divers

I.4 Historique des tentatives d'amélioration.....48

I.5 Conclusion du chapitre I.....53

CHAPITRE II: Prédicteurs, prédictands et expression de la prévision

II.1 Les prédicteurs.....59

II.1.1 Les Radiosondages

II.1.2 L'Analyse en Composantes Principales ou ACP

II.1.3 L'Analyse en Composantes Principales de Processus ou ACPP

a) Formulation de l'ACPP (Braud, 1990)

b) Choix des fonctions de base $e_i(\underline{x})$

c) Résultats

d) Interprétation dans l'espace des Composantes Principales

e) Conclusion

II.1.4 L'interpolation en points de grille

II.2 Les prédictands.....79

II.2.1 Les bassins français

II.2.2 Les bassins catalans (Espagne)

a) Localisation de la Catalogne (Llasat & Puigcerver, 1992)

b) Les stations pluviométriques

c) Les groupements

II.2.3 Les bassins italiens

II.3 Expression et évaluation de la prévision.....88

II.3.1 Expression de la prévision

a) Prévision pluie / non pluie

b) Prévision probabiliste en classes de pluie

II.3.2 Evaluation des performances d'une prévision probabiliste

II.4 Conclusion du chapitre II.....92

CHAPITRE III: Méthodes fondées sur les Composantes Principales

Introduction.....97

III.1 Algorithmes.....98

III.1.1 Méthodologie

a) Validation croisée

b) Méthodes de référence

c) Simplifications

III.1.2 La sélection des variables

a) La sélection ascendante simple

b) La sélection ascendante des K meilleures variables

III.2 Résultat de la sélection…….....107

III.2.1 Sélection ascendante simple sur 36 CP (3*12)

a) Evolution de l'indice de réussite en fonction du nombre de CP retenues

b) Comparaison avec les méthodes de référence

c) Les 8 premières CP retenues

III.2.2 Elimination de l'épaisseur

a) Comparaison 24 / 36 CP

b) Les 8 premières CP retenues

c) Vérification du choix N = 50 analogues

d) Sélection ascendante K=2 sur les 24 CP

e) Conclusion

III.2.3 Les différents types de CP

a) ACPP de corrélation ou de covariance

b) Application à la méthode référence

c) CP combinées

III.2.4 Prise en compte de l'évolution à 24h

III.3 Pondération.....123

III.3.1 Pondération des analogues

a) Méthode de pondération

b) Choix de la pondération

c) Méthode S-12CP

III.3.2 Pondération des CP

a) la méthode de pondération

b) choix de la méthode la plus performante

c) Application sur les 33 bassins

III.4 Application à la prévision probabiliste en classes de pluie.....132

III.4.1 Détermination du nombre N d'analogues

III.4.2 Sélection: méthodes S-8CP et S-12CP

III.4.3 Sélection et pondération

III.4.4 Synthèse des résultats pour les 2 types de prévision

a) En terme de score de réussite

b) En terme de CP retenues

III.5 Conclusion du chapitre III.....137

CHAPITRE IV: Méthodes basées sur les données brutes ou interpolées

Introduction.....141

IV.1 La Sélection des variables brutes (RadioSondages).....142

IV.1.1 Sur 6 bassins témoins

IV.1.2 Essai sur les 33 bassins

IV.1.3 Les RS sélectionnées

IV.1.4 Conclusion

IV.2 Utilisation de grilles et du score de Teweles-Wobus.....149

IV.2.1 Sélection des données en points de grille

IV.2.2 Définition du score de Teweles-Wobus

IV.2.3 Utilisation du score TW pour l'analogie

IV.2.4 Choix du type de données

IV.2.5 Choix des champs

IV.2.6 Optimisation de la grille

a) Taille

b) Localisation

IV.2.7 Conclusion

IV.3 Autres critères de sélection des analogues.....165

IV.3.1 Utilisation d'un critère seul

IV.3.2 Combinaison de critères

IV.3.3 Les anti-analogues

IV.4 Utilisation des réseaux de neurones.....168

IV.4.1 Présentation des modèles neuronaux

IV.4.2 Utilisation du réseau de neurones pour la prévision des pluies

IV-5 Conclusion du chapitre IV.....172

CHAPITRE V: Introduction de nouvelles variables

Introduction….....177

V.1 Variables synoptiques.....177

V.1.1 Présentation des données

V.1.2 Utilisation de ces champs

a) 1 champ, 1 échéance

b) 1 champ, 2 échéances

c) 2 champs, 1 échéance

d) 2 champs, 2 échéances

e) 3 champs, 2 échéances

V.1.3 Conclusion

V.2 Variables à caractère local.....187

V.2.1 Le radiosondage de Nîmes

a) Les données brutes

b) Les différents index : indices d'instabilité et autres paramètres

V.2.2 Méthodologie

V.2.3 Utilisation des données brutes de Nîmes

a) Choix du nombre d'analogues N_2

b) Choix de l'échéance - Détermination du palier

c) Choix de la méthode d'analogie synoptique

d) Les variables retenues

e) Conclusion

V.2.4 Utilisation des différents index

V.2.5 Quelques vérifications

a) L'analogie locale seule

b) Les 2 analogies, synoptique et locale, combinées

c) L'analogie locale avant l'analogie synoptique

V.3 Conclusion du chapitre V.....203

CHAPITRE VI: Extension à des bassin frontaliers - Validation sur les automnes 1994, 1995, 1996

Introduction.....207

VI.1 Extension à des bassins frontaliers.....207

VI.1.1 Les groupements espagnols

a) La méthode S-12CP et les méthodes de référence

b) La méthode S-12RS

c) La méthode TW-GR

d) Utilisation des données brutes de Nîmes

VI.1.2 Les bassins italiens

a) La méthode S-12CP et les méthodes de référence

b) Autres méthodes

c) Utilisation des données brutes de Nîmes

VI.1.3 Conclusion

VI.2 Validation sur les automnes 1994, 1995, 1996.....218

VI.2.1 Les données disponibles

a) Les géopotentiels 700 et 1000 hPa

b) Les données de pluie

VI.2.2 Validation quantitative de la prévision

a) Prévision de la pluie à 24 h

b) Estimation de la prévision des champs à 48, 72, et 96 h

c) Prévision de la pluie à des échéances supérieures

VI.2.3 Validation qualitative de la prévision

a) Représentation de la prévision

b) Prévision à 24 h

c) Prévision à 24 h avec champs prédicteurs observés ou prévus

d) Prévision à des échéances supérieures

e) Extension au Piémont

VI.2.4 Conclusion

VI.3 Conclusion du chapitre VI.....238

CONCLUSION GENERALE.....239

REFERENCES BIBLIOGRAPHIQUES ET BIBLIOGRAPHIE.....247

ANNEXES DES CHAPITRES II, III, IV, V et VI

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Les prévisions de précipitations à échéance de 2 ou 3 jours correspondent à des besoins socio-économiques importants. Les utilisateurs font d'ailleurs preuve, à l'égard de cette variable, d'une exigence au moins équivalente à celles qu'ils affichent pour d'autres, comme par exemple la température ou le vent. En particulier, la prévision des fortes pluies potentiellement génératrices de catastrophes pour les biens et les personnes, suscite une attention particulière: outre une bonne estimation de leur chronologie et de leur intensité, une très bonne localisation dans l'espace est aussi souhaitée.

Or cette prévision est particulièrement difficile à élaborer à l'aide de modèles météorologiques. Elle est notamment plus difficile que pour d'autres variables pronostiques plus classiques car :

- Au-delà des mécanismes hydro et thermodynamiques qui gouvernent ces variables, elle met en jeu des phénomènes supplémentaires, microphysiques, qui sont moins bien instrumentés que les précédents, et d'échelle nettement inférieure.

- Cette échelle est elle-même inférieure aux maillages couramment utilisés, ce qui nécessitent une paramétrisation toujours insuffisante pour bien représenter les phénomènes et notamment leurs extrêmes (cas de la convection).

- Enfin, il y a une interaction forte, mais pas totalement identifiée avec le relief, là encore, à des échelles souvent inférieures à la maille et à la précision topographique des modèles actuels.

Cela explique que les modèles, dont tout le monde reconnaît les progrès en ce qui concerne la prévision de la circulation (champs de pression, de vent, de température), restent encore très décevants en précipitations. Il n'y a qu'à reprendre les derniers grands événements hydrologiques du Sud-Est pour s'en convaincre.

C'est pourquoi, dans le domaine des prévisions automatiques et quantitatives des précipitations, la statistique peut venir en aide à la physique, par exemple par l'intermédiaire d'adaptation statistique en sortie de modèle météorologique. Elle permet, en particulier, de prendre en compte, à travers des séries de données, les particularités du lieu auquel on s'intéresse.
Une démarche courante consiste donc à combiner une modélisation physico-déterministe, pour la partie physiquement bien connue et correctement décrite par des équations, et une approche statistique pour la partie moins connue aboutissant à la variable d'intérêt, ici la précipitation.
C'est un peu la démarche du prévisionniste humain qui exploite les prévisions (qu'il sait robustes) de champs de pression, température etc., pour les transformer, à l'aide de son expérience et de son intuition, en variables du temps dit « sensible » (précipitation, brouillard, verglas, couverture nuageuse...). L'idée sous-jacente dans l'utilisation de l 'expérience est que si l'on reconnaît, dans la situation prévue, une situation déjà observée dans le passé, on peut s'attendre à des effets similaires en terme de temps sensible.
L'inconvénient de cette pratique réside dans la subjectivité humaine, dans le manque de permanence des prévisionnistes qui changent tous les 2 ou 3 jours, et dans la diversité de leur expérience. On a donc cherché à objectiviser cette démarche et à la rendre la plus automatique possible.
Les premières tentatives en ce sens remontent à M. Fontaine du Service Climatologique de la Météorologie Nationale qui, vers la fin des années 50, recherchait dans le passé des cartes ou situations types analogues à celle en cours.

La méthode de prévision quantitative de pluie journalière sur laquelle nous avons travaillé part de ce principe là. Elle est basée sur un traitement statistique des observations déjà effectuées qui revient finalement à formaliser la démarche du prévisionniste : sélectionner dans l'historique météorologique des *situations atmosphériques analogues* (car l'identité stricte n'existe pas) à une situation de référence pour en déduire les effets, à partir des précipitations qu'elles ont engendrées. Cette méthode subjective a été remplacée par une procédure numérique qui consiste à :

i) condenser l'information disponible sous la forme d'un grand nombre de données météorologiques par une Analyse en Composantes Principales,

ii) sélectionner, dans le fichier historique, les situations analogues grâce à un critère de similitude,

iii) calculer une prévision de pluie à partir des précipitations observées lors de ces situations analogues.

Cette méthode a été élaborée dans les années 70 par Duband (1970, 1974, 1981) pour répondre aux préoccupations des Services de Production Hydraulique et des Mouvements d'Energie d'Electricité de France. En effet, pour mieux gérer à court terme les ressources en eau et anticiper les risques de crue sur des bassins français montagneux équipés d'aménagements hydroélectriques, une prévision quantitative des précipitations pour les 96 heures à venir par tranches de 24 h, était nécessaire.
La méthode développée dès 1968 est toujours utilisée de façon opérationnelle depuis bientôt 20 ans.

Le but de cette étude a donc été une reconsidération complète de cette approche, tant sur les données utilisées (prédicteurs pour faire l'analogie et prédictands) que sur les critères de similitude. Il s'agit d'un effort significatif, initié par EDF qui a mis ses archives météorologiques à notre disposition et accompagné par le laboratoire LTHE et le CNRS qui nous ont attribué une bourse de thèse et des crédits de recherche au titre du Programme National des Risques Naturels. Compte tenu de ces moyens, notre travail a plutôt été focalisé sur les épisodes pluvieux intenses qui affectent, en automne, l'arc méditerranéen et occasionne des crues potentiellement très dommageables.
Le travail présenté dans ce mémoire s'articule de la façon suivante :

i) Après une présentation de la méthode actuelle et des différentes tentatives d'améliorations qu'elle a connues, ainsi que d'une description non exhaustive de ce qui se fait ailleurs en terme de prévision par une technique d'analogues dans le *chapitre I*,

ii) nous avons, dans le *chapitre II*, dressé l'inventaire des données disponibles - prédicteurs et prédictands - et travaillé sur l'expression de la prévision et les méthodes d'évaluation de ses performances.

iii) Si dans le *chapitre III*, le critère initial de sélection des situations analogues a été conservé mais significativement optimisé, en conservant toutefois les données condensées comme prédicteurs,

iv) dans le *chapitre IV*, nous sommes revenus aux données brutes et à des données en points de grille, plus proches de ce qui se fait actuellement, et de nouveaux critères de sélection ont été testés.

v) Quant au *chapitre V,* il montre les différentes tentatives d'introduction de nouvelles données météorologiques (locales et synoptiques) par lesquelles compléter l'analogie.

vi) Enfin, le *chapitre VI* présente la validation qualitative et quantitative des meilleurs résultats obtenus sur les 3 derniers automnes, ainsi que l'extension de la méthode à des bassins frontaliers italiens et espagnols.

CHAPITRE I :

LA PREVISION PAR UNE TECHNIQUE D'ANALOGUE

- méthode actuelle

- bibliographie

I.1 Objectifs de la méthode

I.1.1 Les besoins en prévision de précipitation

La prévision quantitative des pluies sur les bassins versants à crue rapide ou exploités pour la production hydroélectrique présente un intérêt crucial en terme d'anticipation du risque et d'optimisation économique. La gamme des échéances souhaitées est très large : inférieure à une heure pour les zones urbaines à quelques jours pour la production d'énergie.
Ainsi :

- *la prévision immédiate (0 - 3 h)* est utile pour les petits bassins versants ou les bassins urbains de quelques dizaines de km^2 qui ont un temps de réponse de l'ordre de l'heure. En effet, une prévision quantitative de la pluie peut alors aider à améliorer la prévision des crues de ces bassins versants à risque. Et en hydrologie urbaine, elle peut aussi contribuer à améliorer la gestion des réseaux d'assainissement pluvial.

- *La prévision à très court terme (6 - 12 h)* permet la connaissance du risque de pluies importantes sur des zones un peu plus grandes (quelques centaines de km^2). Elle est nécessaire afin d'organiser l'alerte du personnel d'astreinte.

- Enfin, *la prévision à court terme (1 à 5 jours)* permet, quant à elle, une mise en alerte au niveau journalier et une mise en vigilance jusqu'à 3 ou 4 jours en avance intéressante et pour la gestion des ressources en eau (lâchers de barrage) et pour le suivi des week-ends. Elle se fait sur des bassins versants de l'ordre de 100 à 1000 km^2.

I.1.2 Les outils de la prévision

Le développement des méthodes de prévision immédiate des précipitations est étroitement lié à l'essor des moyens d'observation de l'atmosphère en temps réel et particulièrement au *radar* et à *l'imagerie satellitaire*. Elles sont basées directement sur les données disponibles, l'approche la plus courante étant l'extrapolation d'imagerie radar qui est cependant rapidement limitée en raison de sa difficulté à prendre en compte la dynamique des phénomènes. En zone montagneuse, et pour des phénomènes convectifs, son échéance maximale n'excède pas 1h30 à 2h. De plus, si cette

extrapolation permet de prévoir le déplacement d'un système pluvieux régulier, en terme de quantité les incertitudes sont encore trop importantes. Enfin, des tentatives ont lieu actuellement pour superposer aux cellules observées un modèle de nuage (1D) capable d'introduire de la dynamique et d'étendre l'extrapolation (Dolcine, 1997; Bell & Moore, 1997).
Les données satellitaires, quant à elle, permettent une bonne transition vers l'échelle synoptique et des échéances plus longues (3 à 6 h) mais elles sont beaucoup plus difficiles à quantifier en terme de cumuls pluviométriques.
A côté de cela, des modèles météorologiques à échelle fine, incluant une description très détaillée des mécanismes de formation des pluies sont susceptibles de fournir une prévision quantitative des précipitations à une telle échéance. Cependant ils présentent encore des anomalies, comme les problèmes liés aux incertitudes des mesures et ceux dus à l'approximation (cf. Ducroc, 1997). Et, pour la plupart, ils sont encore au stade de la recherche, tout comme l'assimilation des données radar et satellite dans ces modèles.

De leur côté, si les *modèles météorologiques opérationnels* autorisent des échéances de 6 à 12 h au minimum, ils ne proposent en sortie des valeurs de précipitations que sur des mailles de 30 x 30 km^2 (PERIDOT), trop lâches par rapport à la taille des bassins versants pour être utiles pour la gestion des crues. Seuls des modèles de recherche, ou en opérationnel, des *modèles à aire limitée* (LAM) descendent jusqu'à des mailles de 5 x 5 km^2. Des recherches sont également menées pour le couplage des données radar à celles obtenues par l'imagerie satellitaire (Browning et Collier, 1989).

Pour une échéance à quelques jours, différents *modèles météorologiques* peuvent être utilisés, qu'ils soient globaux, hemi-globaux ou à aire limitée (LAM) suivant l'échelle souhaitée. Cependant, vu la complexité et le nombre des phénomènes mis en jeu, la statistique a été introduite aussi bien en tant que méthode de remplacement (régression, analyse discriminante) qu'en tant que méthode complémentaire utilisant une *méthode statistique en sortie de modèle météorologique*. Celle-ci se sert d'un modèle statistique pour relier le prédictand aux prédicteurs, issus d'un modèle numérique. On la connaît sous le nom de MOS ou « Model Output Statistic » (Glahn and Lowry, 1972).

C'est une approche de ce type que nous nous proposons d'étudier. En effet, la méthode de prévision par analogue est considérée comme une méthode statistique en sortie de modèle

météorologique puisqu'elle utilise des champs issus de modèles météorologiques pour faire l'analogie, avec des situations archivées sur lesquelles elle appréciera sa statistique.

I.2 Historique de la méthode à EDF et son état actuel

I.2.1 Historique de la méthode

Dès les années 60, le besoin s'est fait sentir dans les Services de la Production Hydraulique et des Mouvements d'Energie de EDF d'avoir chaque matin une connaissance, la plus précise possible, de la quantité de pluie qu'il va tomber dans les prochaines 24, 48 et 72 heures sur des bassins versants équipés d'aménagements hydroélectriques (de 500 à 3000 km^2), bassins essentiellement situés dans les régions montagneuses du Sud de la France (Jura, Alpes, Massif Central, Pyrénées). La réponse de ces hauts bassins variant de quelques heures à 24 ou 36 heures, ce délai permet de passer des pluies aux débits grâce à une relation préétablie.
C'est donc la nécessité de fournir aux exploitants cette information de débit sous forme numérique et graduée en probabilité qui a conduit EDF, à la fin des années 60, à mettre au point une méthode de prévision quantitative des précipitations pour permettre une meilleure gestion à court terme des ressources en eau.

Après quelques années de réflexion, une première méthode de prévision quantitative des précipitations à 24 et 48 heures a été élaborée par Duband (1970, 1971 et 1974). Son principe est apparemment simple. Il reprend et formalise le travail du prévisionniste qui consiste, en étudiant la carte météorologique du jour, à en prévoir les effets. Pour cela, il se sert de son expérience et des situations semblables qu'il a pu voir auparavant et dont il connaît les effets pour en déduire ce qui va se passer. Dans les années 60, Fontaine à Météo-France avait eu l'idée de rechercher manuellement dans une archive cartographiques de telles situations.
La méthode proposée consiste donc à rechercher dans le passé, mais de manière automatique cette fois, les situations les plus ressemblantes à la situation du jour au sens des données météorologiques pour en déduire la quantité de pluie qu'elle peut engendrer.

Le choix des données météorologiques utilisées a été déterminé en fonction de:

- *l'objectif à atteindre:* effectuer une prévision numérique des précipitations localisées d'abord à 24 h puis pour les 3 jours à venir,

- *et par le souci d'être opérationnel* en limitant la collecte des informations à des données facilement et rapidement accessibles.

Ainsi, le niveau de la surface 700 hPa (ou géopotentiel 700 hPa, situé autour de 3000 m d'altitude) à 00 h a été choisi pour caractériser les grandes lignes de la circulation atmosphérique. A cette altitude, la pression présente une inertie et une stabilité plus importante que le champ de pression au sol. Cependant, elle reste sensible à l'influence de perturbations organisées dont la durée d'activité peut varier de 12 à 36 heures.

Quant au champ de pression au sol à 06 h, cette variable et sa variation semblent être déterminantes pour l'intensité de la précipitation à situations atmosphériques comparables au niveau 700 hPa (Duband, 1974). Enfin, le champ de température à 700 hPa est un témoin de l'état thermique de la troposphère moyenne.

Très dépendante des données météorologiques (géopotentiel 700 hPa et pression au sol), la méthode a donc évolué en fonction des données fournies par Météo-France. Ainsi, dans les années 70, pour prévoir la pluie de la journée courante C, la procédure numérique de cette méthode pouvait se résumer en quelques points:

- *constitution d'un fichier historique* (1953-1973) contenant les données météorologiques énumérées ci-dessus, caractéristiques de la situation météorologique de la journée sur l'Europe centrale et de l'Ouest ainsi que les précipitations tombées en 24h (07h-07h) sur 19 groupements pluviométriques des régions montagneuses du sud de la France (voir chapitre II la définition des groupements),

- *récupération des données météorologiques observées* à 00 h (géopotentiel 700 hPa) et 06 h (pression au sol) le jour à prévoir, après dépouillement des radiosondages collectés par Météo-France.

- *condensation de l'information* fournie par les données météorologiques, souvent liées entre elles voire même redondantes pour certaines, par une Analyse en Composantes Principales,

- *sélection*, dans le fichier historique, des 25 journées les plus ressemblantes à la journée C au sens de la forme de la surface de pression 700 hPa,

- *calcul de la prévision de pluie* à attendre dans les 24 prochaines heures pour les 19 groupements pluviométriques. Celle-ci est déterminée grâce à une équation de régression multiple

entre le champ de précipitation P et les champs de pression au sol S et de température à 700 hPa T, calée sur l'échantillon des analogues: $R = f (S , T) + \varepsilon$, puis appliquée à la journée courante.

- pour la prévision à 48 heures, les mêmes opérations sont effectuées sur les valeurs des Composantes Principales (CP) de la journée C+24h. Cependant, les modèles de prévision n'étant encore qu'à leurs balbutiements à cette époque, ces CP étaient calculées par autocorrélation à l'aide des CP à C-24h, C-48h et C-72h.

La quantité de pluie est donnée par l'intermédiaire de la médiane 50% et d'une plage d'incertitude 10-90%.

Au début des années 80 (Duband, 1980 et 1981), quelques modifications et compléments ont été apportés:

- le fichier historique est complété du 1/01/1953 au 31/12/1980,
- le champ de pression au sol à 06 h a été remplacé par le champ de géopotentiel 1000 hPa à 00 h dont l'altitude est proche du niveau du sol,
- les champs de géopotentiels 700 et 1000 hPa sont relevés en 37 stations de radiosondage au lieu de 25 auparavant,
- la prévision est effectuée pour 33 groupements pluviométriques,
- le critère de sélection des journées analogues a pris la forme qu'il a actuellement (cf. § suivant); il est devenu plus sévère puisqu'il introduit en plus d'une distance une notion de similitude avec le champ de géopotentiel 1000 hPa,
- pour la prévision à 48 h, les CP utilisées proviennent cette fois des valeurs des géopotentiels 700 et 1000 hPa prévues par un des premiers modèles numériques de météorologie dynamique, que EDF était d'ailleurs chargé de tester par comparaison des CP calculées par autocorrélation.

Un peu plus tard, une modification du calcul de la prévision de pluie a été apportée: les équations de régression ont été abandonnées au profit d'une prévision probabiliste (quantiles 20, 60 et 90%) par interpolation de la distribution empirique cumulée des précipitations observées sur les analogues. Et ce n'est que lorsque les modèles météorologiques sont devenus plus élaborés et plus fiables, au milieu des années 80, que la prévision s'est étendue à 72 et 96 h.

I.2.2 Etat actuel de la méthode opérationnelle

Nous allons détailler ici le principe de la méthode de prévision quantitative de précipitations journalières proposée par Duband dès 1968 puis élaborée en 1974 pour atteindre sa forme actuelle dans les années 80 (Duband, 1980 et 1981). Cette dernière version est utilisée de façon opérationnelle par le Service Ressources en Eau de la Division Technique Générale d'EDF depuis bientôt 20 ans. Quelques résultats opérationnels obtenus durant l'Automne 1994 illustreront les capacités actuelles du modèle.

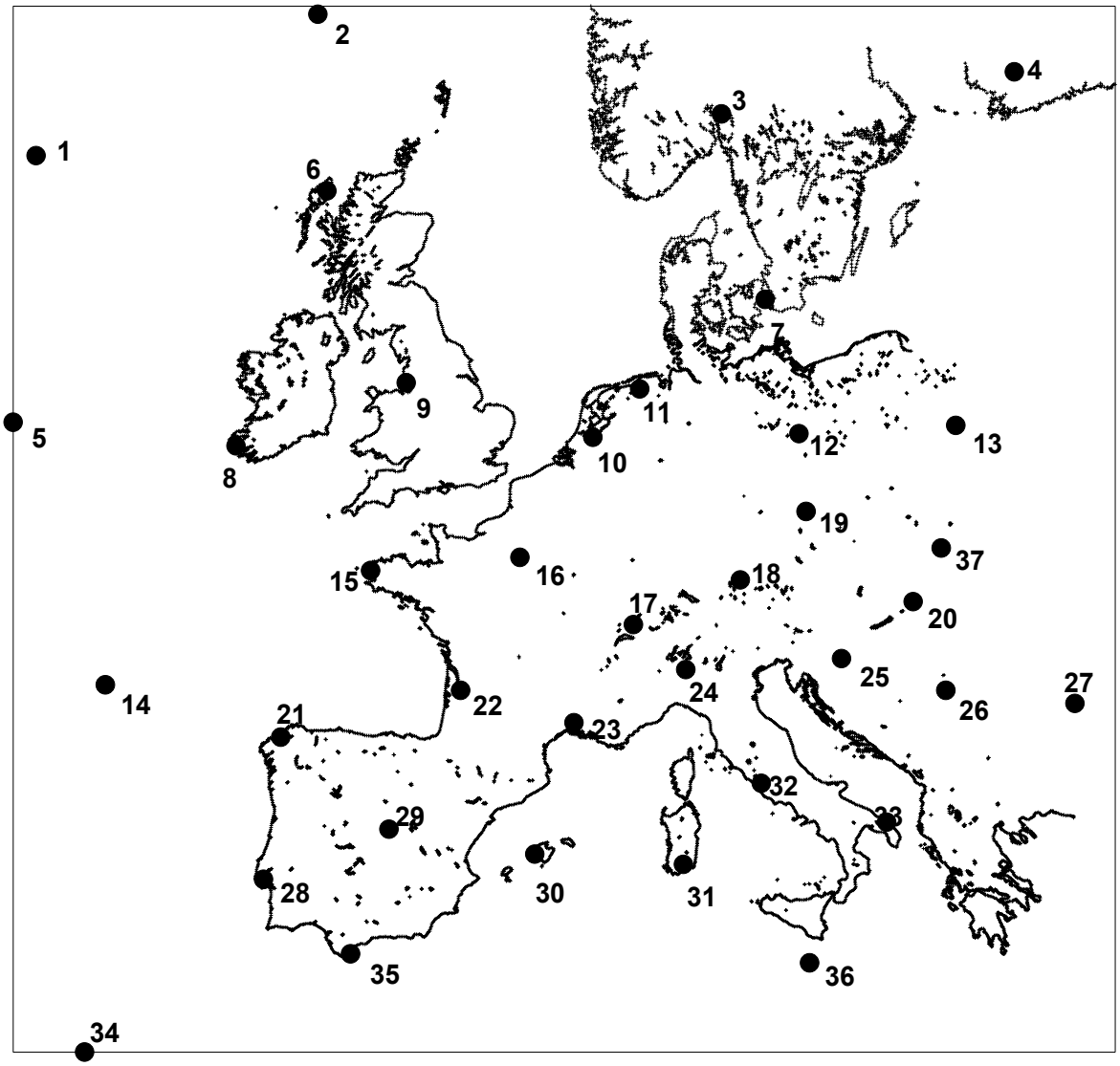
La méthode consiste à sélectionner, à partir d'un fichier historique de données météorologiques et climatologiques qui possède aujourd'hui plus de 40 années de données (1953-1993), un ensemble de situations analogues ou similaires, au sens de la circulation générale, à une journée de référence C, la journée courante où l'on veut faire la prévision.
A cette journée C sont associées les hauteurs de précipitation observées lors des journées retenues comme analogues. Puis, la fonction de distribution empirique cumulée est lissée pour obtenir une prévision des précipitations sous forme de quantiles.

I.2.2.a Les données disponibles

Les données utilisées sont de deux types:

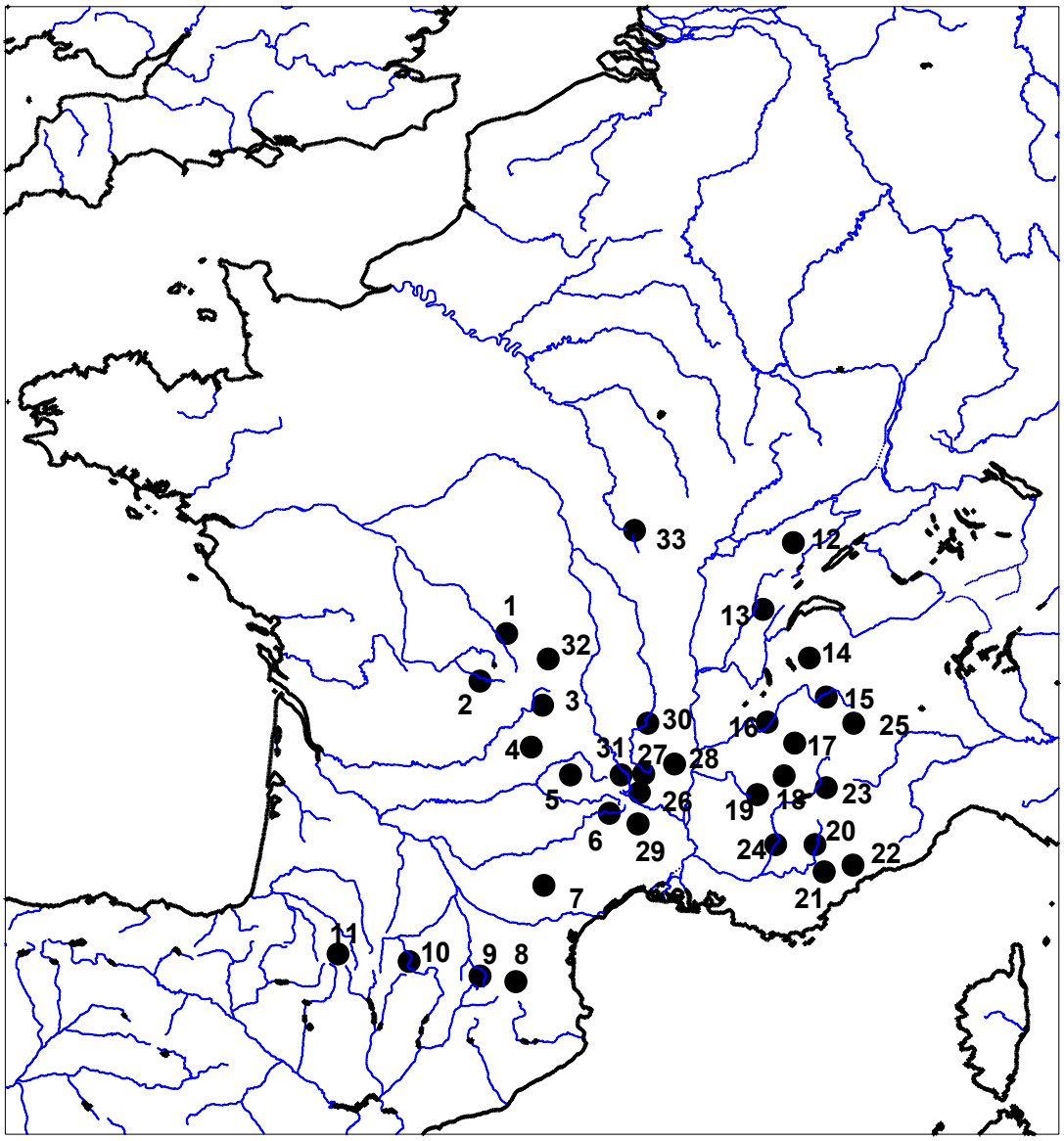
- *celles nécessaires à la sélection des analogues*: les données journalières des champs de géopotentiels 700 et 1000 hPa à 00 TU, collectées en 37 stations de radiosondage réparties sur l'Europe Centrale et l'Europe de l'Ouest (cf. figure I-1).
Cette information a été tout d'abord complétée par les données journalières de l'épaisseur de la couche 700/1000 hPa puis condensée par une Analyse en Composantes Principales (ACP) que nous détaillerons au chapitre II (cf. §II.1.2 et annexe II-1). Les Composantes Principales ou CP ont été préalablement centrées et réduites.

- *celles nécessaires à l'élaboration de la prévision probabiliste*: les relevés quotidiens de pluie sur 173 postes pluviométriques répartis autour et dans les massifs montagneux du sud de la France, moyennés sur 33 groupements ou bassins pluviométriques (cf. figure I-2).



1. Point I	11. Emden	21. La Corogne	31. Cagliari
2. Torshaw'n	12. Lindenberg	22. Bordeaux	32. Rome
3. Oslo	13. Logionowo	23. Nîmes	33. Brindisi
4. Jokioinen	14. Point K	24. Milan	34. Funchall
5. Point J	15. Brest	25. Zagreb	35. Gibraltar
6. Stornoway	16. Trappes	26. Belgrade	36. Malte
7. Kestrup	17. Payerne	27. Bucarest	37. Poprad Tatry
8. Valentia	18. Munich	28. Lisbonne	
9. Aughton	19. Prague	29. Madrid	
10. De Bilt	20. Budapest	30. Palma	

figure I-1: les stations de radiosondage



1. Creuse-Cher	12. Doubs	23. Haute Durance
2. Vézère-Vienne-Thaurion	13. Ain-Valserine	24. Durance moyenne
3. Dordogne	14. Arve-Fier	25. Mont Cenis
4. Cère-Maronne	15. Isère-Doron	26. Chassezac
5. Truyère-Lot inférieur	16. Isère moyenne	27. Loire supérieure
6. Haut Tarn-Haut Lot	17. Romanche-Arc inférieur	28. Doux-Eyrieux
7. Agout-Tarn	18. Drac	29. Gard-Cèze
8. Pyrénées Est	19. Buech-Drôme	30. Loire moyenne
9. Ariège-Vicdessos	20. Verdon	31. Allier supérieur
10. Pique-Garonne-Salat	21. BVI Verdon	32. Sioule
11. Gaves	22. Var-Tinee-Roya	33. Cure

figure I-2: les 33 bassins pluviométriques

Finalement, ces deux types d'information sont organisés en 3 fichiers historiques, un pour chaque saison:

- été du 16/04 au 31/08 (mais les analogues sont recherchés entre le 16/04 et le 15/10),
- automne du 01/09 au 30/11 ,avec des analogues dans cette même période,
- hiver du 01/12 au 15/04 avec des analogues sélectionnés entre le 16/10 et le 15/04,

car, pour être analogues, deux journées doivent se trouver dans la même période de l'année afin que les distributions de l'énergie solaire soient aussi similaires (Lorenz, 1969).
La saison d'automne a été rajoutée afin de mieux cibler les épisodes particulièrement violents dans le Sud-Est de la France pendant cette période
Ces fichiers historiques contiennent, pour chaque jour de 1953 à 1993:

- les 8 premières Composantes Principales (CP) du champ de géopotentiel 700 hPa: Z_1 à Z_8,
- les 8 premières CP du champ de géopotentiel 1000 hPa: S_1 à S_8,
- les 8 premières CP de l'épaisseur de la couche 700/1000 hPa: E_1 à E_8,
- les lames d'eau journalières sur les 33 groupements.

Par la suite seules les 6 premières CP des champs seront utilisées, considérant qu'elles expliquent assez de variance pour reconstituer de façon acceptable l'ensemble de l'information, ce qui peut être discuté (cf. chapitre II, § II.1.2 et II.1.3).

I.2.2.b La sélection des analogues

Elle se fait en deux temps. Après avoir sélectionné les analogues les plus proches dans l'espace (critère de proximité), les journées trop différentes de la journée courante C au niveau de leur forme sont éliminées (critère de corrélation).

Critère de proximité:

C'est une distance euclidienne dans l'espace des Composantes Principales (CP) du champ de géopotentiel 700 hPa. Les **journées J** du fichier historique trop éloignées, au sens de cette distance euclidienne, de la situation de la journée courante ou **journée de référence C** sont éliminées. Et seules celles situées à l'intérieur d'une sphère ou *boule de proximité* de rayon R_b, centrée au point représentatif de la journée de référence **C**, sont conservées (cf. figure I-3):

$$D^2(J,C) = \sum_{i=1}^{6} \left[Z_i(J) - Z_i(C) \right]^2 \leq R^2{}_b(C) \qquad \text{(I-1)}$$

où $Z_i(J)$ est la valeur de la **i^{ème} CP** du champ 700 hPa pour la **journée J**.

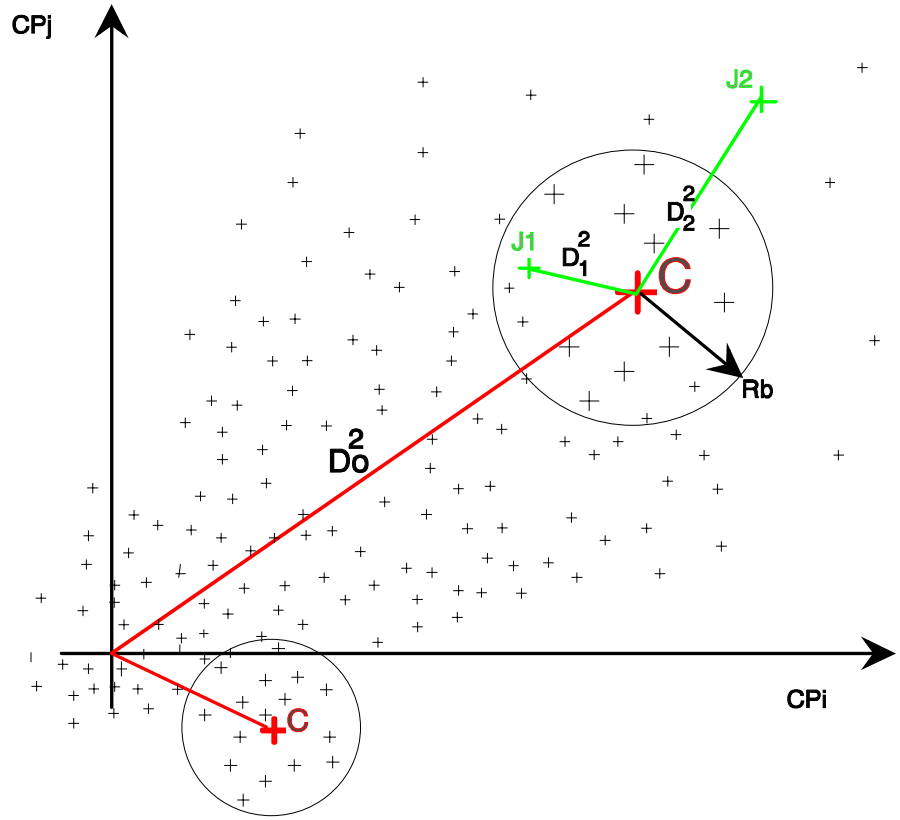


figure I-3: boule de proximité

Le rayon de la boule R_b,

$$R_b(C) = f\left(D_o^2(C) \right) = f\left[\sum_{i=1}^{6} Z_i^2(C) \right] \qquad \text{(I-2)}$$

est fonction de la distance D_o^2 (C) de la journée de référence **C** au barycentre du nuage de points constitué par les journées du fichier historique. Celui-ci se confond d'ailleurs avec l'origine 0 puisque les CP sont centrées réduites.

Et plus **C** est loin de l'origine (D_o^2 grand), plus R_b devient grand. En effet, les journées du fichier historique constituant un nuage de points plus dense à l'origine, plus la journée de référence est éloignée de l'origine, plus elle est « exceptionnelle » et donc plus ses voisins sont « rares » au sens euclidien. Et, comme le nombre d'analogues souhaité est compris entre 10 et 50 (seuils empiriques), R_b va être ajusté afin de trouver un nombre d'analogues compris dans cette fourchette.

Inversement, si la journée **C** est proche de l'origine (D_o^2 petit), le nombre de voisins va être plus élevé et le rayon R_b pourra être diminué.

Cependant, ce critère n'utilise que la seule variable Z. Aussi, un deuxième critère mettant en jeu le champ de géopotentiel 1000 hPa et l'épaisseur de la couche 700/1000 hPa a été construit pour affiner l'analogie.

Critère de corrélation (Duband, 1981):

Soit le vecteur V composé:

- des 6 premières composantes principales (CP) du champ de géopotentiel 700 hPa, Z_1 à Z_6,
- des 6 premières CP du champ de géopotentiel 1000 hPa, S_1 à S_6,
- et de la première CP du champ de l'épaisseur E_1.

Le critère de corrélation est:

$$u^2(J,C) = \frac{D^2(J,C)}{R^2(J,C)} \leq 6 \qquad \text{et } R^2(J,C) > 0.1 \qquad \text{(I-3)}$$

où **R(J,C)** est le coefficient de corrélation entre le vecteur V de la journée de référence **C** et celui d'une situation **J** ,

et **D^2 (J,C)** est la distance calculée avec l'équation (I-1).

Avec ce deuxième critère, seules les journées proches de la situation du jour de référence **C** au sens quadratique, mais aussi de forme analogue sont conservées.

I.2.2.c La prévision des pluies

Pour les 33 groupements de pluie, la moyenne et l'écart-type des pluies des analogues sont calculés afin de se faire une première idée des précipitations possibles.

Si le nombre d'analogues retenus est suffisamment important (seuil empirique fixé à 5), des prévisions stochastiques sont effectuées. Pour chacun des 33 groupements, la distribution empirique des précipitations est alors lissée (par une simple interpolation linéaire) et la prévision est donnée sous la forme des quantiles empiriques de précipitation 20, 60 et 90%.

I.2.2.d Fonctionnement du modèle opérationnel

Chaque matin depuis 1980, EDF reçoit de Météo-France, par ligne téléphonique, les valeurs des géopotentiels 700 et 1000 hPa à 00TU aux 37 points de radiosondage ainsi que leurs prévisions à 24, 48 et 72 h. Cette dernière les a préalablement reconstituées par interpolation à partir des points d'une grille de maille 1.5° x 1.5° pour les données du modèle Européen et de 1.5° x 2° pour celles du modèle Français car certains n'existent plus et les autres ne correspondent pas exactement aux points de grille.

Avec ces données, les Composantes Principales (CP), qui combinent linéairement ces données brutes, sont calculées:

- pour le jour même, à l'aide des valeurs aux points de radiosondage reconstituées à partir des *grilles observées*,
- et pour les 3 jours à venir, avec les valeurs aux points de radiosondage reconstituées à partir des *grilles prévues*.

Vers 6 heures, un bulletin est émis. Il donne pour le jour même et les 3 à venir:

- les 8 premières CP des 3 champs (700 et 1000 hPa, épaisseur de la couche 700/1000 hPa),
- le nombre d'analogues retenus,
- les 10 meilleurs analogues avec leur date, le coefficient de corrélation R^2 (J,C), la distance D^2 (J,C) et les précipitations sur les 33 groupements correspondants,
- la quantité de pluie prévue à 20, 60 et 90% en mm sur les 33 bassins.

I.2.2.e Exemples et performance

Quelques graphes (figure I-4) mettant en regard pour l'automne 1994 la pluie journalière et les prévisions à 20, 60 et 90%, nous donnent une idée qualitative des performances du modèle pour un groupement des Cévennes, correspondant au bassin du Chassezac

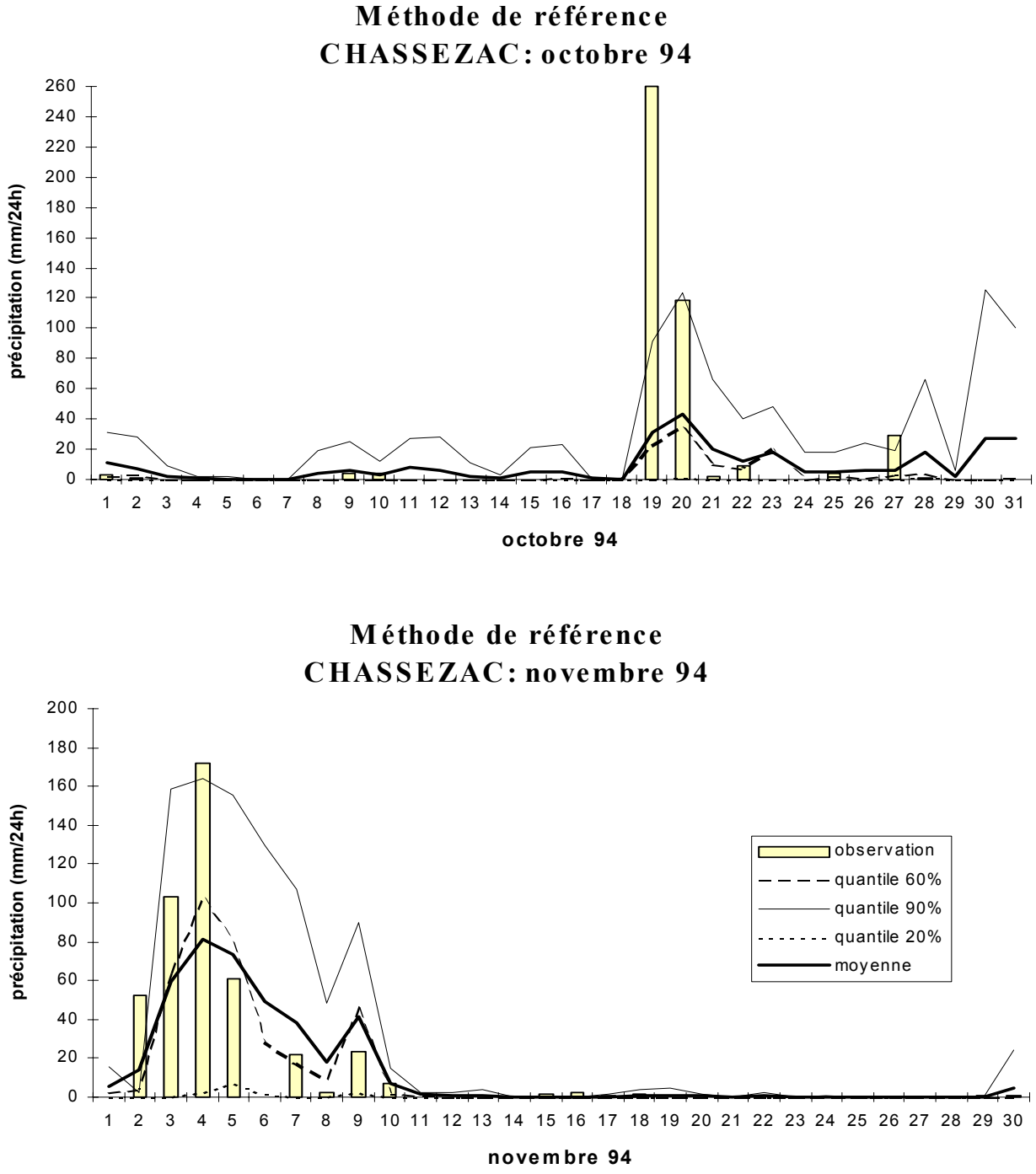


figure I-4: exemples de prévision pour l'automne 1994 et le groupement Chassezac

On peut remarquer que, si le pic de début novembre a été bien vu (quantile 90% supérieur à 160 mm), celui du 19 octobre a été largement sous-estimé. Cependant, les prévisionnistes qui utilisent quotidiennement ce modèle connaissent la réactivité du quantile 90% et lorsque celui-ci se situe autour de 100 mm, ils y prêtent attention et se mettent en vigilance.

Enfin, que ce soit dans la première quinzaine d'octobre, fin octobre ou encore après le pic des 4 et 5 novembre, un certain nombre de fausses alertes sont à déplorer, dues à une trop grande réactivité de la méthode.

Dans l'ensemble, la méthode donne d'assez bons résultats surtout pour les groupements du Nord. Cependant, il existe encore trop de cas où des jours peu pluvieux sortent comme analogues à des journées très pluvieuses (défaut d'alerte) et vice-versa (cas des fausses alertes), en particulier lors des épisodes cévenols d'automne. C'est pourquoi nous nous sommes focalisés sur cette saison, l'automne, et sur les groupements des Cévennes en particulier pour tenter d'améliorer la méthode.
Mais tout d'abord, nous avons voulu présenter quelques applications de l'approche par analogues.

I.3 Approches comparables

Historiquement les méthodes par analogie ont été utilisées à des fins diverses et variées.
Lorenz (1969), au début de la prévision numérique, est le premier à parler de recherche d'analogues et à en introduire les notions essentielles. Il s'en est tout d'abord servi pour évaluer le taux de croissance des erreurs d'analyse et de prévision.
Il a défini que deux états de l'atmosphère étaient analogues s'ils se ressemblaient assez pour que l'un soit équivalent à l'autre, moyennant une erreur raisonnable. Et il les considère comme similaires seulement si:

- les distributions globales en 3 dimensions du vent, de la pression, de la température, de la vapeur d'eau, des nuages,
- et les distributions géographiques de facteurs environnementaux comme la température de la surface de la mer ou la couverture de neige,

sont similaires. Ils doivent, de plus, appartenir à la même période de l'année afin que les distributions d'énergie solaire arrivant dans l'atmosphère soient similaires.

Mais, en général, les jeux de données disponibles ne possèdent pas toutes ces variables. Par exemple, les données de vapeur d'eau et de nuages, si elles sont présentes, sont interpolées de manière assez peu fiable sur les océans. Même les données de vent et de pression ne sont pas des

données indépendantes car elles proviennent de l'analyse de mêmes cartes et les interpolations au-dessus des régions démunies en stations de mesure ne sont, elles non plus, pas toujours très fiables. Lorenz a donc finalement décidé que deux états étaient analogues si les distributions de pression en 3 dimensions sur l'hémisphère Nord étaient ressemblantes. Aussi a-t-il travaillé sur des données de géopotentiel à 200, 500 et 850 hPa à 00 et 12 h, chaque niveau de pression étant sensé représenter 1/3 de la masse de l'atmosphère.

Pour mesurer la différence entre deux états de l'atmosphère, il a calculé les différences de hauteur de chaque niveau de pression à l'aide d'une distance euclidienne D_i^2 (k,l):

Soient p_1 = 200 hPa, p_2 = 500 hPa et p_3 = 850 hPa,

S1, ... , s_n les positions des n points de grille,

t_1, ... , t_p les temps observés par ordre chronologique,

Z_{ij} (k) est l'altitude de la **pression p_i**, au **point s_j** et à la **date t_k**. Et D_i^2 (k,l) s'écrit:

$$D_i^2\ (k,l)\ =\ \sum_{j=1}^{n} \left(Z_{ij}\ (k)\ -\ Z_{ij}\ (l)\right)^2 \qquad \text{(I-4)}$$

Cependant, les géopotentiels variant beaucoup moins en été qu'en hiver, si cette distance était utilisée telle quelle, les analogues seraient bien meilleurs en été. Aussi, une correction a été apportée:

$$E_i\ (k,l) = \frac{c}{2}\left(\log D_i^2\ (k,l) - \log \overline{D_i^2\,(k,l)}\right) \qquad \text{(I-5)}$$

où $\overline{D_i^2\,(k,l)}$ est une estimation de la valeur climatologique normale de $D_i^2\ (k,l)$ pour les saisons auxquelles appartiennent t_k et t_l. Et c a été choisi égal à 16/log2 afin qu'une augmentation de E_{ikl} de 16 points corresponde à une augmentation de D_i (k,l) d'un facteur 2.

Finalement, la *mesure de différence entre 2 états* est la moyenne de E_i (k,l) sur les 3 niveaux de pression:

$$E_{kl} = \frac{1}{3}\sum_{i=1}^{3} E_i\,(k,l) \qquad \text{(I-6)}$$

Et, comme il semble improbable que deux états de l'atmosphère venant de deux saisons différentes puissent être analogues à cause des flux de chaleur, par trop différents, la recherche des analogues se limite aux valeurs de E_{kl} pour lesquelles t_k et t_l appartiennent à la même saison.

Il préconise ensuite d'utiliser la recherche d'analogues pour prévoir le temps (Lorenz, 1980). En effet, cette méthode consiste à sélectionner dans le passé une situation analogue à celle d'aujourd'hui et à prévoir que la situation du lendemain sera la même que celle du lendemain de l'analogue. Elle possède l'avantage, par rapport aux méthodes statistiques classiques, d'introduire explicitement une non-linéarité dans la mesure où l'extrapolation est faite par l'atmosphère, obéissant à ses propres lois.

D'autres travaux ont, par la suite, utilisé la technique de recherche d'analogues pour faire de la prévision. On la retrouve, en particulier, pour prévoir à court terme (24 voire 48 heures), les éléments du temps sensible (température, précipitations, vent) mais aussi des éléments moins communs comme le risque d'avalanche, l'insolation journalière, l'évolution des cyclones... Cependant, nous verrons aussi des exemples de méthodes de prévision utilisant les analogues pour la prévision à long terme, saisonnière ou mensuelle.
Tout un ensemble de méthodes et d'application sont donc à notre disposition. Nous allons en faire le tour en quelques pages, la liste étant bien évidemment non exhaustive.

Remarque: pendant toute cette synthèse bibliographique, nous parlerons de jour (respectivement mois, saison) *de référence* pour les journées (respectivement mois, saison) où la prévision doit être effectuée.

I.3.1 La prévision à court terme

I.3.1.a Prévision de température

Woodcok, en 1980 (Woodcock,1980), a présenté et comparé trois méthodes de prévision du maximum de température journalière en janvier à Sydney en Australie:

i) La première méthode, très traditionnelle, met en place une équation de prévision à partir d'une série historique de prédicteurs de 6 années en faisant de la régression multiple pas à pas.

ii) Pour la deuxième l'équation de prévision est déterminée, toujours par régression multiple mais ajustée à partir d'un sous-ensemble de la série historique. Ce dernier est constitué des 50 journées les plus ressemblantes (ou analogues) à la journée de référence, au sens de la pression moyenne de la surface de la mer car les vents de surface sont un facteur déterminant pour la température maximale, particulièrement sur les côtes, comme à Sydney, où l'influence de la brise de mer est importante.

Les 50 journées analogues sont celles dont le score TW de Teweles-Wobus (Teweles & Wobus, 1954) est le plus faible:

$$TW = 100 \frac{\sum_{i=1}^{n} |E_i|}{\sum_{i=1}^{n} |G_i|} \qquad \text{(I-7)}$$

avec $\mathbf{E_i}$ différence entre le gradient de pression du jour de référence et du jour passé, entre 2 points de grille,

$\mathbf{G_i}$ maximum entre le gradient de pression du jour de référence ou de l'autre, entre les 2 mêmes points de grille,

n nombre de paires de points de grille adjacentes et dans des directions orthogonales.

Ainsi, une équation de prévision différente est obtenue pour chaque jour de référence, contrairement à la *méthode i* où il n'y a qu'une seule équation de prévision.

iii) Cette dernière méthode est la même que la précédente avec 50 analogues tirés au hasard.

Finalement, il apparaît que les meilleures prévisions sont obtenues avec la *méthode ii*.

I.3.1.b Prévision de pluie

Après Duband dans les années 70, c'est **Navarre** (notes) qui, au début des années 1980, a repris au Centre d'Etude de la Neige, un modèle de prévision des précipitations à échelle fine (échelle de la commune) sur les Alpes Françaises. Ce modèle, basé sur la sélection automatique de situations analogues (Vermot-Desroches, 1987; Villé, 1990), doit être considéré comme une interprétation fine des résultats fournis par les modèles de prévision météorologique.

Dans les massifs montagneux comme les Alpes où les répartitions de précipitations peuvent être très variées d'un massif à l'autre, voire d'un versant à l'autre, il paraissait nécessaire de descendre à l'échelle de la commune - ce que ne permettent pas des modèles comme Emeraude (maille d'une centaine de km) ou même Péridot (30 km) - notamment dans le cadre de la prévision des risques d'avalanche. Il est en effet utile pour celle-ci de connaître la quantité et la répartition géographique des précipitations, qu'elles soient sous forme liquide ou solide, à une échelle locale.

Dans ce modèle, la prévision des précipitations repose sur la recherche de journées analogues dans le passé, les journées étant définies par les champs de géopotentiel à 850 et 700 hPa condensés, ici aussi, par une Analyse en Composantes Principales. Des fichiers saisonniers ont été créés, constitués de trois mois glissants. Ainsi, pour prévoir la pluie de janvier, on ne cherchera des analogues que dans les mois de décembre, janvier et février des années passées.
Cette recherche se fait indépendamment sur les 2 niveaux et sur les 15 premières Composantes Principales (CP) de chaque champ, et sur 15 années de données.
Pour chaque niveau, l'évaluation de la proximité entre une journée de référence **C** où l'on veut faire la prévision et les journées du passé est effectuée en calculant la distance euclidienne sur les vecteurs de dimension 15 caractérisant les champs (vecteur d'état climatique):

$$D^2\ (J,C)\ =\ \sum_{i=1}^{15} \left[\ X_i(J)\ -\ X_i(C)\right]^2 \qquad \text{(I-8)}$$

où X_i (J) est la **i^{ème} CP** de la **journée J** du passé.
Les 10 journées avec les distances les plus faibles sont conservées, ceci pour les deux niveaux.
Ensuite, intervient une analyse des analogues retenus pour chaque niveau:

i) Tout d'abord, le coefficient de corrélation entre le vecteur d'état climatique de la journée **J** retenue comme analogue et de la journée de référence **C** est calculé. Il n'intervient pas directement dans la sélection des analogues (qui ne s'effectue qu'avec les distances) mais il apporte une indication sur la ressemblance spatiale des champs.

ii) Puis, pour un analogue $\mathbf{J_0}$ et la journée **C**, les distances entre (Jo+1 et C), (Jo-1 et C), (C+1 et Jo), (C+1 et Jo+1) et (C+1 et Jo+2) sont calculées. Cela permet d'éviter de retenir des journées semblables à 24 h mais dont l'évolution serait trop différente par la suite.
Ce modèle, qui n'est pas utilisé opérationnellement de nos jours, a cependant été exploité quotidiennement pendant toute la durée des Jeux Olympiques d'Albertville en 1992.

Dans les années 80 toujours, le **"Canadian Weather Service"**, organisme chargé des prévisions du temps sur tout le territoire du Canada ainsi que sur les mers environnantes, a utilisé la technique des analogues pour différentes prévisions (Wilson & Yacowar, 1980):

i) prévision quantitative de la pluie toutes les 12 heures sur une échéance de 60 heures.

Les champs de géopotentiels à 500 et 1000 hPa sont utilisés pour faire l'analogie. Les 20 meilleurs analogues, extraits de 10 années de données, sont retenus grâce au score de Teweles-Wobus TW (cf. éq. I-7) et à la corrélation entre la journée de référence et les analogues potentiels. Puis, la prévision probabiliste de pluie en 4 classes est donnée, pour différentes stations, par le rapport entre le nombre d'analogue dans chaque classe et le nombre total d'analogues.

Cette méthode de prévision a été comparée à des méthodes de prévision basées sur la régression multiple ou l'analyse discriminante multiple. Si celles-ci sont toujours meilleures que l'analogie à 24 h, elles perdent beaucoup en efficacité pour des prévisions aux échéances 36, 48 et 60 h. Elles devraient d'ailleurs rejoindre les score de l'analogie pour l'échéance 72 h. Par contre, cette dernière, si elle est moins performante que les autres, a un score qui reste à peu près constant au fur et à mesure que l'échéance augmente.

ii) prévision des chutes de neige dans les stations de ski.

Le fichier historique qui possède 10 années de cartes de temps, a été réduit en 35 types de cartes distincts. Le coefficient de corrélation, calculé entre la carte prévue du jour de référence et les 35 types de temps, permet de classer les types de temps du meilleur (corrélation la plus élevée) au moins bon (corrélation la plus faible). Cela va aider le prévisionniste, en utilisant d'autres prédicteurs si besoin est, à déterminer le risque de neige.

Toujours au Canada et au début des années 80, **Yacowar** (Yacowar, 1975) a aussi travaillé sur une méthode de prévision de pluie utilisant les analogues.

A partir des prévisions baroclines à 36 heures, il extrait d'un fichier historique de 24 années:

- les 20 meilleurs analogues à 500 hPa par corrélation entre la journée de référence et celles du passé (analogues à 500 hPa),
- les 20 meilleurs à 1000 hPa toujours par corrélation (analogues 1000 hPa),
- les 20 meilleurs tels que la somme des corrélations à 500 et 1000 hPa soit minimale (analogues combinés).

Connaissant la pluie des analogues, il calcule alors une fréquence cumulée empirique des précipitations des 20 analogues pour chacune des stations ce qui lui donne une prévision probabiliste de pluie.
Par comparaison avec des méthodes de régression où les équations de régression sont établies à l'aide de 5 années de données, il a clairement montré que les prévisions sont améliorées par l'utilisation des analogues et en particulier des analogues combinés des champs de géopotentiels 500 et 1000 hPa.

Ces exemples de prévision à court terme d'éléments climatiques ne sont bien sûr qu'un échantillon de se qui se fait avec les analogues, mais ils me semblent être assez représentatifs. On peut encore citer Kruisinga et Murphy (1983) qui, aux Pays-Bas, ont appliqué la méthode des analogues à la prévision probabiliste de température, ou Gordon en 1987 qui a utilisé les analogues pour prévoir l'évolution de la brise lors d'une compétition de voile et le vent au niveau d'un aéroport. De même, en Inde des méthodes d'analogues sont utilisées pour la prévision quantitative des précipitations durant la mousson (Singh *et al*., 1980).
Mais nous avons préféré nous en tenir là pour ce qui est de la prévision des précipitations et présenter des travaux quelque peu originaux utilisant aussi une technique d'analogues.

I.3.1.c Prévisions diverses à court terme

La prévision d'avalanches:

* A l'Institut de Mécanique de Grenoble (aujourd'hui le Laboratoire d'étude des Transferts en Hydrologie et Environnement, LTHE), dans les années 1970, **Obled**, en collaboration avec EDF-DTG, a adapté la méthode à la prévision des avalanches sur le domaine de Davos en Suisse (Obled & Good, 1980): pour un jour de référence, des situations analogues au sens nivométéorologique sont extraites du passé (période 1951-1976) et l'on regarde le nombre et le type d'avalanches qu'elles ont connues.
Il a donc tout d'abord fallu caractériser la situation nivométéorologique sur un domaine de 50 à 100 km^2 (massif ou grand domaine skiable). Or, cette situation nivométéorologique contient à la fois des variables de forçage (réchauffement, vent, chute de neige en cours) et des variables d'état du manteau neigeux (nombre de couches, stabilité etc...) qui ont, bien sûr, des unités différentes. Et

l'on pressent qu'il ne suffit pas de les standardiser au sens statistique pour les rendre comparables comme peuvent l'être deux géopotentiels à deux stations de radiosondage différentes.
Cela a donc souligné de manière encore plus aiguë que dans les travaux de Duband le problème des redondances entre variables et surtout celui de la notion de « distance » ou critère de similitude à utiliser. Est apparu notamment le besoin de pondérer ces variables de manière différente.
Cette méthode, sous le nom des plus proches voisins, fonctionne depuis 1983 sur le site de Davos.

* **Bolognesi** a présenté en 1993 (Bolognesi, 1993) un modèle hybride pour le diagnostic spatial des risques d'avalanche qui combine un raisonnement par analogie et un raisonnement par déduction.
L'objectif de ce modèle, en cours de développement par l'IFENA de Davos et le CEMAGREF de Grenoble, est de fournir en temps réel une estimation fiable des risques de déclenchement accidentel d'avalanches. Il est encouragé par certains directeurs de sécurité qui utilisent déjà des systèmes d'aide à la décision pour le déclenchement préventif d'avalanches.
Ce modèle reprend notamment la méthode dite "des plus proches voisins" pour fournir une liste des 10 journées passées les plus proches de la journée présente au sens nivo-météorologique, avec leur activité avalancheuse correspondante. On obtient donc, *par analogie*, un premier diagnostic des risques d'avalanche pour la journée. Une deuxième évaluation est faite en utilisant des principes d'intelligence artificielle: à partir de bases de données expertes, des diagnostics de risque sont délivrés par inférences logiques, reproduisant ainsi un *raisonnement déductif*.
Le diagnostic final est le résultat d'un arbitrage entre les diagnostics obtenus par analogie et par déduction.

La prévision d'insolation:

Au début des années 1980, au LTHE, une application beaucoup plus proche de la prévision des pluies a été faite pour prévoir les durées d'insolation journalières. Elle a permis de montrer l'intérêt de sélectionner les variables à prendre en compte dans l'analogie et de les pondérer différemment (Bois, Obled et Thalamy, 1981; Thalamy, 1981). Malheureusement, l'arrêt des programmes sur l'énergie solaire mit fin à ces recherches que nous décrirons plus en détail dans le paragraphe suivant (§ I.4).

La prévision de trajectoire de cyclones:

En Australie, **Woodcok et Keenan** (Woodcock & Keenan, 1979) ont fait de la prévision à 24 h de trajectoire des cyclones tropicaux en utilisant des cyclones passés similaires en distance et en trace.

i) Similitude en trajectoire:

Considérons le cyclone de référence dont on veut prévoir l'évolution. Il est positionné en A (cf. fig. I-0) et la veille il était en B. Tous les cyclones qu'il va être possible de sélectionner comme analogues doivent donc être centrés en B et leur position 24 heures plus tard est indiquée par un C sur la figure I-5. Seuls ceux dont le C est dans le cercle de centre A et de rayon R seront analogues par la trace.

R a été choisi égal à 1600 km, ce qui permet d'éliminer à peu près 1/3 des journées cycloniques possibles sur 915 au départ.

ii) Sélection par la similitude:

Ensuite, sur les 2/3 restantes, 50 journées analogues sont sélectionnés par le score TW de Teweles-Wobus sur le champ de géopotentiel 700 hPa.

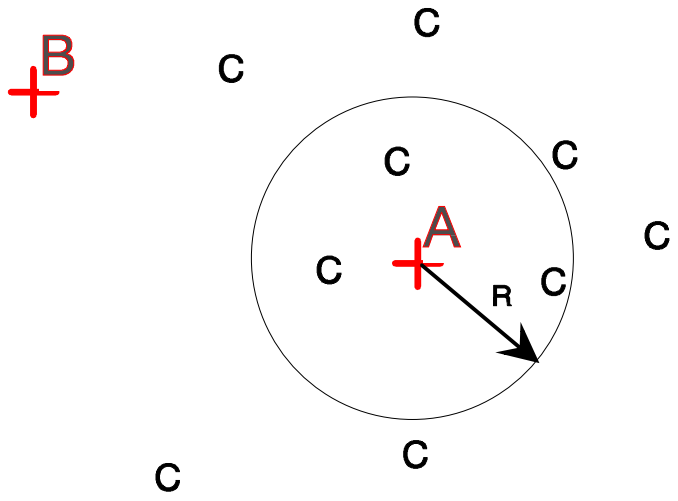


figure I-5: schéma de sélection des analogues par la trace
(tiré de Woodcock & Keenan, 1979)

Enfin, une équation de prévision est calculée par régression multiple pas à pas à partir de divers prédicteurs potentiels, dérivés des valeurs des champs de géopotentiel 1000, 700, 500 et 300 hPa, du mouvement des 24 h précédentes...
L'utilisation de ces analogues semble apporter une amélioration de la prévision en terme d'erreur moyenne absolue et de variance expliquée mais à cause de la faible taille de l'échantillon - la vérification a été faite sur une quarantaine de prévisions - aucun test de signification n'a pu être appliqué.

I.3.2 Prévision à long terme

L'utilisation de la technique des analogues pour une prévision à long terme (saisonnière ou mensuelle) est fondée sur l'hypothèse que des situations se répètent aussi pour ce qui concerne les variations saisonnières (ou mensuelles) d'un système climatique. On suppose donc que si le système atmosphère-océan se ressemble entre deux automnes, alors les deux hivers seront eux aussi similaires. Ainsi, s'il existe dans le fichier historique un bon analogue de l'automne en cours, l'hiver suivant cet automne analogue sera donné comme prévision de l'hiver à venir.
Cette idée est assez nouvelle en prévision à long terme. Cela a été mis en application de façon opérationnelle pour des prévisions mensuelles dans de nombreux services météorologiques dans le monde: au Canada (Shabbar & Knox, 1986), en Hongrie (Toth, 1988), aux Pays-Bas (Nap *et al*., 1981), en Union Soviétique (Gruza & Rankova 1980) ou encore en Angleterre (Murray, 1974; Bowen, 1976). Cependant, pour des prévisions saisonnières, seules quelques expériences ont été menées, notamment aux USA par Barnett et Preisendorfer (1978), Bergen et Harnack (1982) et ensuite Livezey et Barnston (1988, 1989).

Barnett et Preisendorfer (1977, 1978) ont présenté différentes méthodes de prévision saisonnière utilisant les analogues et pouvant être appliquées à la température, la pression, les précipitations, etc...
Ils ont formalisé le concept d'un *vecteur d'état climatique* proposé par Lorenz (1969) dont le mouvement représenterait l'évolution temporelle du système climatique. Un certain nombre de métriques, associées à l'espace dans lequel évolue ce vecteur, servent à sélectionner des états climatiques passés analogues à un état de référence. Et les prévisions sont basées sur l'évolution passée de ce vecteur d'état climatique.

Les données sélectionnées pour représenter un état climatique sont les moyennes mensuelles en différentes stations:

- du géopotentiel 700 hPa,
- de l'épaisseur de la couche 700/1000 hPa,
- de la température de l'air en surface,
- des précipitations,
- de la température de surface de la mer (SST).

Toutes ces séries temporelles, connues sur 20 ans, sont ensuite combinées pour ne former qu'un seul vecteur, le *vecteur d'état climatique*. A son tour, il est décomposé en fonctions empiriques orthogonales et les p premières, A_1 à A_p, sont conservées pour représenter l'ensemble de l'information.

Trois approches de sélection d'analogues ont été testées par une validation croisée sur les 20 années de données disponibles:

i) approche classique: on recherche la situation t* telle que l'état climatique soit le plus proche de celui de la situation de référence t où l'on veut donner une prévision.

La distance utilisée pour mesurer la différence entre les deux états est alors une distance euclidienne de la forme:

$$d_1^2\,(t,t') \;=\; \sum_{i=1}^{p} \left[\, A_i(t) \;-\; A_i(t') \,\right]^2 \qquad \text{(I-9)}$$

La date t* est telle que d^2_1 (t,t*) soit la plus petite. La prévision pour $t+\Delta t$ est donc ce qui a été observé à $t^*+\Delta t$.

Dans ce cas-là, on considère que c'est *l'état actuel du système climatique* qui est le plus important pour son évolution future.

ii) pour cette deuxième approche, *l'évolution du vecteur d'état climatique* est considéré comme aussi important pour la prévision que son état de référence. La distance utilisée est alors:

$$d_2^2\,(t,t') \;=\; \sum_{i=1}^{p} \left[\, \left(A_i(t) - A_i(t')\right)^2 \;+\; \left(A_i(t-\Delta t) - A_i(t'-\Delta t)\right)^2 \,\right] \qquad \text{(I-10)}$$

iii) en troisième approche, en plus de l'état actuel du vecteur climatique son *déplacement* est pris en compte par l'intermédiaire de sa dérivée temporelle:

$$d_3^2\,(t,t') \;=\; \sum_{i=1}^{p} \left[\left(A_i(t) - A_i(t')\right)^2 \;+\; \left(\dot{A}_i(t) - \dot{A}_i(t') \right)^2 . \Delta t^2 \right] \qquad \text{(I-11)}$$

Des essais sur la prévision saisonnière de la température de la surface de la mer de une à quatre saisons à l'avance, ont donné des résultats intéressants pour l'été mais pas aussi probants que ce qui était espéré. Cependant, les capacités de prévision sont très limitées par la faible longueur du fichier de données historiques (24 années de données soit 24 données par saison). Enfin, les résultats montrent que c'est à la fois l'état actuel de l'atmosphère et son passé récent qui sont importants pour prévoir son évolution future (méthode ii).

Une dizaine d'années plus tard aux Etats-Unis, **Livezey et Barnston** (1988,1989) ont utilisé une idée relativement récente dans la littérature (Harnack et al., 1986; Barnston & Livezey, 1986; Van den Dool, 1987): les *anti-analogues* en combinaison avec les analogues.
En quelques mots, une saison anti-analogue est une saison dont les données climatiques de base sont opposées à la saison de référence; ainsi, si l'on considère la température comme donnée climatologique de base, un hiver est anti-analogue à un hiver de référence froid (i.e. température en-dessous de la moyenne) si sa température est, au contraire au-dessus de la moyenne. La prévision pour le printemps suivant est alors l'inverse de celle du printemps suivant l'hiver anti-analogue, soit en-dessous de la moyenne si elle était au-dessus et vice-versa.
Cette utilisation d'analogues et d'anti-analogues pour la prévision (système mixte) présente l'avantage d'augmenter les chances de trouver des cas appropriés dans un fichier historique relativement succinct.

Ils ont appliqué cette méthode à la prévision saisonnière de température en 3 classes équiprobables (en-dessous, autour et au-dessus de la moyenne).
Les données utilisées pour rechercher l'analogie sont les suivantes:

- le champ de géopotentiel à 700 hPa dans l'hémisphère Nord extratropical,
- le champ de l'épaisseur de la couche 700/1000 hPa dans l'hémisphère Nord extratropical,
- la température de la surface de la mer dans des zones extratropicales et tropicales de l'hémisphère Nord,
- la température de l'air aux USA,

- l'indice des oscillations australes E+R, qui est la somme des pressions de surface à l'île de Pâques et à celle de Rapa.

Ces données ont été choisies car, tout en possédant un historique assez grand (35 ans de données), elles permettent de bien décrire l'atmosphère. Chacune des séries a ensuite été réduite en un petit nombre de séries temporelles pour créer le vecteur d'état climatique.

Ils ont ensuite considéré que les analogues peuvent être ou positifs (analogues) ou négatifs (anti-analogues). Et le signe de l'analogie correspond au signe du produit scalaire entre le vecteur d'état climatique de la saison de référence et celui de la saison analogue.

Quand les anti-analogues sont identifiés, leurs vecteurs d'état climatique sont multipliés par (-1) pour devenir des analogues positifs, tout en restant répertoriés comme anti-analogues.

Ensuite, la proximité de tous les analogues et anti-analogues est mesurée grâce:

- soit à la distance euclidienne classique entre les 2 vecteurs d'état climatique (cf. eq. I-6),
- soit à la *métrique ii* (éq. I-7) de Barnett et Preisendorfer où l'analogie est déterminée non pas sur une mais sur deux saisons consécutives antérieures.

Les N meilleurs analogues sont conservés (les N plus petites distances), en gardant toujours en mémoire s'ils sont analogues ou anti-analogues. Puis, la prévision de température en chaque station est déterminée de la manière suivante:

soit P^i_j la probabilité d'être dans la **classe i** pour la **station j** et pour **N** analogues:

$$P^i_j = \sum_{k=1}^{L} w_k \, . \, P^i_{kj} \qquad \text{(I-12)}$$

où P^i_{kj} = 1 si la **classe i** a été observée à la **station j** pour la **saison suivant k**,

= 0 sinon,

w_k est le poids donné à la saison analogue k.

La prévision pour chaque station j est alors la classe avec le P_j^i maximal.

La détermination d'une pondération des analogues et du nombre N d'analogues à retenir a donc constitué un des objectifs majeurs de cette étude. Des tests de validation croisée sur trois mois glissants (janvier-février-mars, février-mars-avril etc...) ont été effectués sur trois méthodes:

i) le système mixte (analogue et anti-analogue) avec une pondération décroissante des N analogues par l'inverse de la distance euclidienne utilisée, en normalisant de manière à ce que la somme des poids des analogues soit égale à 1,

ii) le système mixte avec une pondération uniforme des N analogues en 1/N,

iii) l'utilisation des seuls analogues, avec des poids décroissants.

Pour chacune, la sensibilité

- au nombre N d'analogues,
- au nombre de composantes P du vecteur d'état climatique,
- et aux variables à utiliser,

a été étudiée.

Les résultats les plus importants ont été récapitulés ci-dessous:

- les meilleurs scores en terme de prévision ont été trouvés pour les saisons froides (janvier-février-mars et février-mars-avril) puis pour les saisons chaudes (juin-juillet-août et les deux suivantes). Les saisons de transition, comme l'automne et le printemps, donne des résultats moindres car l'état climatique peut varier de manière très soudaine pendant ces périodes, rendant les prévisions beaucoup plus délicates,
- 3 des 5 variables utilisées dans la formation du vecteur d'état climatique (le géopotentiel 700 hPa, l'indice E+R et la température de l'air aux USA) sont nécessaires pour augmenter le gain de prévision. L'épaisseur semble être une variable redondante et la température de surface de la mer est beaucoup moins informative pour la prévision que les autres variables,
- le système mixte analogue / anti-analogue avec des poids décroissants donne les meilleurs résultats.
- le nombre optimal N d'analogues à retenir est assez variable, suivant les saisons et les méthodes.

Pour le système mixte avec des poids décroissants, un nombre N=10 d'analogues a été retenu dans les prévisions opérationnelles avec ce système quelle que soit la saison. Cependant ce nombre est fonction de la nature des données mais aussi de la taille du fichier d'apprentissage.

I.3.3 Divers

Enfin, on trouve des utilisations un peu particulières de la technique des analogues. Ainsi, au Centre d'Etude de la Neige à Grenoble, elle a été utilisée dans le cadre du développement d'outils

permettant d'estimer les conséquences d'un changement de climat sur l'enneigement des Alpes françaises (Martin, 1995).

A un état simulé par un modèle de circulation générale (MCG) est associée une journée réelle, sélectionnée dans un fichier de référence (1981-1995), et présentant un état analogue. Puis la simulation de l'enneigement est faite (par Safran et Crocus) en utilisant les données observées de la journée analogue, pour laquelle on dispose de toutes les données nécessaires.

I-4 Historique des tentatives d'amélioration

Trois études ont été faites par des élèves de l'Ecole Nationale de la Météorologie en 1985 et 1986 à l'occasion de stages à EDF-DTG ou au LTHE dans le but d'améliorer cette méthode de prévision quantitative des pluies appelée SENALOG.

Un autre travail utilisant la technique d'analogue a été effectué au LTHE en 1981, non pas sur la prévision quantitative des pluies, mais sur une méthode de prévision de durée d'insolation. Néanmoins cela nous a paru utile de l'évoquer ici à nouveau car de nombreuses idées nous ont paru intéressantes et applicables à notre méthode de prévision de pluie. Ses travaux ont d'ailleurs influencé les études précédemment citées, c'est pourquoi nous en parlerons en premier lieu.

① **En 1981, Thalamy** (Bois, Obled & Thalamy, 1981; Thalamy, 1981), a travaillé sur la prévision des durées d'insolation par une technique d'analogues.

Le fichier historique dans lequel les analogues sont recherchés contient 700 journées (2 mois/an de 1959 à 1969) représentées par 56 paramètres:

- la date: année, mois, jour,
- les rapports d'insolation RI (rapport entre le nombre d'heures d'ensoleillement de la journée et la durée d'ensoleillement astronomique à la date considérée) des journées J, J+1 et J-1, classés en 3 groupes (RI faible, moyen et fort),
- les 8 premières CP du géopotentiel 700 hPa des jours J, J-1 et J+1,
- les 6 premières CP du géopotentiel 1000 hPa des jours J, J-1 et J+1,
- les 3 premières CP de la température à 700 hPa des jours J, J-1 et J+1.

Le critère de sélection des journées analogues est la distance euclidienne entre la journée de référence C et les journées J du fichier historique pour un certain nombre de Composantes Principales choisies:

$$D^2(J,C) = \sum_{\text{CP choisies}} \left[X_i(J) - X_i(C) \right]^2 \qquad \text{(I-13)}$$

Les N meilleures journées avec les N distances les plus petites (N prédéfini) sont sélectionnées. Puis, on affecte à la journée de référence **C** le groupe de rapport d'insolation le plus présent dans les analogues.

Par la suite, Thalamy a tenté plusieurs améliorations:

i) L' introduction du rapport d'insolation de la veille en critère de sélection de deuxième niveau: dans les analogues sélectionnés par le critère de distance précédemment cité, seuls ceux dont le rapport d'insolation est dans le même groupe que celui de la veille du jour de référence C sont conservés. Cela a nettement amélioré les résultats.

ii) La sélection des variables nécessaires pour extraire les analogues dans le critère de distance (cf. éq. I-13) par différentes méthodes: sélection ascendante séquentielle avec ou sans remise en cause, sélection descendante séquentielle. Des méthodes plus systématiques n'ont pu être mises en oeuvre faute de moyens informatiques suffisamment performants. Mais c'est elle qui a, pour la première fois, testé la sélection automatique de variables.
Après s'être faite une idée plus précise sur les variables intéressantes pour la sélection des analogues elle a tenté d'agir sur le critère de distance de sélection pour améliorer la qualité des prévisions.

iii) L'optimisation de la distance:

- en prenant en compte l'évolution temporelle des journées. Une variable d'évolution est ajoutée dans la distance de sélection des analogues qui devient:

$$D^2(J,C) = \sum_{\text{CP choisies}} \left[X_i(J) + \Delta X_i(J) - \left\{ X_i(C) + \Delta X_i(C) \right\} \right]^2 \qquad \text{(I-14)}$$

avec $\Delta X_i(J) = X_i(J) - X_i(J\text{-}1)$: variable d'évolution entre J et J-1.
Elle fait ici implicitement l'hypothèse que l'évolution entre les journées J-1 et J est la même qu'entre J et J+1. Mais les résultats n'ont pas été concluants.

- en introduisant des pondérations de manière systématique sur les variables utilisées dans la distance de sélection dans le but de maximiser le taux de réussite (rapport entre le nombre de bonnes prévisions et le nombre de prévisions effectuées), soit en cherchant à annuler sa dérivée (mais alors on peut tomber sur un minimum), soit en procédant par tâtonnement. Les meilleurs résultats donnent un gain de l'ordre de 1% du taux de réussite.

iv) D'autres essais d'améliorations comme:

- la recherche du nombre optimal N de voisins à retenir (N=30),

- la pondération des voisins selon leur proximité à la journée de référence qui n'a pas donné de résultats intéressants.

Signalons enfin que Thalamy a utilisé un système d'apprentissage glissant ou validation croisée, où toutes les journées du fichiers, antérieures et postérieures à la journée de référence C (sauf celles de l'année en cours), sont utilisées comme analogue potentiel.

② **Mandon** (Mandon, 1985), au LTHE à Grenoble, a tout d'abord remarqué un certain manque d'homogénéité des pluies des analogues d'une même journée. En particulier, dans le cas de pluies intenses de type "cévenol" un nombre non négligeable de défaut d'alerte et de fausses alertes est à noter.

Pour remédier à cela, elle a cherché à introduire un critère de sélection intervenant à un deuxième niveau de sélection et applicable seulement en cas de mise en alerte. Ce critère devait permettre de séparer les analogues peu ou pas pluvieux des autres, plus intenses, dans le cas d'épisodes "cévenols".

Pour cela, 2 méthodes distinctes ont été testées:

i) L'étude du caractère discriminant de nouveaux paramètres à caractère local comme le vent, l'humidité et la température de l'air à Nîmes ou encore la température superficielle de la Méditerranée. Cela a permis la mise en oeuvre d'un test sur le produit vent∗humidité à 700 hPa.

ii) L'élaboration d'un deuxième critère barométrique plus approprié à la détection des épisodes "cévenols" l'a amenée à différencier les situations d'été, d'automne et d'hiver et à effectuer un test spécifique pour chaque saison car les Composantes Principales discriminantes ne sont pas les mêmes. En effet, alors qu'en automne le gradient de pression Nord-Sud permet de différencier les épisodes "cévenols" des autres, en hiver, c'est plutôt le gradient Est-Ouest.

Deux travaux ont ensuite été menés en parallèle, à Grenoble et à Toulouse, sur cette méthode:

③ **A** Toulouse à EDF-DTG, **Vallée** (Vallée, 1986) s'est penché sur la région du Sud-Ouest du Massif Central et sur les Pyrénées Orientales pour:

i) refaire les groupements pluviométriques de la zone considérée grâce à l'Analyse en Composantes Principales, à la classification hiérarchique ascendante et à des considérations d'ordre pratique. Mais ceux-ci n'ont pas, en terme de qualité de prévision, donné de meilleurs résultats que les anciens.

ii) introduire une pondération des radiosondages très irrégulièrement répartis (réseau dense sur l'Europe Centrale et beaucoup plus lâche sur l'Atlantique, cf. fig. I-1) avant d'effectuer l'Analyse en Composantes Principales (CP). Cette pondération a été optimisée en fonction de la surface d'influence de chaque radiosondage (polygone de Thiessen associé) et de son éloignement au centre de la zone d'étude.

L'information a ainsi été mieux condensée sur les premières CP: le pourcentage de variance expliquée est passé de 92.2 à 96.3% pour les 6 premières CP du champ de géopotentiel 700 hPa en automne. Et cela a permis de faire passer le pourcentage de variance expliquée par la prévision de 11,7 à 14,7.

iii) optimiser la sélection des analogues de 2 façons:

- La pondération des CP utilisées dans le critère de proximité (cf. eq. I-1) a tout d'abord été optimisée:

$$D^2(J) = \sum_{i=1}^{8} k_i \cdot \left(Z_i(J) - Z_i(C) \right)^2 \leq R^2_b \qquad \text{(I-15)}$$

En prenant 8 CP au lieu de 6, la meilleure pondération est (1.4 , 0 , 0 , 0 , 0 , 0 , 0.5 , 1) qui a encore permis de gagner 1.5 points (14,7 à 16,1%).

- Puis il a, à l'instar de Mandon, créé un critère de sélection de deuxième niveau avec le vent de Nîmes à 700 hPa à 12 h.

Soit Vs la composante sud du vent de Nîmes à 700 hPa et P la pluie prévue finale:

si Vs < 10 m/s, $P = 0.7 *$ (pluie moyenne des analogues)

si Vs > 10 m/s, $P = 0.7 *$ (pluie prévue à 90%).

Celui-ci n'est activé que lorsque le modèle prévoit un risque de fortes pluies sur les Cévennes (pluie prévue à 90% supérieure à 65 mm). De bons résultats ont été obtenus pour les bassins les plus touchés par les pluies intenses de flux de sud.

④ A la même époque, à Grenoble, au LTHE et à l'EDF-DTG, **Berlin et Cendrier** (Berlin & Cendrier, 1986) ont travaillé essentiellement sur les prédictands et les prédicteurs de la méthode. Ils ont essayé de:

i) redéfinir certains groupements pluviométriques du bassin de la Moyenne Durance où les résultats du modèle sont inférieurs à ceux des autres bassins. Mais là non plus, cela n'a rien apporté à la qualité de la prévision.

ii) prendre en compte l'organisation géographique du réseau de radiosondages afin de mieux condenser l'information météorologique en effectuant:

- en première approche, une Analyse en Composantes Principales sur un réseau de radiosondage plus équilibré (ACP équilibrée): le réseau initial allégé de certaines stations dans les zones les plus denses en stations.

- puis, une Analyse en Composantes Principales de Processus (ACPP), qui possède la particularité, par rapport à l'ACP classique, d'introduire des pondérations différentes aux stations irrégulièrement réparties.

Cela faisait suite aux travaux de Obled (1979) et Bouhaddou (1984) sur la théorie de l'ACP appliquée à des processus temporels ou spatiaux.

Pour la période 01/01/1953 - 31/05/1976 et pour l'automne et l'hiver, le pourcentage de variance expliquée par les 6 premières CP du champ 700 hPa est de:

* 93,1 % pour l'ACP classique,
* 92,8 % pour l'ACP régulière,
* 92,8 % pour l'ACPP (pondération par polygones de Thiessen)
* 96,3 % pour l'ACCP (pondération par triangularisation).

Enfin, durant mon DEA en 1994 (Guilbaud, 1994), différentes tentatives d'amélioration ont été effectuées:

i) pondération des CP utilisées dans la sélection des analogues: les performances du modèle augmentent lorsque l'on donne plus de poids à l'information proche de la zone étudiée,

ii) pondération des pluies des analogues en fonction de leur distance au jour de référence. Les résultats n'ont pas été encourageants.

iii) élimination de la boule de proximité qui limite le nombre d'analogues retenus: les résultats sont équivalents, avec un nombre d'analogues supérieurs. Cela permettra, par la suite, d'effectuer une deuxième sélection sur ces analogues.

iv) modification de la distance de similitude (cf. équation I-1) en introduisant les différences premières des CP ou les CP des différences premières des champs de géopotentiels eux-mêmes. Les résultats étaient plutôt positifs mais il y avait trop de jours sans analogue si l'on voulait que les méthodes restent sélectives et en même temps opérationnelles.

I-5 Conclusions du chapitre I et objectifs de nos travaux

Les différentes tentatives d'amélioration de la méthode de prévision de pluie par recherche d'analogues exposées précédemment sont autant de voies de recherche envisageables. Certaines n'ont été explorées que partiellement à cause des limitations dues aux capacités de calcul de l'époque. Des gains ont cependant été obtenus. Mais ils ont été jugés insuffisants (quelques % en terme de qualité de prévision) pour justifier la remise en cause d'un système de prévision opérationnelle assez lourd.

Aussi notre but est une **reconsidération majeure de cette approche**, de la manière la plus systématique possible, en nous inspirant des essais déjà effectués, des idées explorées par Thalamy et des résultats de notre recherche bibliographique pour aller plus loin dans l'amélioration de cette méthode.

Tout d'abord, **pour le prédictand** qu'est la pluie (cf. chapitre II), nous chercherons à:

- *étendre la prévision* à un plus grand nombre de bassins,
- *exprimer la prévision autrement* qu'en quantiles car il n'existe pas de score adéquat pour mesurer les qualités d'une telle prévision.

Pour les prédicteurs, il s'agit:

- *de recalculer les Composantes Principales* (CP) par une ACP de Processus afin de prendre en compte l'organisation géographique du réseau de mesure et de les comparer par rapport aux CP issues de l'ACP classique (chapitre II),

- *d'utiliser les données brutes ou interpolées sur une grille* et de ne plus passer nécessairement par les CP pour sélectionner les analogues (chapitre IV),

- *d'utiliser des données météorologiques nouvelles*, peut-être plus corrélées aux pluies, et construites à partir des champs déjà disponibles comme l'Instabilité Barocline, indicateur des zones d'instabilité (chapitre V),

- *d'injecter des informations plus locales* issues de radiosondages voisins de la zone d'étude comme Nîmes ou Palma en critère de deuxième niveau (chapitre V).

Avec toutes ces données et grâce aux moyens de calcul aujourd'hui disponibles, différents moyens seront utilisés pour tenter d'améliorer les performances de la prévision:

- *des méthodes de sélection et/ou pondération* des variables seront utilisées afin d'optimiser la distance de sélection des analogues (chapitre III),

- *de nouveaux critères de sélection des analogues* seront testés (chapitre IV).

Enfin, des tests de validation sur les derniers automnes (1994, 1995 et 1996) seront effectués pour différents bassins et pour les meilleures méthodes retenues (chapitre VI).

Note à l'attention du lecteur « pressé » :

Un grand nombre d'essais ont été réalisés durant cette étude, une certaine exhausitivité dans l'exploration faisant partie des objectifs de l'étude. Naturellement, tous ces essais ne se sont pas révélés également fructueux. Si une partie d'entre eux nous ont permis d'avancer et de construire une nouvelle méthode, certains n'ont apporté que des réponses partielles et d'autres encore se sont révélés sans intérêt véritable.

Cependant, nous avons tenu à les décrire tous dans ce mémoire, et ceci pour 2 raisons :

- tout d'abord, quelque soit leur résultat, ils ont tous représenté un travail parfois significatif,

- et ensuite, parce qu'ils ont été réalisés, pour certains, à l'instigation de collègues et de partenaires, parfois "très convaincants", qui trouveront ici une réponse, au moins partielle, à leurs interrogations.

Néanmoins, pour permettre une lecture plus sélective, nous avons marqué :

- de 2 astérisques (**) les essais infructueux qui se sont révélés sans aucun intérêt pour la suite du travail. Quelques lignes de synthèse placées en début du paragraphe traitant de l'essai (si celui-ci dépasse la demi page) résument alors l'idée testée. Cela permet de passer rapidement au paragraphe suivant.

- d'un seul astérisque (*) les essais qui n'ont pas abouti mais dont les conclusions restent importantes que ce soit dans la suite de l'étude ou en terme de perspectives. Dans ce cas, une lecture rapide est quand même suggérée.

CHAPITRE II :

PREDICTEURS, PREDICTANDS

et

EXPRESSION DE LA PREVISION

II.1 Les prédicteurs

La qualité d'une approche par analogue est directement tributaire du contexte des données disponibles. On entre donc d'emblée dans un compromis: celui entre la durée des archives sur lesquelles on peut s'appuyer, et leur richesse d'information. Disons que si l'on veut une archive longue, on doit se limiter aux champs de géopotentiels standards, ainsi qu'aux températures. Si l'on veut des données plus diversifiées, elles proviendront de modèles d'assimilation qui ont beaucoup évolué au fil du temps, même si le Centre Européen pour la Prévision Météorologique à Moyen Terme (CEPMMT) a fait un archivage homogène de 1981 à 1997.

Le choix des paramètres décrivant la circulation atmosphérique à court terme a donc été fait par Duband (1970) dès les années 70 en fonction de la méthode utilisée, la prévision par analogue nécessitant une période d'archivage des paramètres la plus longue possible, mais aussi de:

- *l'objectif à atteindre:* la prévision quantitative des précipitations sur des bassins d'environ 1000 km^2 pour les 48 h à venir,

- et par le souci de respecter des *contraintes opérationnelles* : collecte des données limitée à des données facilement accessibles.

II.1.1 Les radiosondages

Pour les différentes raisons évoquées ci-dessus, Duband a choisi les niveaux des surfaces 700 et 1000 hPa (ou géopotentiels 700 et 1000 hPa) et l'épaisseur de la couche 700/1000 hPa à 00 heure TU. En effet, le géopotentiel 700 hPa est utile pour caractériser les grandes lignes de la circulation atmosphérique. Le géopotentiel 500 hPa (situé autour de 5 000 m d'altitude) aurait aussi pu convenir car il est situé à peu près au milieu de l'atmosphère donc il représente bien la circulation générale. Cependant, au niveau 700 hPa se forment en moyenne les précipitations d'été et d'automne (850 hPa en hiver). Quant au géopotentiel 1000 hPa, il peut donner (cf. Duband, 1974) une indication sur l'intensité des précipitations, à situations atmosphériques comparables à 700 hPa. D'autre part, l'épaisseur de la couche 700/1000 hPa peut être considérée comme un témoin de l'état thermique de l'atmosphère.

Pour ce qui est des points de mesure, au nombre de 37 radiosondages à partir de 1980 (cf. fig. I-1), ils ont été choisi en fonction du réseau existant à l'époque et plus étendu vers l'Ouest que vers l'Est, le type de temps moyen étant essentiellement un régime d'Ouest. Malheureusement, certains radiosondages ont été supprimés depuis sur l'océan Atlantique (points bateau) et les valeurs en ces points sont reconstituées par analyse a posteriori du champ de pression connu de nos jours en une grille de 1.5 x 1.5 ° pour les données du modèle européen et 1.5 x 2 ° pour celles du modèle français.

Jusqu'à présent, ces données, archivées de 1953 à 1993, ont été utilisées après transformation par une Analyse en Composantes Principales ou ACP. Mais d'autres types de traitements, peut-être plus appropriés, peuvent leur être appliqués, comme l'Analyse en Composantes Principales de Processus (ACPP) ou une interpolation aux noeuds d'un réseau régulier.

II.1.2 L'Analyse en Composantes Principales ou ACP

Dès les débuts de cette approche (Duband, 1970), il a été considéré que la pluie sur un bassin était potentiellement influencée par l'ensemble des 3 champs disponibles (géopotentiels 700 et 1000 hPa et épaisseur de la couche 700/1000 hPa) car cette pluie peut venir aussi bien du Nord-Est, de l'Est, du Sud etc...
Pourtant, ces champs sont caractérisés par des mesures en 37 points de radiosondage, mesures visiblement très intercorrélées et donc partiellement redondantes. D'où l'idée de condenser préalablement cette information par une technique intrinsèque comme l'Analyse en Composantes Principales ou ACP. Cette technique est intrinsèque dans la mesure où elle ne prend pas en compte de variables exogènes (et notamment le prédictand), comme le ferait la corrélation multiple, l'analyse factorielle discriminante ou l'analyse canonique.

Aussi, pour chaque saison (été, automne, hiver) et pour la période 1953-1993, une ACP a été effectuée sur chacun des 3 ensembles de 37 variables. De nouvelles variables orthogonales et sans dimension (normées par leur variance ou valeur propre associée λ_k), les Composantes Principales ou CP, ont donc été calculées:

$$Z_{ij} = \frac{1}{\sqrt{\lambda_k}} \sum_{k=1}^{37} \alpha_{kj} \frac{X_{ik} - M_k}{\sigma_k} \qquad \text{(II-1)}$$

avec j = 1 à 8 au lieu de 37,

i = 1 à n observations quotidiennes disponibles,

α_{kj} (k= 1 à 37), cosinus directeurs de la CP d'ordre j,

λ_j variance de la CP d'ordre j ou valeur propre,

M_k, σ_k moyenne et écart-type de la variable d'origine X_k calculés sur la période disponible (1953-1993) du radiosondage k pendant la saison considérée.

La technique de l'ACP est expliquée dans le détail dans l'annexe II-1.

Dans un premier temps, les 37 radiosondages ont donc été considérés comme un « paquet » de 37 variables dont nous avons recherché les facteurs dominants et statistiquement indépendants grâce à une Analyse en Composantes Principales classique.

Ce faisant, on risquait de considérer comme dominant un facteur sensible à un effet de taille: si l'on met dans le paquet plusieurs fois la même variable, la technique va détecter cet ensemble de variables comme un facteur très dominant dans le paquet alors que c'est tout à fait artificiel. C'est d'ailleurs ce qui se produit quand les variables sont des stations de mesure. Il suffit alors de mettre beaucoup de stations dans une zone pour que cet ensemble de variables soit considéré comme facteur dominant du champ étudié alors que ce n'est dû qu'à une hétérogénéité locale dans la densité de stations.

Il apparaît donc nécessaire de prendre en compte cette densité d'échantillonnage dans l'Analyse en Composantes Principales d'un champ aléatoire qui est intrinsèquement continu dans l'espace. Cet effet serait moindre si l'échantillonnage était homogène en densité (à pas constant) mais ce n'est pas le cas dans notre réseau de radiosondage. En effet, ce dernier donne, par exemple, beaucoup trop de poids au centre de l'Europe et pas assez au-dessus de l'Atlantique, même si sur mer la corrélation d'un point à l'autre est supérieure (cf. fig. I-1). Il va donc falloir en quelque sorte "réhomogénéiser" le réseau.

Nous allons donc utiliser l'ACP de Processus qui est l'équivalent de l'ACP classique pour des champs continus.

II.1.3 L'Analyse en Composantes Principales de Processus ou ACPP

Pour traiter des variables continues dans l'espace et/ou dans le temps, une approche de type processus aléatoire comme l'Analyse en Composantes Principales de Processus (ACPP), utilisant une formulation intégrale, est mieux adaptée, même si, dans la pratique, on est obligé de discrétiser le problème pour effectuer les calculs.

En outre, elle possède l'avantage de prendre explicitement en compte:

- la forme du domaine étudié,
- la typologie du réseau de mesure par l'intermédiaire de pondérations différentes si les stations ou points de mesure sont irrégulièrement répartis.

II.1.3.a Formulation de l'ACPP (Obled, 1979 ; Braud, 1990)

Soit $X(\xi,\underline{x})$ un processus aléatoire connu sur un domaine borné D.

Supposons-le centré ($E[X(\xi,\underline{x})] = 0$) et de variance $E[X^2(\xi,\underline{x})]$ finie afin que la fonction de covariance $C(\underline{x},\underline{x}') = E[X(\xi,\underline{x}) \cdot X(\xi,\underline{x}')]$ existe.

Soit $\underline{x}(x,y)$ un vecteur des coordonnées de l'espace et ### un processus aléatoire.

L'ACPP cherche à décomposer le processus de la manière suivante:

$$X(\xi,\underline{x}) = \sum_{k=1}^{\infty} Z_k(\xi) \cdot F_k(\underline{x}) \qquad \text{(II-2)}$$

où ξ est l'index de la réalisation considérée,

$\underline{\mathbf{x}}$ représente les coordonnées spatiales du point courant,

$Z_k(\xi)$ sont les coefficients aléatoires,

$F_k(\underline{x})$ sont les composantes spatiales déterministes ou EOF (Empirical Orthogonal Functions).

La détermination des Z_k et F_k se fait grâce à la méthode des moindres carrés: on cherche à minimiser

$$E\left[\int_D \left(X(\xi,\underline{x}) - X_k(\xi,\underline{x}) \right)^2 d\underline{x} \right] \qquad \text{(II-3)}$$

où $X_k(\xi,\underline{x}) = \sum_{k=1}^{K} Z_k(\xi) \, . \, F_k(\underline{x})$ est la décomposition limitée à K termes du processus.

Et l'on voit bien que, au niveau de l'intégrale (éq. II-3), cette optimisation est spécifique au domaine imposé D, d'où le terme *empirique* dans Empirical Orthogonal Functions, car le domaine est arbitraire.

De plus, pour assurer l'unicité de la solution, les fonctions $F_k(\underline{x})$ sont normées à 1:

$$\int_D F_k(\underline{x}) d\underline{x} = 1$$

Et finalement, on aboutit aux résultats suivants:

i) Les $F_k(\underline{x})$ sont des fonctions orthonormées sur D:

$$\int_D F_k(\underline{x}) \, . \, F_m(\underline{x}) \, d\underline{x} = \delta_{km} \qquad \text{où } \delta_{km} = 1 \text{ si } k = m,$$

$$= 0 \text{ sinon.}$$

Ce sont les fonctions propres de la matrice de covariance C intégrée sur D, associées aux valeurs propres λ_k:

$$\int_D C(\underline{x},\underline{x}').F_k(\underline{x}')d\underline{x}' = \lambda_k . F_k(\underline{x}) \qquad \text{(II-4)}$$

ii) les $Z_k(\xi)$ sont des variables aléatoires nommées Composantes Principales (CP) ou encore Fonctions Empiriques Orthogonales (EOF), décorrélées et de variance λ_k décroissante:

$$E\left[Z_k(\xi).Z_m(\xi)\right] = \lambda_k . \delta_{km}$$

obtenues par projection de $X(\xi,\underline{x})$ sur la kième fonction propre:

$$Z_k(\xi) = \int_D X(\xi,\underline{x}).F_k(\underline{x})d\underline{x} \qquad \text{(II-5)}$$

Nous avons donc formulé le problème théorique de manière intégrale. Les développements théoriques complets, esquissés par Obled (1979), ont été faits par Bouhaddou (1984) à une dimension et Braud (1990) à deux dimensions.

Cependant, pour calculer F_k et Z_k à l'aide des équations (II-4) et (II-5) il faut connaître X de façon continue sur l'ensemble du domaine D. Or, dans la pratique, X n'est connu que ponctuellement en P points qui sont les stations de mesure de ce domaine. Une résolution numérique, entraînant nécessairement une interpolation et donc une approximation, s'avère indispensable.

Nous utiliserons la méthode proposée à une dimension par Deville (1974) et généralisée à 2 dimensions par Obled (1979) et Obled et Creutin (1986) en s'inspirant des techniques d'éléments finis. Celle-ci est détaillée en annexe II-2. Seuls les résultats sous forme matricielle sont présentés ici.

Soit la base de fonction $e_i(\underline{x})$, pour i=1 à P, sur laquelle on interpole le processus X(###,$\underline{x}$). On obtient en discrétisant l'équation (II-4):

$$C.E.F = F.\Lambda \qquad \text{(II-6)}$$

avec **C** = (C_{ij}) i=1 à P, j=1 à P matrice carrée des covariances entre 2 points du réseau,

E = (E_{jm}) j=1 à P, m=1 à P matrice carrée des produits scalaires construits à partir des $e_i(x)$,

F = (F_{mk}) m=1 à P, k=1 à P matrice carrée des valeurs des fonctions propres aux points de mesure.

Λ matrice diagonale des valeurs propres $\boldsymbol{\lambda_k}$ pour k=1 à P.

L'équation (II-5) devient, pour l'estimation des $Z_k(\underline{\xi})$:

$$\mathbf{Z = X.E.F} \qquad \text{(II-7)}$$

où **X** = $X(\xi_i,\underline{x}_j)$ i=1 à N, j=1 à P matrice des réalisations de $X(\xi,x)$ (les N champs observés sur P stations),

et **Z** = $Z_k(\underline{\xi}_i)$ i=1 à N, k=1 à P matrice des Composantes Principales.

En résumé, on voit par l'intermédiaire des équations (II-6) et (II-7) que la seule différence avec l'ACP classique (annexe II-1) se situe dans l'introduction d'une matrice E contenant les produits scalaires des fonctions de base, intégrés sur le domaine D.

II.1.3.b Choix des fonctions de base $e_i(x)$

A deux dimensions, elles sont le plus souvent de 2 types:

i) fonctions constantes par morceau (fig. II-1a)

$e_i(\underline{x})$ = 1 sur le polygone de Thiessen associé à la station i (tous les points du polygone sont plus proches de la station i que de toutes les autres stations),

= 0 ailleurs.

Dans ce cas, la matrice E des produits scalaires est diagonale et les termes diagonaux sont égaux aux surfaces des polygones, d'où l'interprétation pratique : on donne à chaque station un poids égal à sa surface d'influence.

ii) fonctions linéaires par facette (fig. II-1b)

$e_i(\underline{x})$ = 1 à la station i,

= 0 aux autres stations,

avec une décroissance linéaire sur les facettes reliant les différents sommets.

Une triangularisation du domaine D est donc nécessaire.

Le calcul des termes E_{ij} de la matrice des produits scalaires utilise des techniques d'éléments finis (Norrie & De Vries, 1973). Il a été décrit par Obled et Creutin (1986) puis repris dans la thèse de Braud (1990) que l'on pourra consulter pour le calcul détaillé de cette matrice E.

En résumé, si S_{imk} est l'aire du triangle (i,m,k) avec i < m < k, on ajoute

- $S_{imk}/6$ aux termes E_{ii}, E_m, E_{kk},
- $S_{imk}/12$ aux termes E_{ik}, E_{im}, E_{mk}.

Et l'on somme pour tous les triangles.

L'utilisation d'une fonction spline d'ordre supérieur (fig. II-1c) peut aussi être envisagée mais les calculs deviennent vite très importants et des problèmes d'incertitude apparaissent lors du choix des frontières d'intégration du domaine. Par conséquent, nous n'en parlerons pas plus.

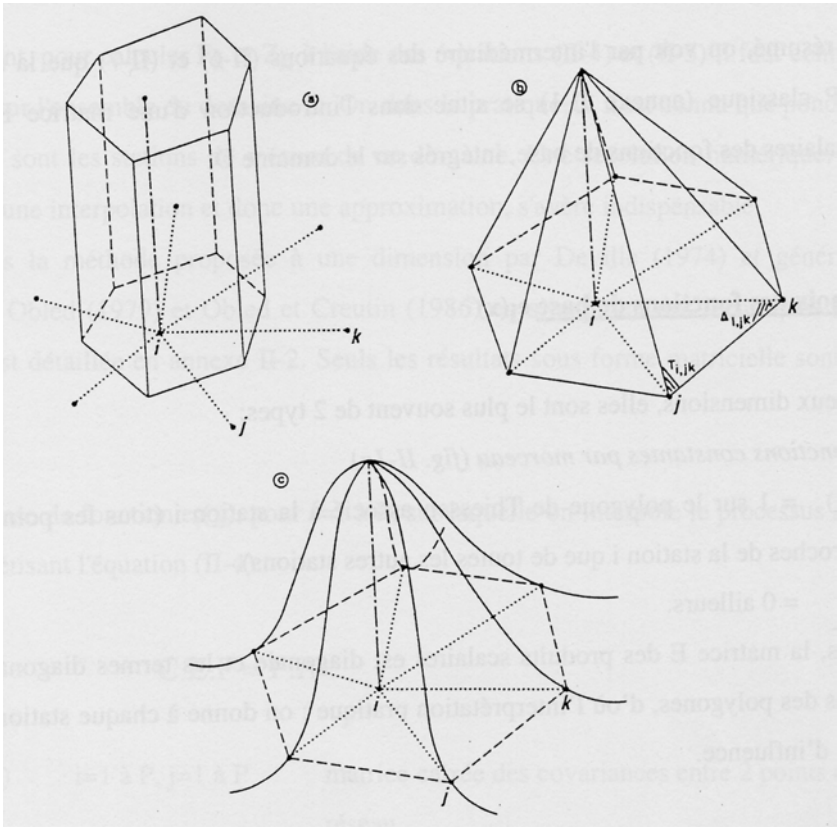


figure II-1: fonctions de bases typiques
a) constantes par morceau, b) linéaires par facette, c) fonctions splines
(tiré de Obled & Creutin, 1986)

Pour finir, il faut noter que la frontière du domaine intervient de façon déterminante dans le calcul des intégrales. Or le choix de cette frontière peut être arbitraire. Ce n'est pas toujours quelque chose d'intrinsèque au phénomène, elle n'est donc pas unique. De plus, sur le plan numérique, il n'existe pas non plus une seule triangularisation possible d'un domaine. Par conséquent, nous avons dû faire des choix.

II.1.3.c Résultats

Nous avons effectué une Analyse en Composantes Principales de Processus sur les 3 saisons et pour les 3 processus dont nous disposions, à savoir: les champs de géopotentiels 700 et 1000 hPa et l'épaisseur de la couche 700/1000 hPa pour l'été, l'automne et l'hiver, sur la période 1953-1993.

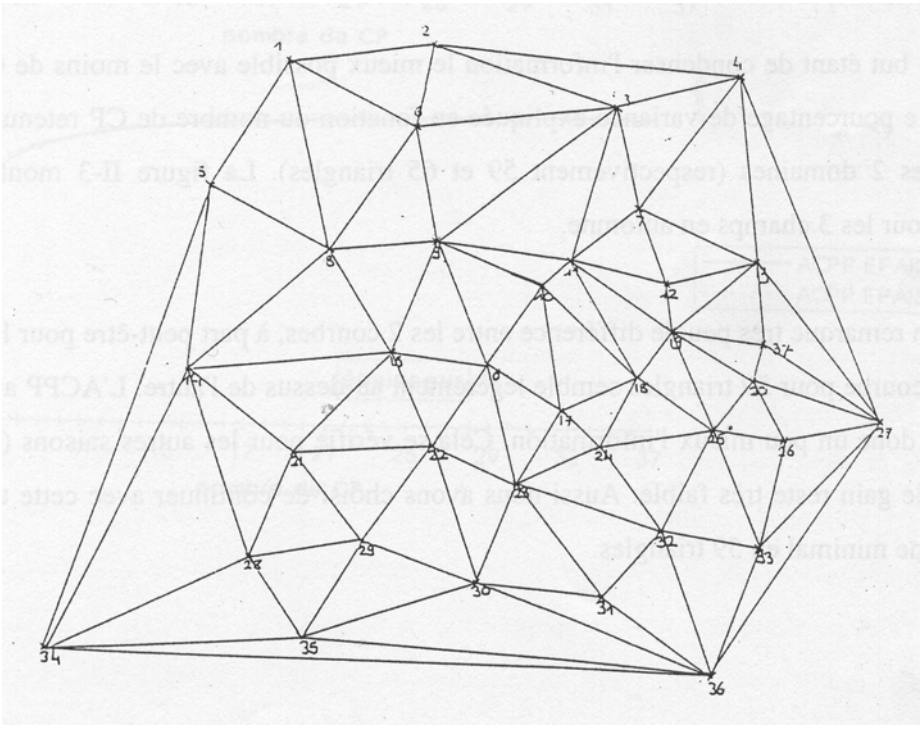
Notre choix de fonctions de base s'est porté sur les fonctions linéaires par facette, ceci pour deux raisons. Tout d'abord, il nous a semblé que les résultats obtenus par ces fonctions étaient

meilleurs en terme de variance expliquée pour un même nombre de CP retenues (Obled et Creutin, 1986; Braud, 1990). De plus, pour un réseau irrégulier, la triangularisation est plus facile à mettre en oeuvre que le calcul des surfaces des polygones de Thiessen (fonctions constantes par morceau) qu'il faut le plus souvent faire à la main.

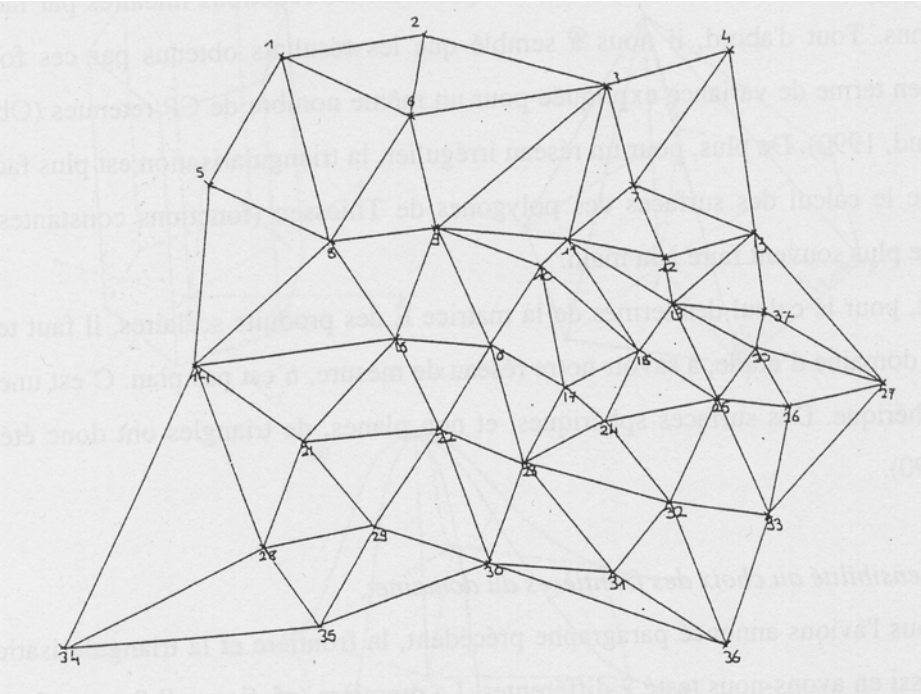
Cependant, pour le calcul des termes de la matrice E des produits scalaires, il faut tenir compte du fait que le domaine d'étude, à savoir notre réseau de mesure, n'est pas plan. C'est une portion d'une surface sphérique. Des surfaces sphériques, et non planes, de triangles ont donc été calculées (cf. Braud, 1990).

i) Sensibilité au choix des frontières du domaine:

Comme nous l'avions annoncé paragraphe précédent, la frontière et la triangularisation ne sont pas uniques aussi en avons-nous testé 2 différentes. La première (cf. figure II-2a) représente le domaine convexe avec 65 triangles alors que la deuxième (figure II-2b) se limite à un domaine minimal connexe avec 59 triangles seulement.



a) domaine convexe - 65 triangles



b) domaine connexe - 59 triangles

figure II-2: 2 triangularisations différentes du domaine

Le but étant de condenser l'information le mieux possible avec le moins de CP, nous avons comparé le pourcentage de variance expliquée en fonction du nombre de CP retenues pour l'ACPP et pour les 2 domaines (respectivement 59 et 65 triangles). La figure II-3 montre les résultats obtenus pour les 3 champs en automne.

On remarque très peu de différence entre les 2 courbes, à part peut-être pour les 10 premières CP où la courbe pour 59 triangles semble légèrement au-dessus de l'autre. L'ACPP avec 59 triangles condense donc un peu mieux l'information. Cela se vérifie pour les autres saisons (cf. annexe II-3) même si le gain reste très faible. Aussi nous avons choisi de continuer avec cette triangularisation du domaine minimal en 59 triangles.

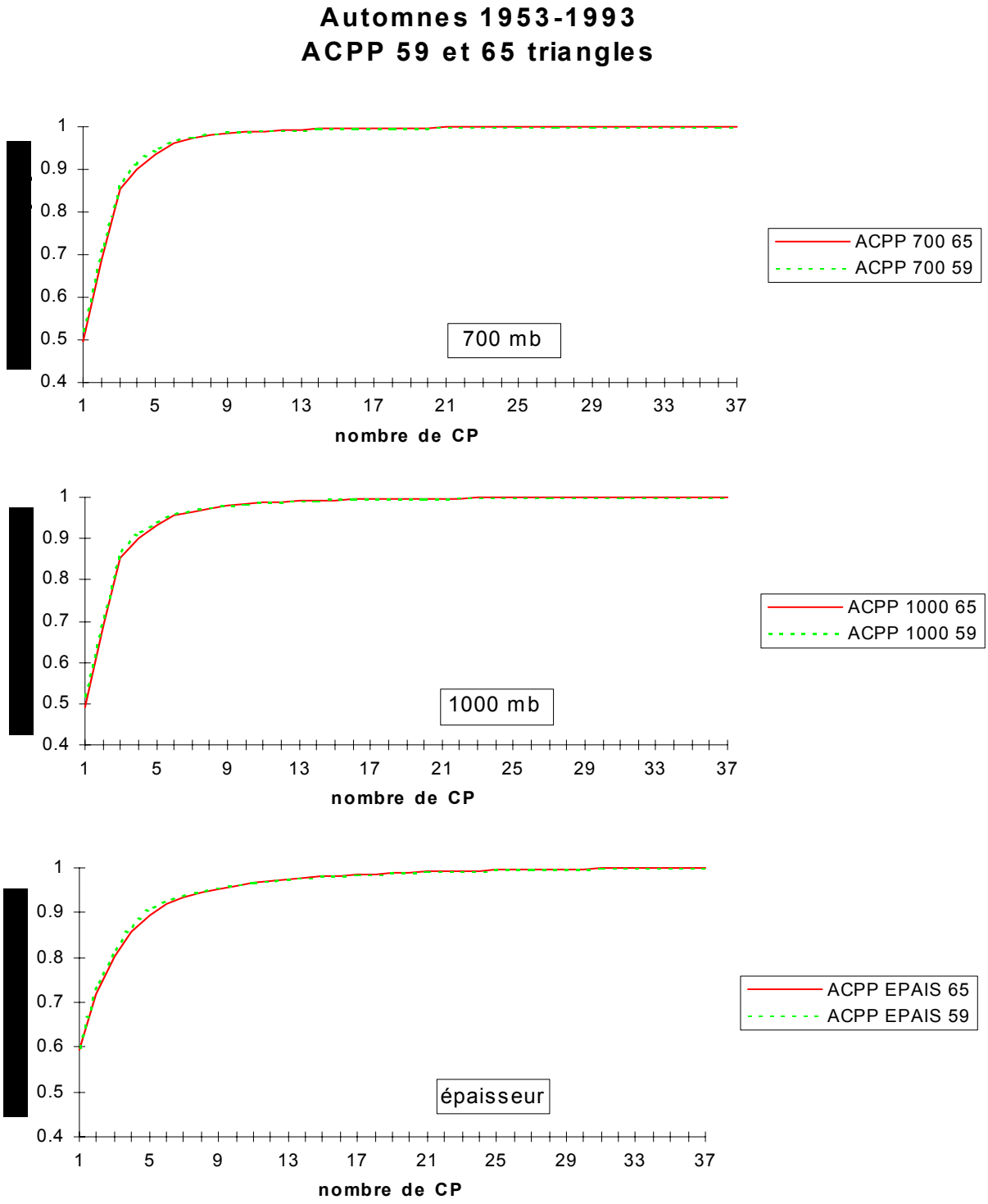


figure II-3: % de variance expliquée en fonction du nombre de CP obtenues par ACPP Comparaison des 2 domaines - automne

ii) Comparaison avec l'ACP classique:

Pour chacun des 3 champs le pourcentage de variance expliquée a été tracé en fonction du nombre de CP calculées par ACP et ACPP. Sur la figure II-4 se trouvent les courbes pour l'automne et en annexe II-4 vous trouverez ceux pour l'été et l'hiver.

Un gain non négligeable de variance expliquée est à noter jusqu'à 15-20 CP lorsque l'on calcule les CP par ACP de Processus. En particulier, sont consignés dans le tableau II-1 les pourcentages de variance expliquée par ACP et ACPP pour le nombre de CP utilisées dans la sélection des analogues:

	6 CP du champ 700 hPa	6 CP du champ 1000 hPa	1 CP de l'épaisseur 700 / 1000 hPa
ACP	92.6	90.7	53.7
ACPP	96.7	96.1	60.3

tableau II-1: comparaison des % de variance expliquée ACP / ACPP, automnes 1953-1993

Donc l'ACPP permet une meilleure condensation de l'information puisque, pour un nombre de CP donné, le pourcentage de variance expliquée est supérieur.

iii) Le nombre de CP significatives

Lorsque l'on effectue une Analyse en Composantes Principales, qu'elle soit classique ou de Processus, le but est de condenser l'information en en perdant le moins possible. On cherche donc à expliquer le maximum de variance tout en ne conservant que K vecteurs propres (ou CP) significatifs avec K << P, où P est le nombre de variables de départ.

De nombreuses règles permettant la détermination du nombre K de CP significatives existent (cf. Obled, 1979). Nous n'allons pas les décrire ici, ceci n'étant pas le but de notre étude. Mais une méthode classique sera utilisée pour déterminer K: la méthode LEV (Log-Eigen-Value).
On trace le logarithme des valeurs propres $\#\#\#_k$ en fonction de leur rang k. Il apparaît en général plusieurs segments rectilignes avec des cassures nettes qui séparent le plus souvent des valeurs propres multiples. C'est à l'endroit de ces cassures que l'on peut déterminer un nombre significatif de CP à retenir.

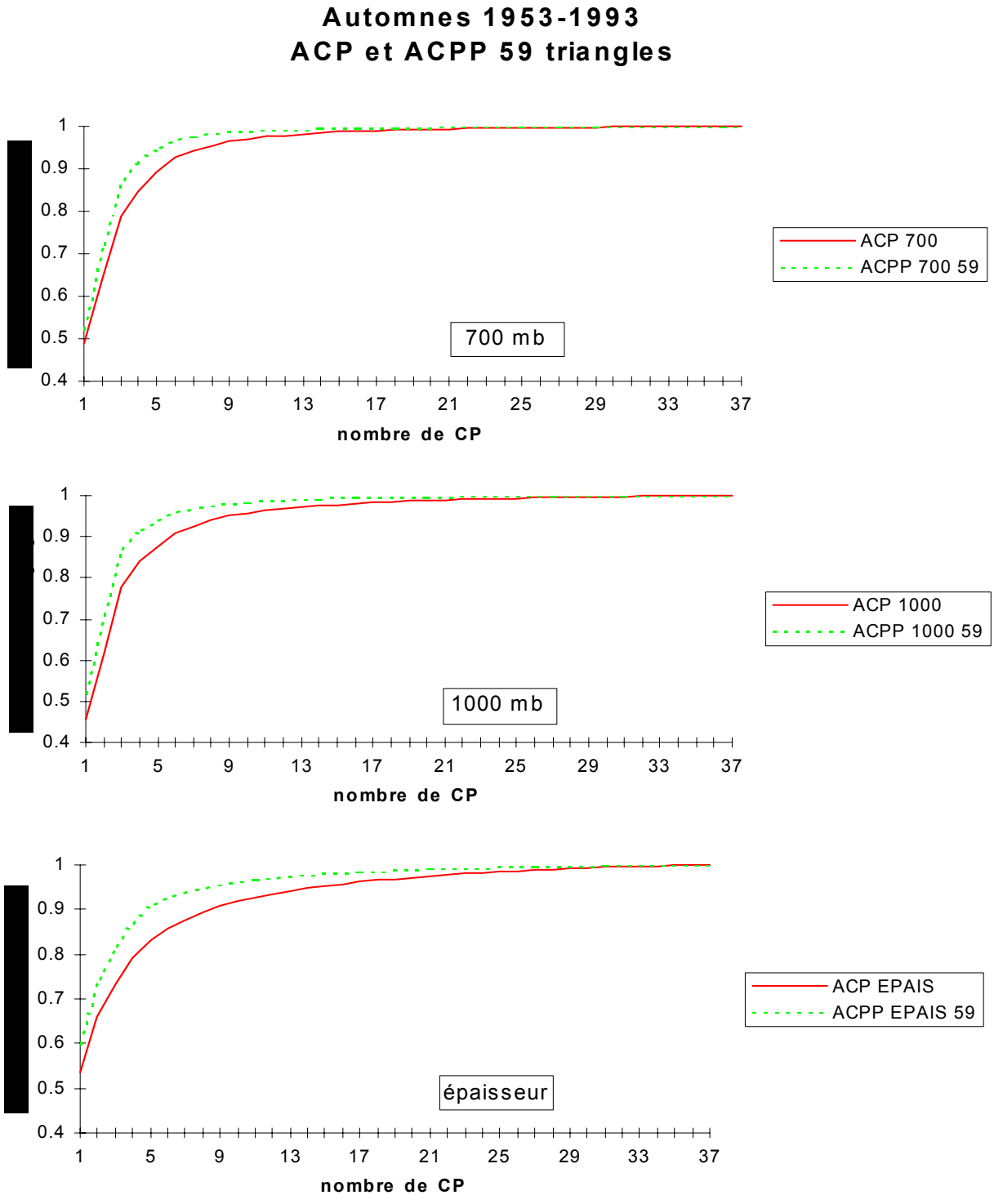


figure II-4: comparaison ACP et ACPP pour les 3 champs - automne

Dans notre cas, pour chaque champ nous possédons 37 stations de mesure soit 37 CP au maximum. Si l'on trace log ($\#\#\#_k$) = f (k) pour l'automne (figure II-5) différentes coupures sont possibles.

Pour les champs 700 et 1000 hPa, le fait de retenir 12 CP semble un bon compromis: le nombre de CP est assez faible et le pourcentage de variance expliquée est de 99% environ. Pour le champ de l'épaisseur la coupure se trouve plutôt à 14 CP mais afin d'harmoniser les 3 champs nous n'en prendrons que 12, ce qui permet d'expliquer plus de 97% de la variance.

Les résultats pour l'été et l'hiver sont donnés en annexe II-5.

II.1.3.d Interprétation dans l'espace des Composantes Principales

Plusieurs modes de représentation sont utilisés pour visualiser les résultats numériques et en faire une analyse. Dans notre cas, les vecteurs X_j de la matrice $[X_{np}]$ représentent des valeurs de niveau à 700 ou 1000 hPa en un point donné de l'espace. On peut donc se servir de la représentation cartographique des cosinus directeurs (lignes d'isovaleurs) sur un fond de carte où sont représentées les stations de mesure.

Les cartes ainsi obtenues donnent une interprétation spatiale des vecteurs propres étudiés qui peuvent éventuellement traduire des tendances physiques des niveaux retenus. Cette cartographie est une façon indirecte de reconnaître que l'on travaille sur un processus continu.

Les figures II-6 et II-7 donnent respectivement les cartes des 6 premiers cosinus directeurs calculés par ACP et ACPP du champ 700 hPa pour les automnes de 1953 à 1993 (autres champs en annexes II-6). Nous ne ferons ici que l'interprétation des cartes pour le champ 700 hPa, les autres champs amenant à des commentaires identiques.

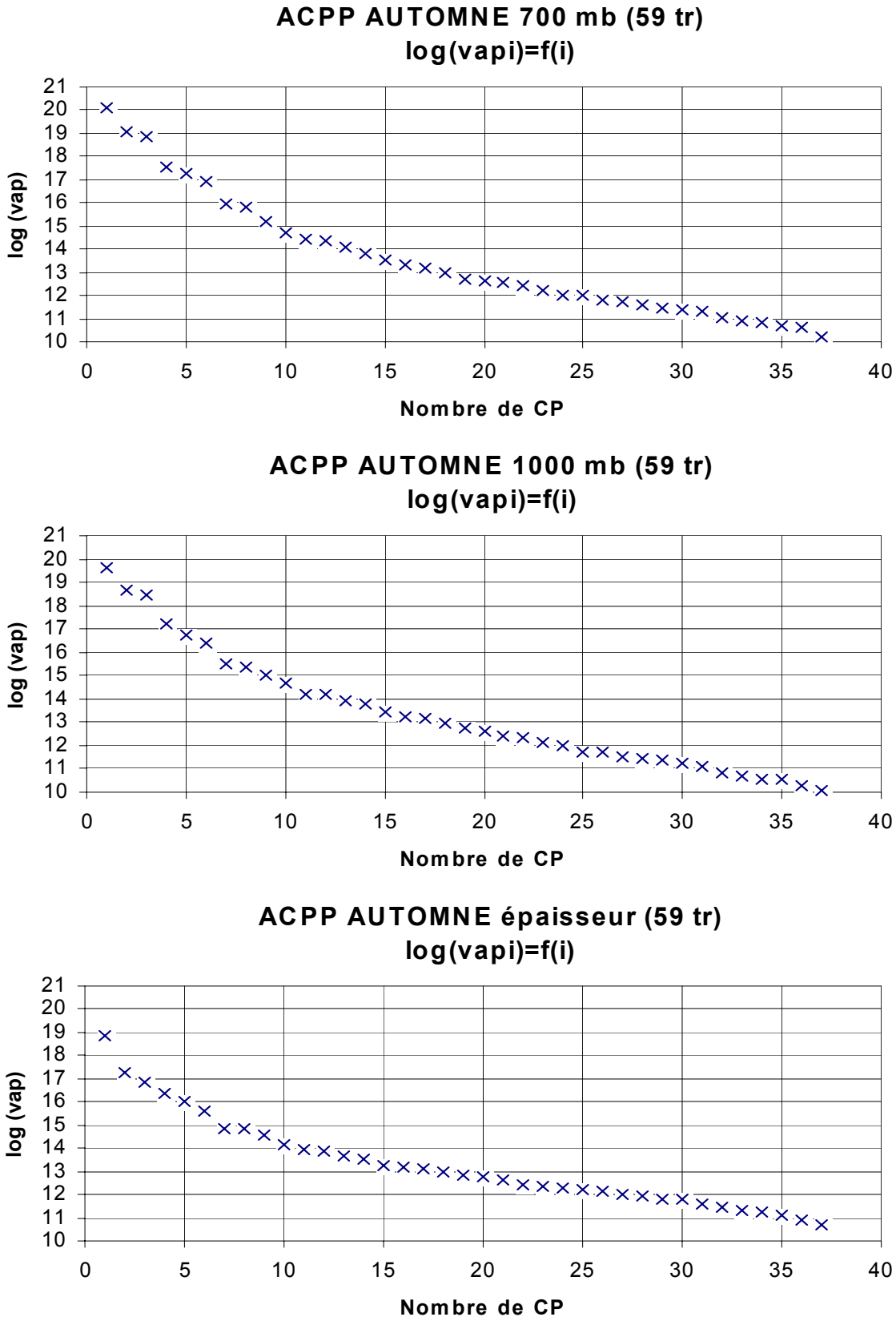


figure II-5: méthode LEV pour les 3 champs - automne

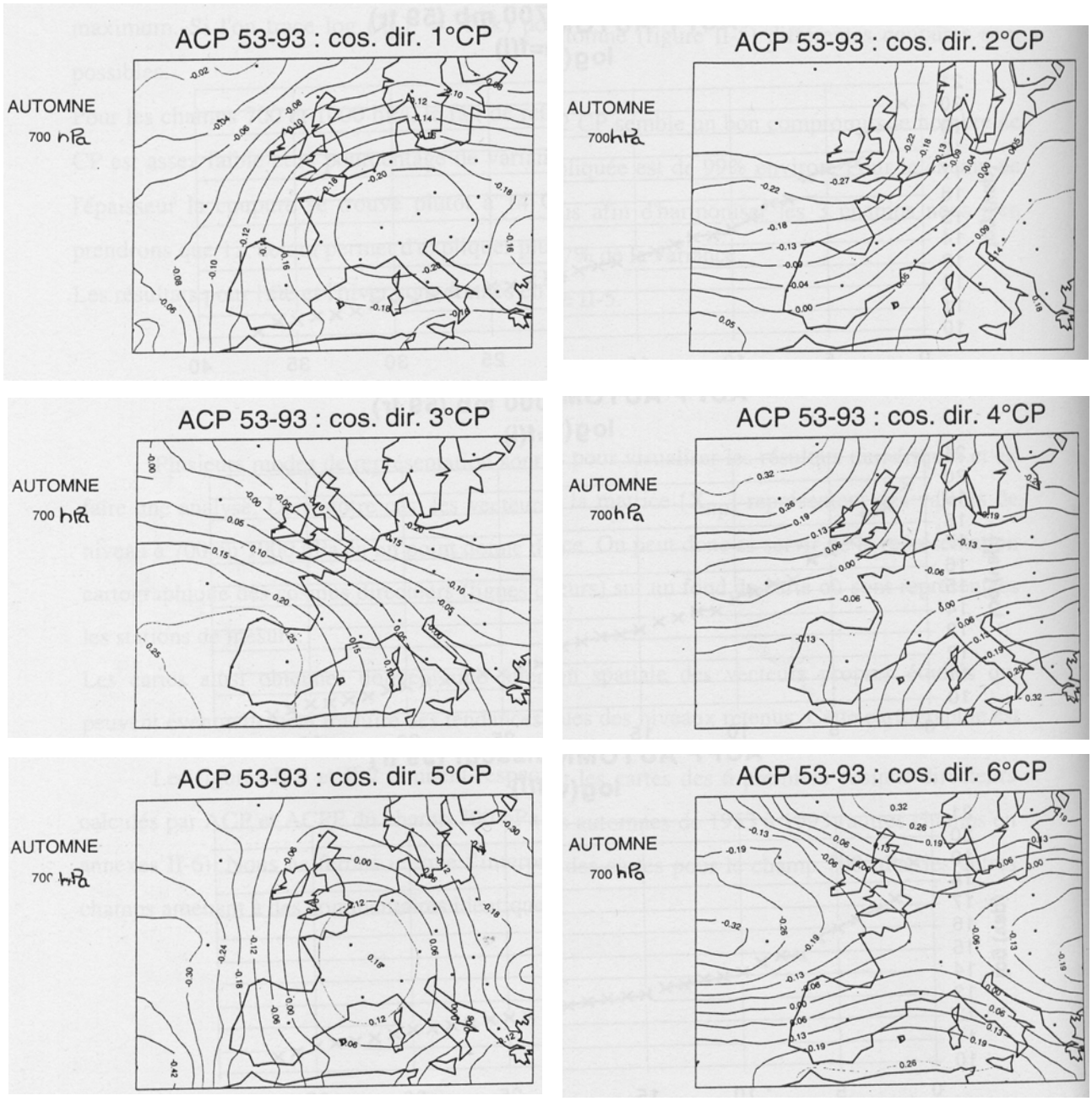


figure. II-6: carte des 6 premiers cosinus directeurs pour l'ACP

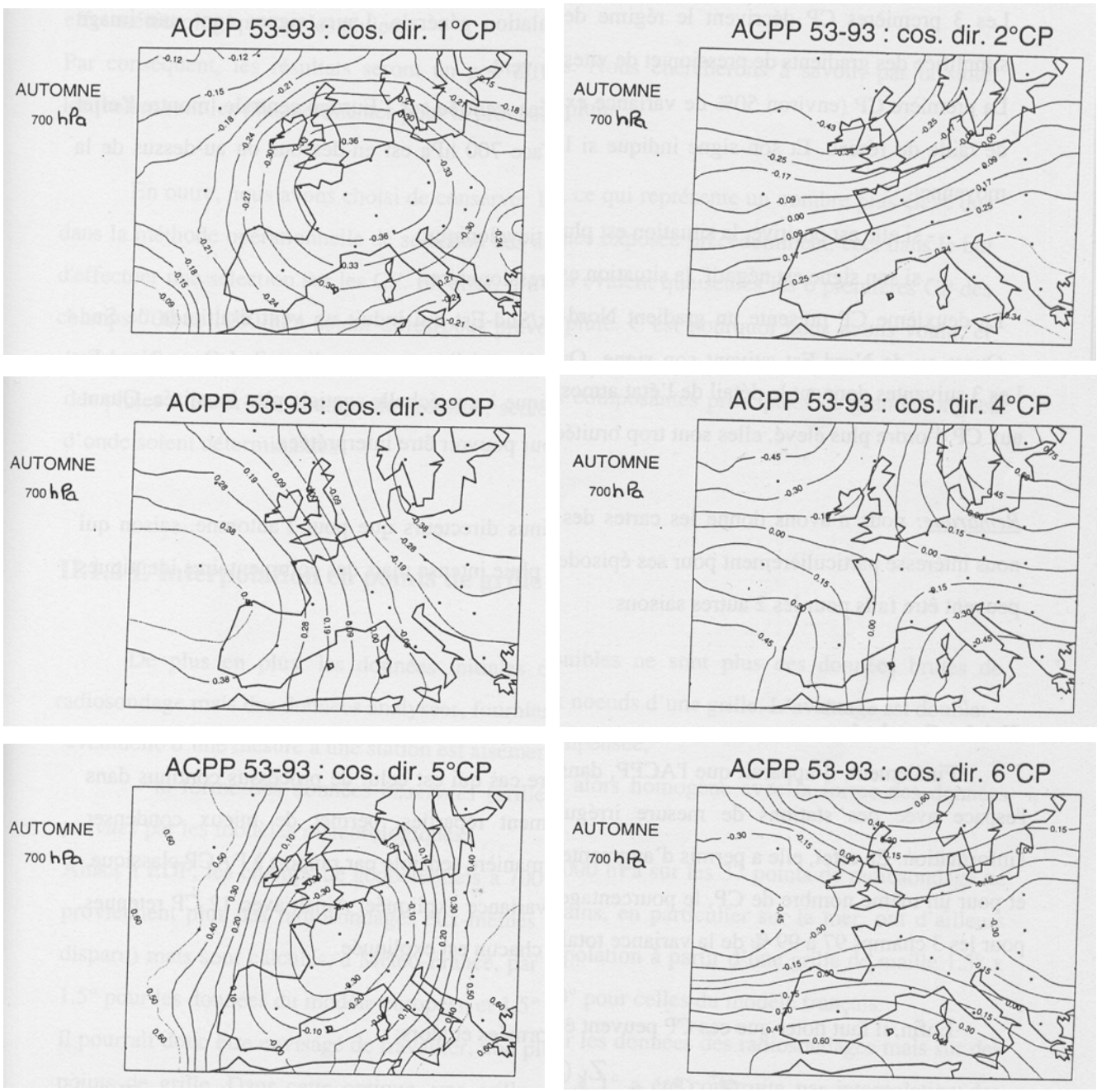


figure II-7: carte des 6 premiers cosinus directeurs pour l'ACP de Processus

En premier lieu, une grande similitude des tendances (signe et forme des isolignes) est à noter entre les cartes obtenues par ACP et ACPP.

Les 3 premières CP décrivent le régime de circulation générale. Leurs signes sont une image simplifiée des gradients de pression et de vitesses de vent.

La première CP (environ 50% de variance expliquée), centrée sur l'Europe centrale, montre l'effet de taille du réseau. Et son signe indique si la surface 700 hPa est en-dessous ou au-dessus de la moyenne:

- si elle est positive, la situation est plutôt anticyclonique,
- si son signe est négatif, la situation est dépressionnaire.

La deuxième CP présente un gradient Nord-Ouest/Sud-Est qui induit un vent d'altitude de Sud-Ouest ou de Nord-Est suivant son signe. Quant à la troisième, son gradient Sud-Ouest/Nord-Est induit un flux de Sud-Est ou de Nord-Ouest toujours suivant son signe.

Les 3 suivantes donnent le détail de l'état atmosphérique à une échelle spatiale plus localisée. Quant aux CP d'ordre plus élevé, elles sont trop bruitées pour pouvoir être interprétées.

Remarque: nous n'avons donné les cartes des cosinus directeurs que pour l'automne, saison qui nous intéresse particulièrement pour ses épisodes de pluie intense mais des commentaires identiques peuvent être faits pour les 2 autres saisons.

II.1.3.e Conclusion

Finalement, il apparaît que l'ACPP, dans notre cas qui est celui de processus continus dans l'espace avec des stations de mesure irrégulièrement réparties, permet de mieux condenser l'information. En effet, elle a permis d'augmenter de manière sensible par rapport à l'ACP classique, et pour un même nombre de CP, le pourcentage de variance expliquée. Ainsi, avec 12 CP retenues pour les 3 champs 97 à 99 % de la variance totale de chacun est expliquée.

Enfin, il faut noter que ces CP peuvent être normées et alors:

$$Z_k^*(\xi) = \frac{Z_k(\xi)}{\sqrt{\lambda_k}} \qquad \text{(II-8)}$$

Elles peuvent aussi être calculées avec la matrice de corrélation - on parlera alors d'ACPP de corrélation - au lieu de la matrice de covariance des données de départ (ACPP de covariance). Dans

ce cas, il suffit que les données de départ soient centrées mais aussi normées. Et cela entraîne une homogénéisation de la variance des champs. Or celle-ci variant selon les zones, cela peut avoir un effet météorologique sur les pluies.
Par conséquent, les résultats seront un peu différents. Nous chercherons à savoir, par la suite, lesquelles sont les plus pertinentes pour la prévision de pluie.

En outre, nous avons choisi de conserver 12 CP, ce qui représente un nombre plus élevé que dans la méthode opérationnelle de sélection des analogues exposée précédemment, ceci dans le but d'effectuer une sélection sur les CP. En effet, il n'est pas évident que seules les 6 premières CP des champs 700 et 1000 hPa soient utiles pour prévoir la pluie. C'est pourquoi nous avons voulu, et c'est ce que nous présenterons dans les prochains chapitres, effectuer une sélection sur ces CP elles-mêmes, CP qui serviront ensuite à extraire les analogues. Car il n'est pas sûr que dans la génération des pluies à partir des champs de pression, seules les composantes principales de grande longueur d'onde soient déterminantes.

II.1.4 L'interpolation en points de grille

De plus en plus, les données initiales disponibles ne sont plus des données brutes de radiosondage mais des données analysées, fournies aux noeuds d'une grille. L'avantage est double:

- à travers le processus d'analyse et d'assimilation, ces données sont critiquées et l'absence éventuelle d'une mesure à une station est aisément compensée,

- la forme des données *observées* (grille) est alors homogène avec la forme des données *prévues* par les modèles météorologiques.

Ainsi, à EDF, les champs de géopotentiels à 700 et 1000 hPa sur les 37 points de radiosondage ne proviennent plus des radiosondages eux-mêmes (certains, en particulier sur la mer, ont d'ailleurs disparu) mais sont calculés, à Météo-France, par interpolation à partir d'une grille de maille 1.5° x 1.5° pour les données du modèle européen et 1.5° x 2.0° pour celles du modèle français.
Il pourrait donc être envisagé de travailler, non plus sur les données des radiosondages mais sur des points de grille. Dans cette optique, une grille de 1° x 1° a été construite par interpolation des données de radiosondage grâce à une fonction spline. Elle comporte 19 x 22 soit 418 points de grille (cf. figure II-8).

C'est sans doute moins bien que des données de grille vraiment analysées, mais il n'en existe pas d'homogènes pour la période 1953-1993.

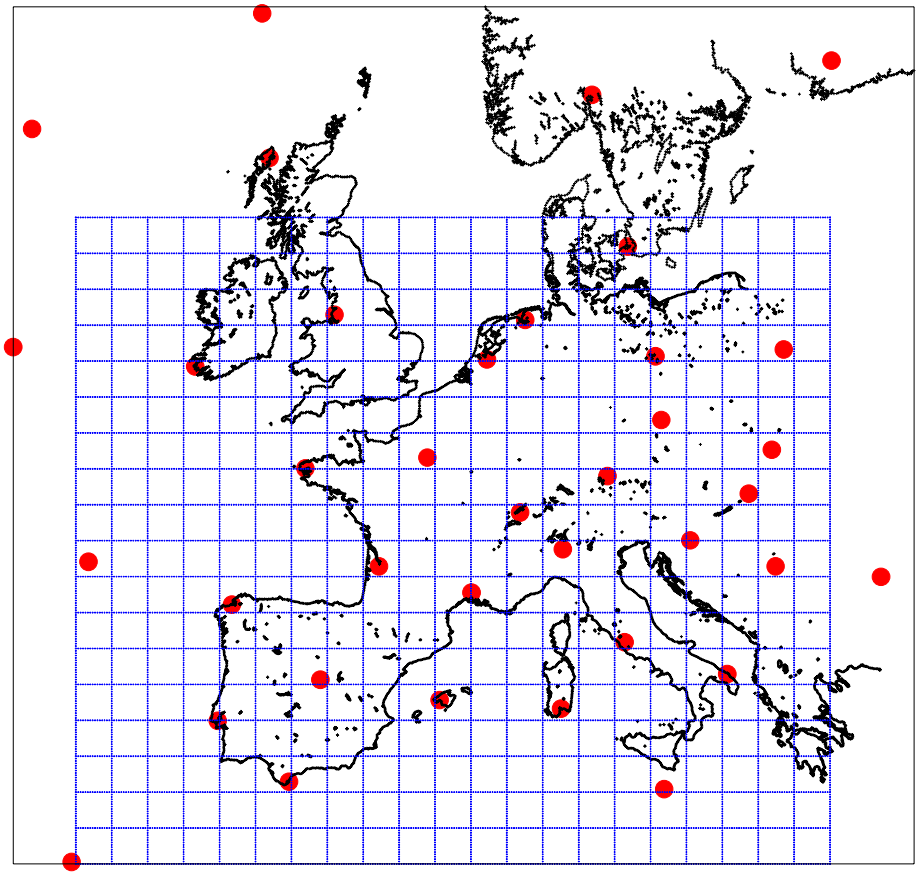


figure II-8: grille 1° x 1°

En conclusion, même s'ils sont issus essentiellement d'une seule source (les radiosondages), les prédicteurs sont donc disponibles sous 3 formes différentes pour la période 1953-1993:

- **les niveaux bruts** sur un réseau irrégulier de 37 stations,
- **les niveaux interpolés** aux noeuds d'un réseau régulier de 418 points,
- **les niveaux condensés** en 12 CP par une ACP de Processus.

II.2 Les prédictands

La variable à prévoir, le prédictand, est la lame d'eau attendue en 24 heures (de 07 h à 07 h TU) sur 33 bassins hydrologiques opérationnellement intéressants situés sur les massifs montagneux français.
Préoccupés surtout par le défi posé par les épisodes intenses d'automne, il nous est paru évident que leur prévision n'intéresse pas uniquement le Sud-Est de la France mais tout l'arc méditerranéen. C'est pourquoi, durant notre étude, quelques bassins frontaliers italiens (Ligurie et Piémont) et espagnols (Catalogne) ont été rajoutés.

Les raisons de cette introduction sont évidemment multiples. La première réside dans la forte similitude climatique de l'arc méditerranéen occidental. Ce sont souvent les mêmes systèmes qui balaient la côte catalane puis le littoral français jusqu'aux crêtes du Massif Central, la côte provençale et les Alpes du Sud puis la Riviera italienne. Inversement, en cas de retour d'Est, on a d'abord de fortes pluies sur le bassin du Pô, parfois suivies de débordements importants sur la zone alpine ou la Côte d'Azur. Il y a donc une forte similitude de préoccupation au niveau de ces régions, et plus particulièrement pour les épisodes d'automne.
Un autre aspect est lié à l'organisation administrative qui, en Italie du Nord par exemple, donne beaucoup d'autonomie mais aussi de responsabilité à la « Région », notamment pour la sécurité civile. Si, dans tous les cas, la prévision synoptique reste du ressort de l'Etat, la prévision hydrométéorologique est de celui des régions. Elle doivent donc se doter des moyens nécessaires et recherchent des méthodes adaptées à leurs besoins spécifiques, ici la prévision hydrométéorologique et l'alerte aux précipitations intenses.
A titre d'exemple, la région Ligurie a créé, auprès de l'université de Gênes, son propre centre de prévision hydrométéorologique !

II.2.1 Les bassins français

Le fichier de base des précipitations journalières est constitué des lames d'eau observées de 07 h à 07 h TU en 173 stations pluviométriques des Alpes, du Jura, du Massif Central, des

Cévennes et des Pyrénées. D'après ces données, un fichier de 33 groupements pluviométriques a été établi en calculant la moyenne arithmétique de 4 à 10 stations situées sur un même bassin versant, d'une surface pouvant aller de quelques centaines à quelques milliers de km^2. La position des groupements est donnée sur la figure I-2 du chapitre I.

Ces bassins, répartis dans les régions montagneuses de la moitié Sud de la France ont été choisis pour leur intérêt hydrologique dans le cadre de la production d'énergie hydroélectrique et/ou comme témoins de régions climatologiquement homogènes. Il ne s'agit donc pas forcément d'un bassin mais parfois de plusieurs situés dans des zones pluviométriquement homogènes sur lesquelles le mélange d'influences est minimum. Ainsi, en zone montagneuse, ces groupements ont souvent des formes allongées le long des reliefs.

Pour la zone française considérée, ce zonage a été proposé « manuellement » par Duband (1970), en s'aidant de la représentation des coefficients de corrélation des $2^{ème}$ et $3^{ème}$ Composantes Principales des pluies des 173 stations. Il a ensuite été confirmé sur les Cévennes à l'aide de techniques automatiques par Creutin et Obled (1980). Plus récemment, Champeaux et Tamburini (1996) ont étendu cette approche à l'ensemble du territoire métropolitain, mais à l'aide du seul réseau climatique d'état. Ils retrouvent néanmoins, dans la zone Sud-Est, des groupements d'assez petites tailles et présentant de fortes élongations le long des reliefs.

On notera que ces groupements homogènes en pluviométrie ne correspondent pas non plus à la structure maillée des modèles météorologiques. Cela pose éventuellement un problème d'interprétation de la pluie moyenne prévue sur une maille puisqu'une même maille de 30 x 30 km peut recouvrir partiellement 2 ou 3 zones homogènes.

En résumé, les données obtenues sont les lames d'eau journalières de 07 h à 07 h TU en 33 groupements pluviométriques pour la période 1953-1993. Cependant, la reconstitution de ces groupements est chaque année plus problématique car depuis 1986 les données manquantes sont de plus en plus nombreuses, certaines stations du réseau de Météo-France étant abandonnées. Jusqu'à présent la stratégie utilisée a été de remplacer la station manquante par une régression utilisant les stations encore disponible du groupement. Elle a été préférée à l'idée, pourtant plus facile à mettre en oeuvre, de calculer une moyenne avec les seuls témoins disponibles.
Cependant, une révision complète des groupements est envisagée dans un futur proche à EDF, afin de pallier le problème des stations abandonnées.

II.2.2 Les bassins catalans (Espagne)

Dans le cadre d'une action intégrée PICASSO entre le LTHE et l'équipe de Physique Atmosphérique de l'Université de Barcelone, cette dernière nous a fourni des données pluviométriques journalières sur la Catalogne provenant de l'Institut de la Météorologie Espagnole, dans le but de tester la méthode et ses tentatives d'améliorations sur quelques bassins catalans (Gibergans, 1995).

II.2.2.a Localisation de la Catalogne (Llasat, Puigcerver, 1992)

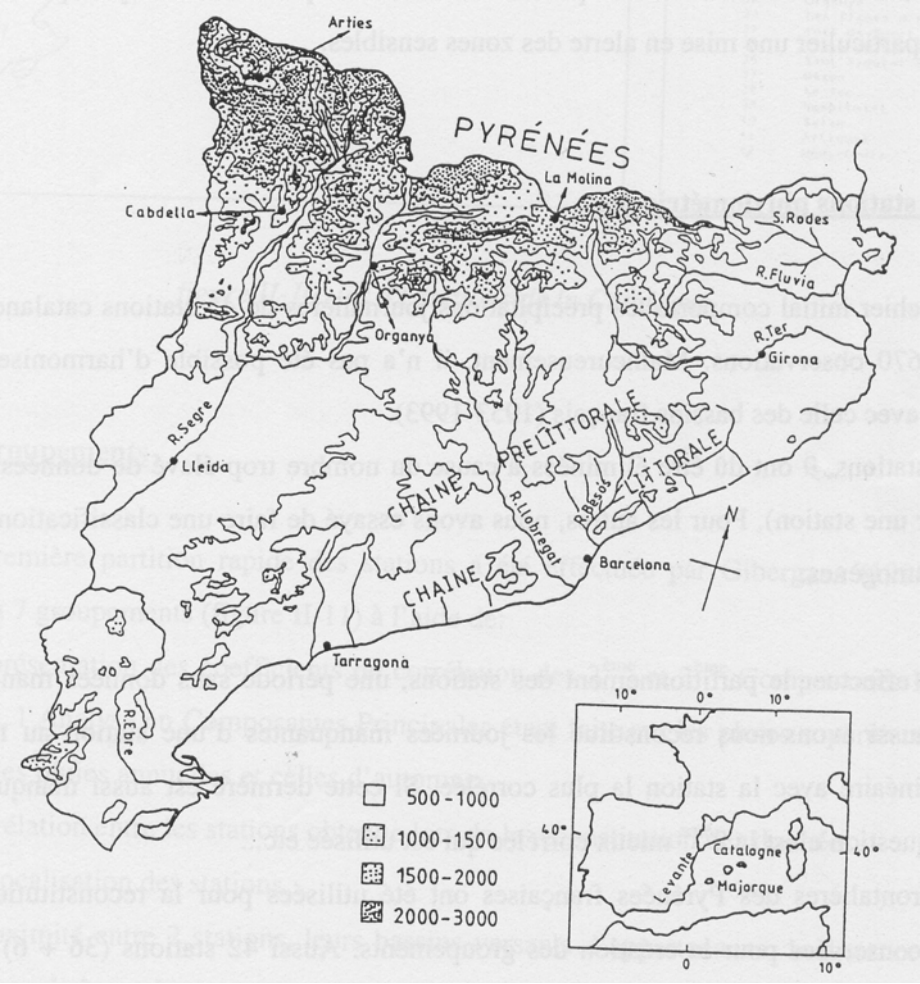


figure II-9: la Catalogne (tiré de Llasat & Puigcerver, 1992)

La Catalogne est située au Nord-Est de la Péninsule Ibérique et caractérisée par une orographie accidentée (cf. figure II-9), marquée principalement par trois chaînes de montagnes (littorale, prélittorale et pyrénéenne) et deux dépressions (littorale et centrale). Les chaînes littorale et prélittorale, ainsi que la côte, ont une orientation Sud-Ouest et Nord-Est. Et bien que le massif pyrénéen soit globalement orienté Ouest-Est, une grande partie de ses chaînes et de ses contreforts est parallèle à la côte méditerranéenne.

Ces facteurs géographiques sont déterminants pour le déclenchement et la distribution spatiale des pluies. Et de fortes précipitations convectives qui peuvent engendrer des crues catastrophiques se produisent relativement fréquemment dans cette région (en particulier sur la côte méditerranéenne) en automne à la tombée du jour (cf. Llasat et Puigcerver, 1992)

Aussi les Catalans sont-ils intéressés par une méthode de prévision des précipitations pouvant entraîner en particulier une mise en alerte des zones sensibles.

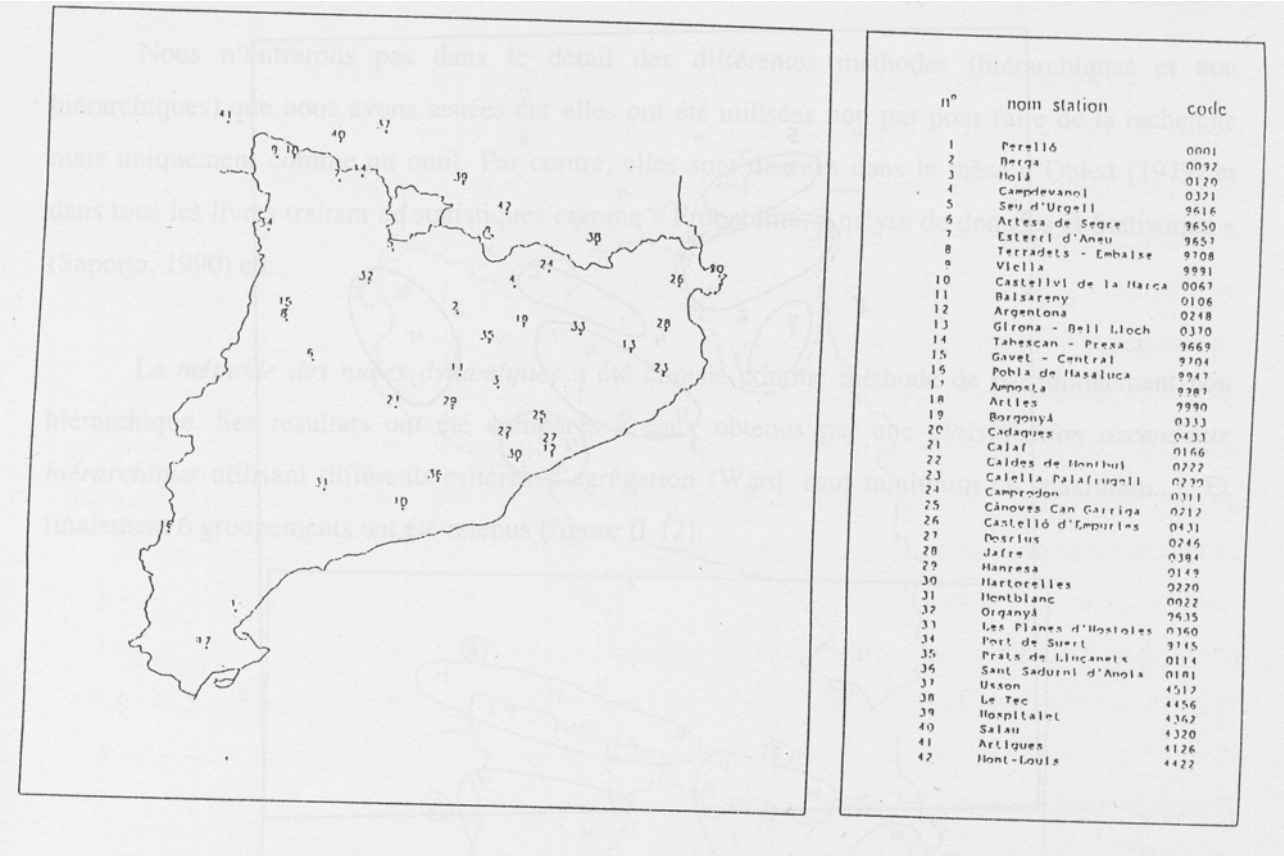
II.2.2.b Les stations pluviométriques

Le fichier initial comporte les précipitations journalières de 45 stations catalanes de 1970 à 1990 soit 7670 observations. Malheureusement, il n'a pas été possible d'harmoniser la période d'archivage avec celle des bassins français (1953-1993).

Sur ces 45 stations, 9 ont dû être éliminées à cause du nombre trop élevé de données manquantes (>4500 pour une station). Pour les autres, nous avons essayé de faire une classification des stations par zones homogènes.

Pour effectuer le partitionnement des stations, une période sans données manquantes était nécessaire aussi avons-nous reconstitué les journées manquantes d'une station au moyen d'une régression linéaire avec la station la plus corrélée. Si cette dernière est aussi manquante pour la journée en question c'est la $2^{ème}$ mieux corrélée qui est utilisée etc...

6 stations frontalières des Pyrénées françaises ont été utilisées pour la reconstitution. Elles ont ensuite été conservées pour la création des groupements. Aussi 42 stations (36 + 6) sont à notre disposition pour la période 1970-1990 (cf. figure II-10).



n°	nom station	code
1	Perelló	0001
2	Berga	0032
3	Olà	0120
4	Campdevanol	0321
5	Seu d'Urgell	9616
6	Artesa de Segre	9650
7	Esterri d'Aneu	9657
8	Terradets - Embalse	9708
9	Vielha	9991
10	Castellvi de la Marca	0067
11	Balsareny	0106
12	Argentona	0248
13	Girona - Bell Lloch	0370
14	Tabescan - Presa	9669
15	Gavet - Central	9704
16	Pobla de Hasaluca	9947
17	Amposta	9987
18	Arties	9990
19	Borgonyà	0333
20	Cadaques	0431
21	Calaf	0166
22	Caldes de Montbui	0222
23	Calella Palafrugell	0239
24	Campredon	0311
25	Cànoves Can Garriga	0212
26	Castelló d'Empuries	0431
27	Desrius	0246
28	Jafre	0384
29	Manresa	0149
30	Martorelles	0220
31	Montblanc	0022
32	Organyà	9635
33	Les Planes d'Hostoles	0360
34	Port de Suert	9745
35	Prats de Lluçanes	0114
36	Sant Sadurni d'Anoia	0181
37	Usson	4512
38	Le Tec	4456
39	Hospitalet	4362
40	Salau	4320
41	Artigues	4126
42	Mont-Louis	4422

figure II-10: les 42 stations de la Catalogne

II.2.2.c Les groupements

Une première partition rapide des stations a été effectuée par Gibergans (1995). Il a pu construire 6 ou 7 groupements (figure II-11) à l'aide de:

- la représentation des coefficients de corrélation des 2ème et 3ème Composantes Principales des 42 stations, l'Analyse en Composantes Principales étant faite sur les pluies supérieures à 20, 40 et 50 mm, sur les pluies annuelles et celles d'automne.

- la corrélation entre les stations obtenue lors de la reconstitution des données,

- ou la localisation des stations.

En effet, la proximité entre 2 stations, leurs bassins versants d'appartenance, la distance à la mer, l'orientation des chaînes de montagnes sont autant de facteurs méso-échelle décisifs pour la localisation et la distribution des pluies.

Les résultats détaillés de cette étude peuvent être consultés dans le rapport de Gibergans (1995).

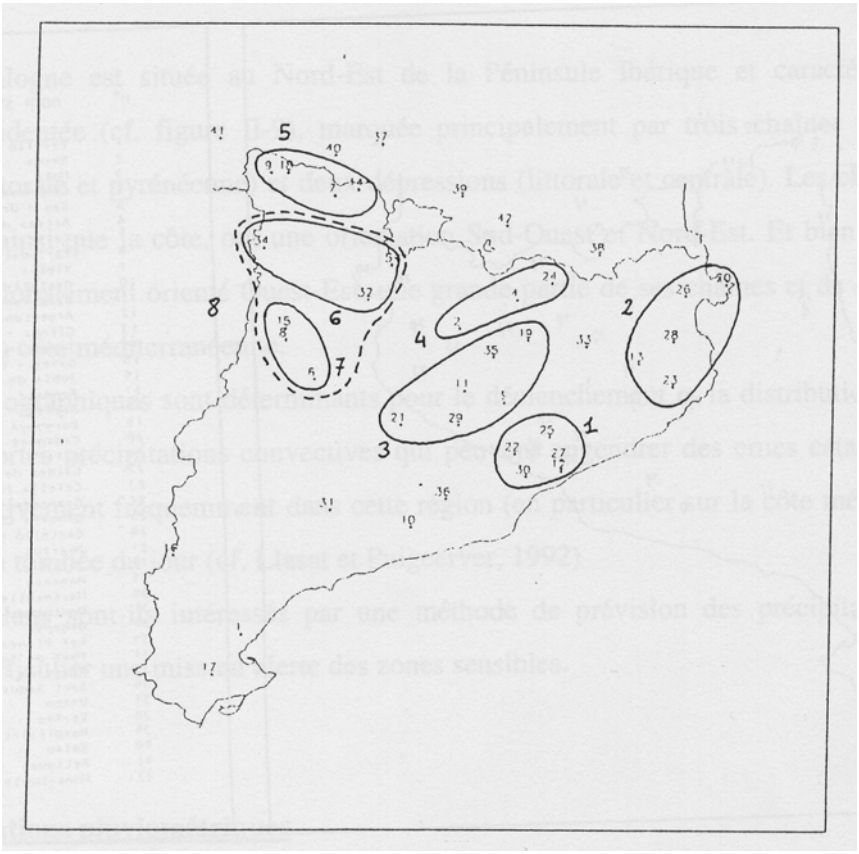


figure II-11: 1ère partition des stations catalanes (Gibergans, 1995)

Pour créer des groupements les plus robustes possibles dans le but de les introduire dans notre étude, nous avons aussi utilisé plusieurs méthodes de classification automatique, et donc non subjectives ou visuelles, tout en conservant en tête le fait qu'il doivent être construits par rapport à la variable d'intérêt: la pluie journalière.

Bref rappel sur les méthodes de classification automatique:

Elles ont pour but de regrouper les individus en un nombre restreint de classes homogènes et peuvent se répartir en 2 grands types de méthodes:

- les méthodes hiérarchiques qui produisent des suites de partitions en classes de plus en plus nombreuses, mais dont le principal problème réside dans le choix du critère d'agrégation entre 2 classes c'est-à-dire la *distance interclasse*,

- les méthodes non hiérarchiques qui produisent directement une partition en un nombre fixé de classes et donc qui supposent la définition d'un *critère global* mesurant la proximité des individus d'une même classe (distance intraclasse) et le choix d'un *nombre de classes* dès le début.

Nous n'entrerons pas dans le détail des différentes méthodes (hiérarchiques et non hiérarchiques) que nous avons testées car elles ont été utilisées non pas pour faire de la recherche mais uniquement comme un outil. Par contre, elles sont décrites dans la thèse d'Obled (1979) et dans tous les livres traitant de statistiques comme « Probabilité, Analyse de données et Statistique » (Saporta, 1990) etc...

La *méthode des nuées dynamiques* a été choisie comme méthode de partitionnement non hiérarchique. Ses résultats ont été comparés à ceux obtenus par une *classification ascendante hiérarchique* utilisant différents critères d'agrégation (Ward, saut minimum ou maximum...). Et finalement 6 groupements ont été retenus (figure II-12):

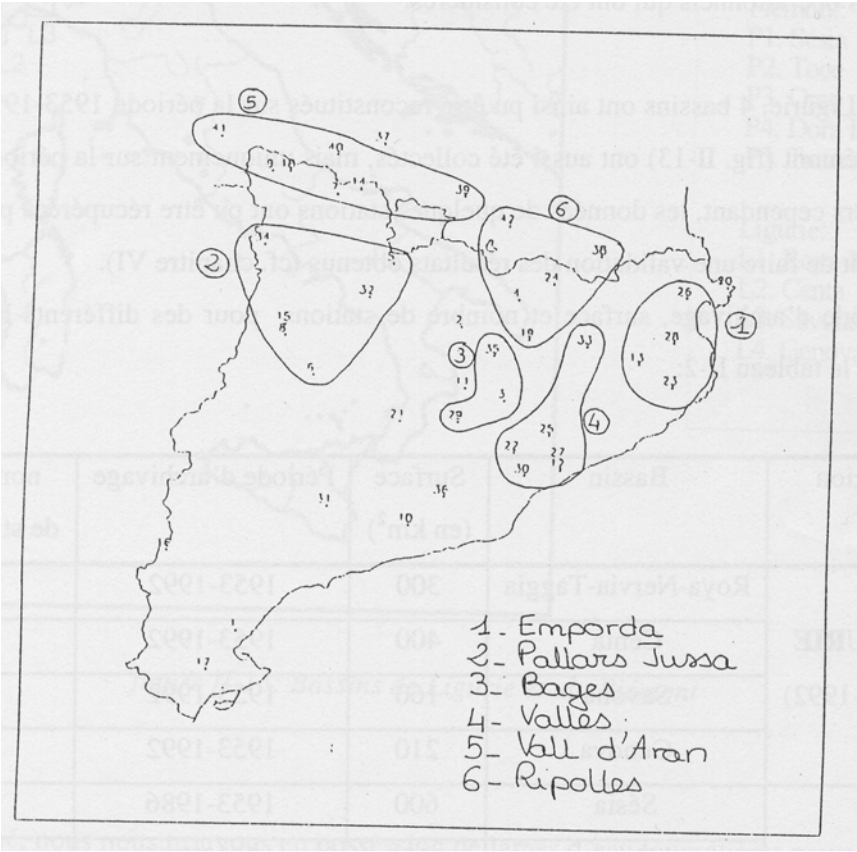


figure II-12: les groupements retenus par des méthodes de classification automatique

On remarque que les groupements 1, 2 et 8 de Gibergans correspondent pratiquement à des regroupements obtenus par classification automatique. Le groupement 5 a été un peu étendu. Par contre, les groupements 3 et 4 ont été remaniés. Ce sont ces derniers groupements qui ont été retenus.

II.2.3 Les bassins Italiens

L'Italie aussi est soumise à de forts épisodes pluvieux, en particulier sur les régions frontalières de la France. Et des partenaires italiens du Piémont (Regione Piemonte) et de la Ligurie (Université de Gênes) ont manifesté leur intérêt et souhaité, à titre expérimental, que cette méthodologie soit testée en différé sur des bassins qui les préoccupent.
Ainsi, des lames d'eau journalières en un certain nombre de stations pluviométriques ont été mises à notre disposition. Par contre, ces stations nous ont été données bassin par bassin. Il n'a donc pas été possible de faire un zonage préalable en régions pluviométriquement homogènes. Ce sont donc les bassins versants opérationnels qui ont été considérés.

Pour la Ligurie, 4 bassins ont ainsi pu être reconstitués sur la période 1953-1992 (fig. II-13). 5 bassins du Piémont (fig. II-13) ont aussi été collectés, mais uniquement sur la période 1953-1986. Pour ces derniers cependant, les données de quelques stations ont pu être récupérées pour l'automne 1994 dans le but de faire une validation des résultats obtenus (cf. chapitre VI).
Les nom, période d'archivage, surface et nombre de stations pour des différents bassins ont été consignés dans le tableau II-2:

Région	Bassin	Surface (en km^2)	Période d'archivage	nombre de stations
LIGURIE (1953-1992)	Roya-Nervia-Taggia	300	1953-1992	9
	Centa	400	1953-1992	6
	Savona	160	1953-1992	5
	Genova	210	1953-1992	8
PIEMONT (1953-1986)	Sésia	600	1953-1986	9
	Toce	1200	1953-1986	9
	Orco	800	1953-1986	4
	Dora Riparia	700	1953-1986	6
	Tanaro	1700	1953-1986	7

tableau II-2: les bassins du Piémont et de Ligurie

La lame d'eau pour un bassin considéré a été calculée à l'aide d'une moyenne arithmétique sur les stations disponibles d'un même bassin. Si pour un jour donné, une station est manquante, la moyenne est calculée sans cette station. Contrairement aux bassins français et espagnols les stations manquantes n'ont pas été reconstituées par les stations voisines car les corrélations étaient relativement médiocres.

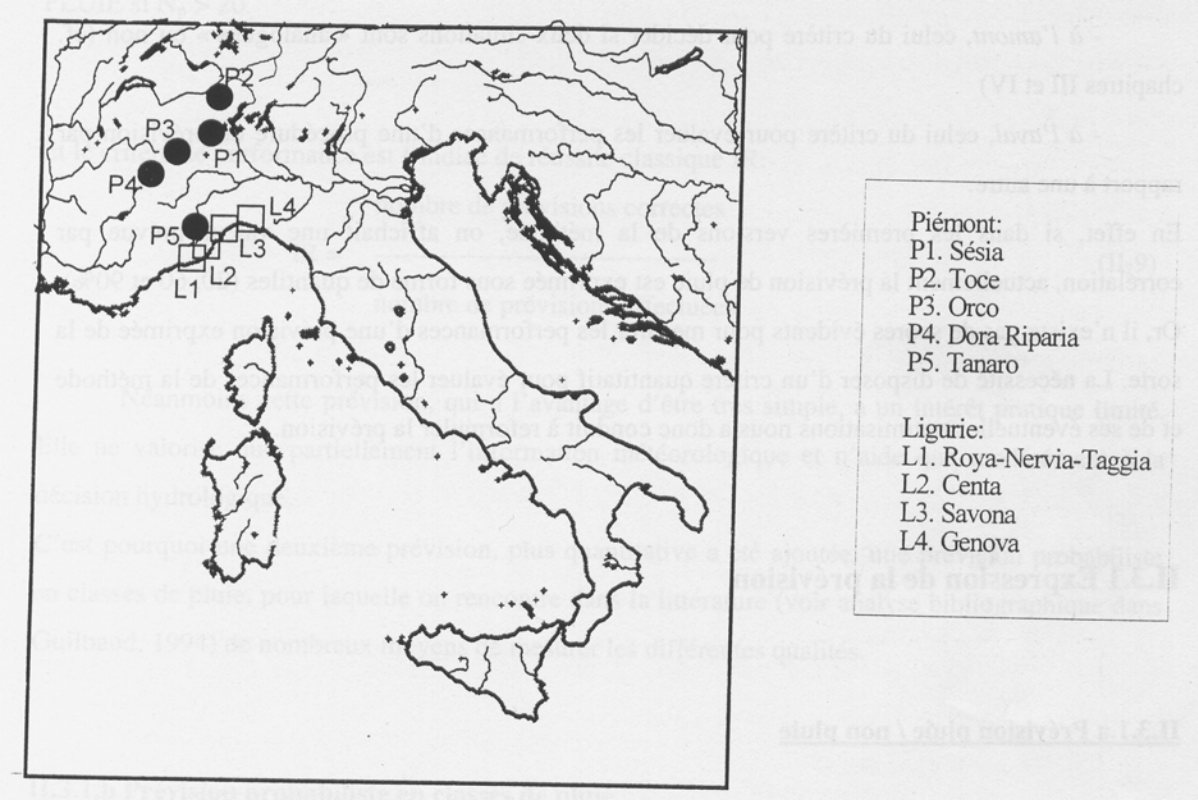


figure II-13: Bassins de Ligurie et du Piémont

En résumé, nous nous trouvons en possession de lames d'eau journalières pour:

- 33 groupements français pour la période 1953-1993,
- 6 groupements catalans pour la période 1970-1990,
- 5 bassins du Piémont 1953-1986,
- 4 bassins de Ligurie 1953-1992,

Seulement, avant d'entamer le chapitre des recherches, il a encore fallu résoudre un problème, celui de la mesure des performances de la méthode de prévision.

II.3 Expression et évaluation de la prévision

Si le choix des prédicteurs relève d'un arbitraire fortement contraint par la disponibilité des données brutes, deux autres choix se sont révélés cruciaux:

- *à l'amont*, celui du critère pour décider si deux situations sont « analogues » ou non (cf. chapitres III et IV)

- *à l'aval*, celui du critère pour évaluer les performances d'une procédure de prévision par rapport à une autre.

En effet, si dans les premières versions de la méthode, on affichait une valeur prévue par corrélation, actuellement la prévision de pluie est exprimée sous forme de quantiles (20, 60 et 90%). Or, il n'existe pas de scores évidents pour mesurer les performances d'une prévision exprimée de la sorte. La nécessité de disposer d'un critère quantitatif pour évaluer les performances de la méthode et de ses éventuelles optimisations nous a donc conduit à reformuler la prévision.

II.3.1 Expression de la prévision

II.3.1.a Prévision pluie / non pluie

La prévision la plus simple est la prévision catégorique en Pluie / Non Pluie, calculée de la manière suivante:

soit **N** analogues retenus pour la journée courante **C**, dont $\mathbf{N_p}$ avec pluie,

et **P** le pourcentage climatologique de jours sans pluie du bassin où l'on veut faire la prévision,

alors,

si $\mathbf{N_p} \leq N \times P$ prévision: NON PLUIE

si $\mathbf{N_p} > N \times P$ prévision: PLUIE

Le fait de prendre en compte la climatologie permet de privilégier un peu les pluies (et en particulier les fortes) qui sont moins fréquentes que les jours secs (cf. climatologie des différents groupements en annexe II-7).

Exemple: pour N = 50 analogues et pour un bassin avec 60 % de jours sans pluie, la prévision sera PLUIE si $N_p > 20$.

Et le critère de performance est l'indice de réussite classique IR:

$$IR = \frac{\text{nombre de prévisions correctes}}{\text{nombre de prévisions effectuées}} \qquad \text{(II-9)}$$

Néanmoins cette prévision, qui a l'avantage d'être très simple, a un intérêt pratique limité. Elle ne valorise que partiellement l'information météorologique et n'aide que modérément à la décision hydrologique.
C'est pourquoi une deuxième prévision, plus quantitative a été ajoutée: une prévision probabiliste en classes de pluie, pour laquelle on rencontre dans la littérature (voir analyse bibliographique dans Guilbaud, 1994) de nombreux moyens de mesurer les différentes qualités.

II.3.1.b Prévision probabiliste en classes de pluie

Soient N = 50 analogues sélectionnés pour le jour courant **C**. S'il y en a N_i dont la pluie est dans la classe i, la méthode prévoit comme probabilité d'être dans la classe i, le pourcentage d'analogues dans cette classe:

$$p_i = \frac{N_i}{N} \qquad \text{(II-10)}$$

Il a donc tout d'abord fallu déterminer le nombre de classes de pluie et leurs limites.
Pour les automnes de 1953 à 1993, environ la moitié des journées n'ont pas reçu de pluie et ceci sur l'ensemble des 33 bassins. De par là même, il devenait impossible de faire des classes équiprobables ou même d'égale amplitude et il s'imposait de faire une première classe de journées sans pluie.

Ensuite, nous avons essayé de faire des classes d'égal effectif pour les journées pluvieuses. Mais là encore nous avons dû y renoncer car dans ce cas, la dernière classe correspondait aux pluies supérieures à 15 mm dans le meilleur des cas. Or, comme nous voulons nous focaliser sur les épisodes pluvieux intenses, des classes avec plus de pluie sont nécessaires.
Et finalement, le découpage s'est fait en 8 classes (cf. figure II-14) avec un nombre de journées par classe diminuant progressivement (cf. annexe II-7). Le choix des limites de classes s'est fait en s'inspirant des échelles de couleurs des radaristes.

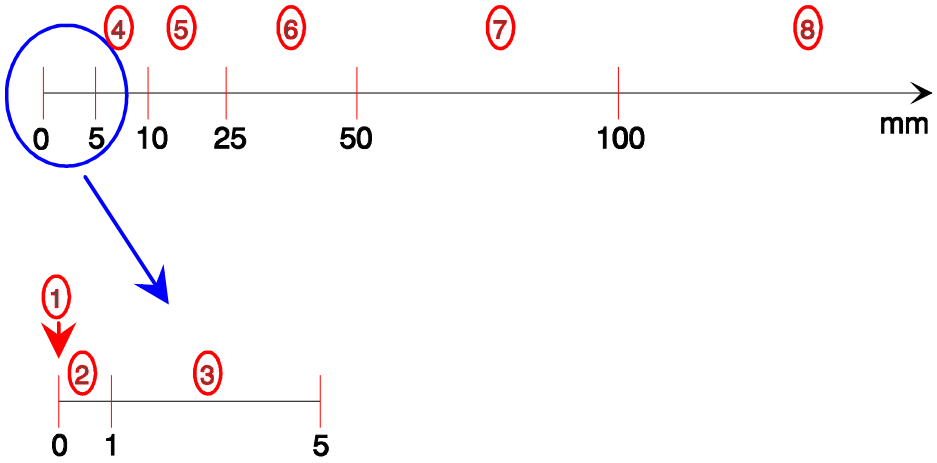


figure II-14: Les classes de pluie

Les 3 premières classes correspondent à des traces. Ensuite, le découpage essaie de correspondre à différents seuils de pluie et d'alerte :

- 5-10 mm pluie faible,
- 10-25 pluie modérée,
- 25-50 pluie forte (mise en vigilance),
- 50-100 pluie intense (mise en alerte),
- > 100 mm pluie extrême.

Idéalement, il aurait fallu choisir les classes par famille de bassin voire bassin par bassin mais nous avons préféré simplifier en prenant un compromis valable pour tous les bassins, ce qui explique en particulier la dernière classe (> 100 mm) surtout utile pour les bassins cévenols sujets à d'intenses épisodes pluvieux. Ce compromis, s'il paraît un peu simpliste, reste cependant acceptable si l'on intègre le fait que la pluie sur un bassin est déjà la moyenne des précipitations observées sur un nombre de stations qui diffère suivant les bassins.

II.3.2 Evaluation des performances d'une prévision probabiliste

Dans le cadre de mon DEA (Guilbaud, 1994), une analyse bibliographique a été effectuée. Elle passait en revue un certain nombre de scores mesurant les qualités d'une prévision probabiliste. Le plus connu est sans aucun doute le Score de Probabilité (PS) de Brier (Brier, 1950):

$$PS = \frac{1}{N} \sum_{n=1}^{N} \sum_{g=1}^{G} (p_{gn} - \delta_{gn})^2 \qquad \text{(II-11)}$$

avec **N** le nombre de prévisions effectuées,

G le nombre de classes,

$\mathbf{p_n} = (\mathbf{p_{1n}},....,\mathbf{p_{Gn}})$ où p_{gn} est la probabilité prévue d'être dans la classe g, lors de la $n^{ième}$ prévision,

$\boldsymbol{\delta_n} = (\boldsymbol{\delta_{1n}},....,\boldsymbol{\delta_{Gn}})$ où $\delta_{gn} = 1$ si la classe g est observée et 0 sinon, lors de la $n^{ième}$ prévision.

Ce score, compris entre 0 et 2, vaut:

- 0 si chaque prévision est catégorique et correcte ($p_{gn} = 1$ et classe g observée $\forall$ n),
- 2 si chaque prévision est catégorique et fausse ($p_{gn} = 1$ et classe $j \neq g$ observée $\forall$ n).

Par conséquent, contrairement au pourcentage de réussite, la prévision est d'autant meilleure que le score PS se rapproche de zéro.

Néanmoins, pour quantifier les performances d'une telle prévision, le score qui nous a semblé le plus approprié est le Ranked Probability Score (RPS) introduit par Epstein en 1969 (Epstein, 1969). Il mesure les performances d'une prévision probabiliste pour des *classes ordonnées* comme le sont les classes de pluie, c'est-à-dire qu'il considère que proposer la classe voisine de celle observée est moins pénalisant que de proposer des classes plus éloignées.

Ce score peut s'écrire de différentes façons, à des facteurs multiplicatifs près (Epstein, 1969; Murphy, 1970; Dequé *et al*., 1988).

Nous retiendrons l'expression la plus simple (Dequé *et al*., 1988) qui est l'erreur quadratique entre les vecteurs des probabilités cumulées observées et prévues. Avec les mêmes notations que pour le score de Brier, cela donne:

$$RPS = \frac{1}{N} \sum_{n=1}^{N} \sum_{g=1}^{G} \left[\sum_{k=1}^{g} (p_{kn} - \delta_{kn}) \right]^2 \qquad \text{(II-12)}$$

avec $0 \leq RPS \leq G-1$.

Et, comme pour le score de Brier PS, *plus il est proche de 0, meilleure est la prévision*.

Par rapport au score de Brier, ce score possède l'avantage d'introduire une notion de distance. Ainsi une erreur sur des classes voisines est moins grave que sur des classes éloignées: il tient compte du fait qu'on peut faire une prévision presque correcte en proposant la classe voisine de celle observée.

***Exemple* (Epstein, 1969) :**

Illustrons cette qualité par un petit exemple en considérant deux prévisions probabilistes de quatre classes de température: $T \leq 0°$, $0° < T \leq 20°$, $20° < T \leq 40°$ et $T \geq 40°$,

prévision n°1: (0.1 , 0.3 , 0.5 , 0.1),

prévision n°2: (0.5 , 0.3 , 0.1 , 0.1),

et l'on observe la classe 4: (0 , 0 , 0 , 1).

Intuitivement, la première prévision semble meilleure car la classe la plus probable est plus proche de la classe observée que dans la prévision n°2. Le RPS fera de même (RPS1 < RPS2). Par contre, le score de probabilité PS ne départagera pas les 2 prévisions (PS1 = PS2) car il prend juste en compte la probabilité prévue de la classe observée qui est la même dans les 2 cas.

II.4 Conclusion du chapitre II

Finalement, nous disposons:

- pour les prédicteurs, d'un fichier historique journalier de 1953-1993, contenant les champs de géopotentiels 700 et 1000 hPa et l'épaisseur de la couche 700/1000 hPa à 00 TU en 37 points de radiosondage couvrant l'Europe de l'ouest et l'Europe centrale. Ces données peuvent être utilisées soit (cf. fig. II-8):

* brutes,
* après ACP ou ACPP en 12 Composantes Principales par champ,
* après interpolation en 418 points d'une grille de pas 1° x 1°,

- pour les prédictands (cf. fig. II-15), d'un fichier historique contenant les lames d'eau journalières:

* en 33 groupements pluviométriques français, de 1953 à 1993,
* en 4 bassins de Ligurie (Italie), de 1953 à 1992,
* en 5 bassins de Piémont (Italie), de 1953 à 1986,
* en 6 groupements catalans (Espagne), de 1970 à 1990.

Quant à la prévision de pluie, exprimée dans un premier temps en quantiles, elle a dû être reformulée afin de disposer de critères permettant l'évaluation des performances de la méthode et de ses éventuelles optimisations. Ainsi, elle peut s'exprimer alternativement :

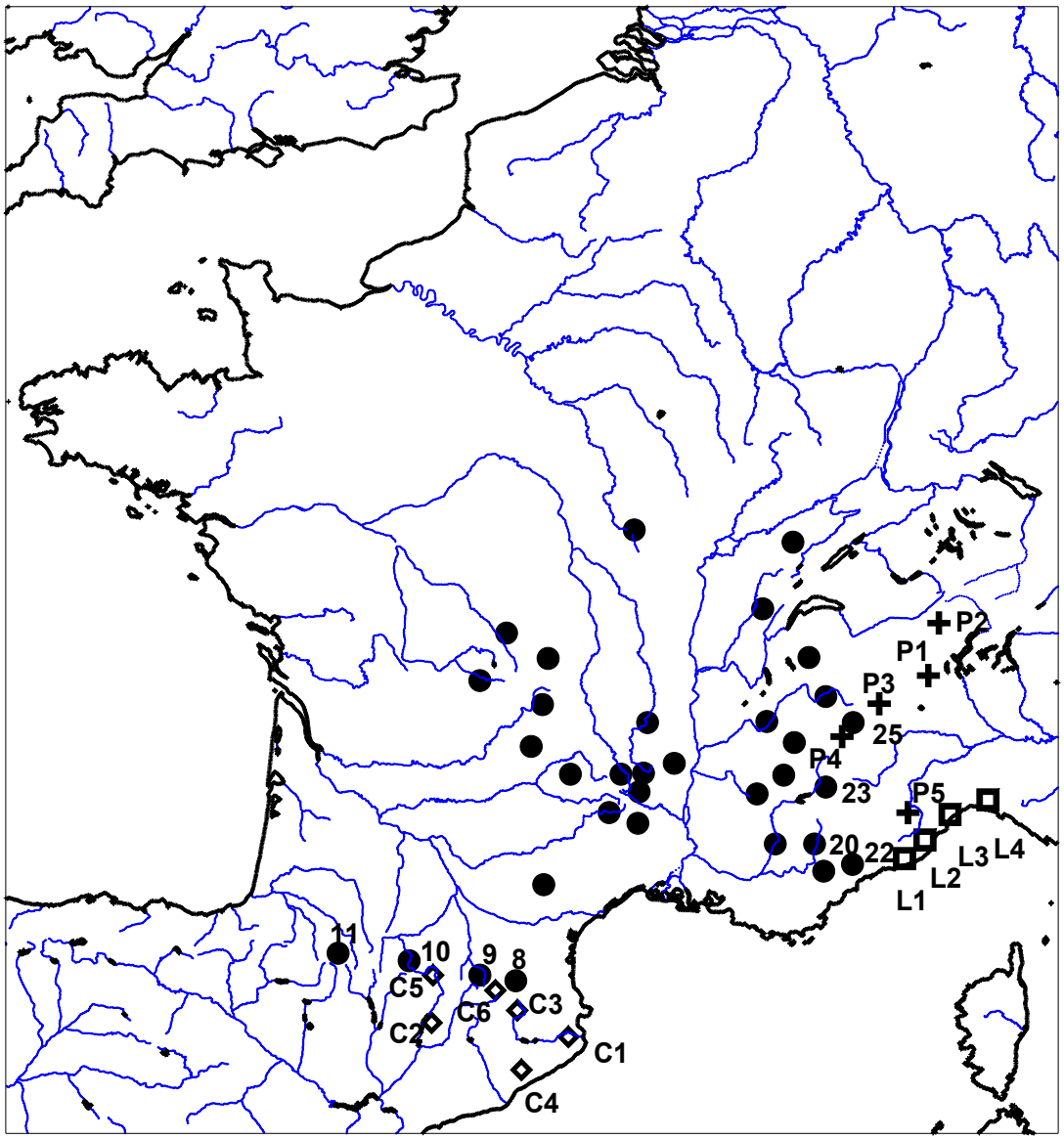
- *en quantiles* (20, 60 et 90 %), ce qui permet une représentation graphique et une évaluation qualitative sur quelques épisodes intéressants (cf. figure I-5),

- *en pluie / non pluie*, prévision qualitative avec un score très simple, le pourcentage de prévisions correctes,

- *en probabilités de classes de pluie*, prévision plus quantitative et donc plus intéressante, avec le Ranked Probability Score pour évaluer les performances des prévisions.

A partir de là, disposant d'un critère d'évaluation des performances, l'optimisation peut se faire soit en modifiant :

- le critère d'analogie (chapitres III et IV)
- la forme des prédicteurs (chapitres III et IV),
- ou encore en introduisant de nouveaux prédicteurs (chapitre V).



Catalogne:
C1. Emporda
C2. Pallars Jussa
C3. Bages
C4. Valles
C5. Vall d'Aran
C6. Ripolles

Ligurie:
L1. Roya-Nervia-Taggia
L2. Centa
L3. Savona
L4. Genova

Piémont:
P1. Sésia
P2. Toce
P3. Orco
P4. Dora Riparia
P5. Tanaro

figure II-15: les prédictands

CHAPITRE III :

METHODES FONDEES

sur les

COMPOSANTES PRINCIPALES

Introduction

Le nombre de variables utilisées pour décrire la situation météorologique peut d'emblée poser problème. Historiquement, le problème venait surtout des limitations en puissance de calcul. Mais plus théoriquement, il vient aussi des redondances importantes qui existent entre les variables prédicteurs utilisées.

Un exemple classique consiste à proposer d'une part le champ de température à 850 hPa et d'autre part l'épaisseur entre les niveaux 1000 et 700 hPa. Or on sait que cette dernière est fortement liée à la température moyenne de la couche: ces deux variables représentent donc à peu près le même phénomène et sont redondantes.
Un autre exemple concerne la prise en compte de plusieurs états antérieurs (par exemple les géopotentiels à J - 24 h ou J - 48 h) pour inclure dans la description de la situation un aspect évolutif. Malheureusement, ces données sont temporellement très corrélées et donc redondantes.
Un dernier exemple concerne la prise en compte de 2 stations de radiosondage voisines. C'est la redondance spatiale, qui a déjà été traitée par l'Analyse en Composantes Principales (ACP) mais qui pourrait l'être aussi par la sélection.

Ces redondances ne sont cependant pas forcément fâcheuses; néanmoins, elles le deviennent quand elles donnent à un phénomène unique un poids important justifié uniquement par la répétition arbitraire des variables qui le décrivent.
Il y a à cela deux corrections possibles:

- identifier les phénomènes réellement indépendants décrits par l'ensemble de variables. C'est l'orthogonalisation du jeu de données, effectuée par ACP ou ACP de Processus (ACPP) qui nous a permis d'extraire un sous-ensemble de 12 Composantes Principales (CP) par champ, statistiquement décorrélées.

- procéder par sélection de variables en demandant à un algorithme d'apprentissage de choisir les plus pertinentes pour expliquer la pluie (cf. chapitre IV).

Dans le cas de la première correction, comme l'orthogonalisation est intrinsèque au jeu de données, elle ne prend pas en compte la variable à prévoir. Il y a donc ensuite une sélection à faire

parmi ces variables indépendantes pour détecter les plus informatives vis à vis du prédictand, à savoir la quantité de pluie journalière sur les différents bassins disponibles.

Dans ce chapitre, nous allons donc mettre en oeuvre des méthodes de sélection et de pondération des CP à utiliser pour extraire les analogues. La méthodologie employée, les quelques modifications nécessaires pour mettre en place la sélection des CP, les différentes méthodes de sélection et de pondération seront décrites dans un premier temps, avant de donner, pour les 2 types de prévision retenus, des résultats par l'intermédiaire des critères de performance:

- *le Ranked Probability Score RPS* pour la prévision probabiliste en classes,
- *et l'indice de réussite IR* pour la prévision en pluie / non pluie.

Tout au long de notre étude, nous nous sommes focalisés sur le cas précis des **automnes** (période du 1/09 au 30/11), en particulier à cause des épisodes pluvieux intenses qui se produisent à cette époque. Aussi avons-nous travaillé uniquement sur cette saison.

III-1 Algorithmes

III.1.1 Méthodologie

III.1.1.a Validation croisée

Afin de tester les différentes évolutions de la méthode sur le plus grand nombre de journées possibles, nous avons utilisé la validation croisée sur la période 1953-1993 (cf. figure III-1):

- pour chaque journée test d'automne de 1953 à 1993 soit 3731 observations,
- N analogues sont extraits parmi toutes les journées d'automne de 1953 à 1993,
- en excluant la possibilité de tirer un analogue dans l'année auquel appartient le jour test.

Le critère de performance est ensuite calculé sur l'ensemble de ces 3731 prévisions.

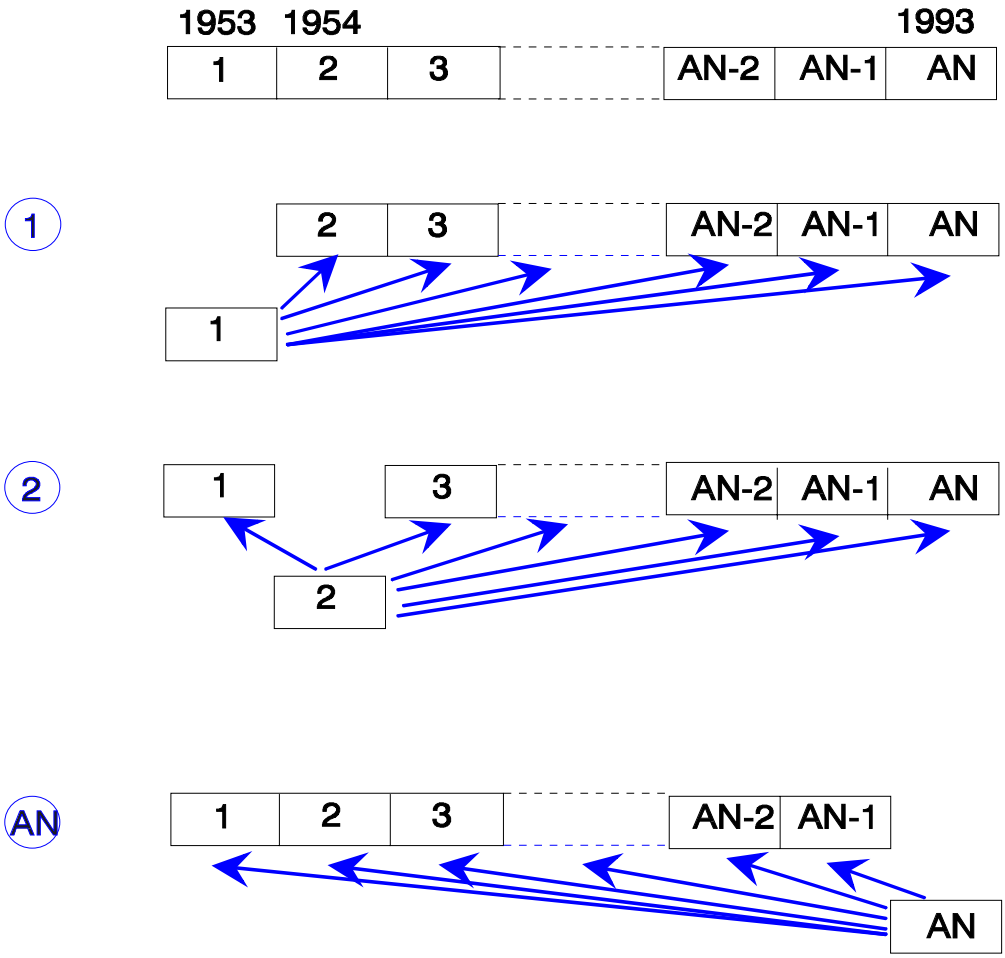


figure III-1: la validation croisée

III.1.1.b Choix des méthodes de référence

En terme de comparaison, plusieurs méthodes classiques de prévision de pluie feront office de référence. Tout d'abord, la méthode élaborée par Duband (1980), dite « de référence », sera la base de notre étude et notre référence principale. Ses performances, pour les deux types de prévision, sont données dans le tableau III-1, en moyenne sur les 33 bassins français pour la période 1953 à 1993. On y trouvera aussi celles de:

- *la climatologie* qui donne comme prévision la situation la plus probable climatologiquement sur la période 1953-1993 pour chaque bassin, c'est-à-dire:

* pour la prévision en pluie / non pluie, *non pluie* pour les bassins où il y a plus de jours secs que de pluvieux et *pluie* pour ceux avec un plus grand nombre de journées pluvieuses,

* pour la prévision probabiliste, les probabilités *a priori* d'être dans chacune des classes,

- *la persistance* qui consiste à prévoir pour aujourd'hui la situation de la veille (soit pluie / non pluie, soit une classe de pluie),

- *le hasard*: pour calculer la prévision, N=50 analogues sont tirés de manière aléatoire.

	Prévision pluie/non pluie IR	Prévision en classes RPS * 100
climatologie	56.5	75.4
persistance	72.8	98.6
hasard	53.4	76.6
méthode de référence	**72.2**	**59.7**

tableau III-1: performances des différentes méthodes

On peut noter que, si pour la climatologie et le hasard, les résultats sont nettement moins bons que ceux de la méthode de référence, ce n'est pas le cas pour la persistance qui donne une performance légèrement supérieure en prévision pluie / non pluie. C'est une méthode de prévision non négligeable à cause de la durée moyenne des épisodes pluvieux et non pluvieux, puisqu'elle ne se trompe qu'aux changements de temps.

Par contre, elle est nettement moins performante pour la prévision en classes car dans tous les cas, c'est une prévision catégorique soit complètement correcte, soit complètement fausse. Par conséquent, elle est beaucoup plus pénalisante, si elle est fausse, qu'une prévision probabiliste non catégorique avec une erreur sur la classe la plus probable dans le calcul de RPS. Elle est d'ailleurs moins bonne que la climatologie, qui reste une prévision probabiliste (la même pour chaque jour et pour un bassin donné, à savoir le pourcentage de jours d'automne de 1953 à 1993 dans chaque classe).

III.1.1.c Simplifications

Le fait d'utiliser deux critères comme dans la méthode de référence (cf. § I.2.2.b, les critères de proximité et de corrélation) pose un problème pour mettre en oeuvre des méthodes de sélection automatique de variables à utiliser dans ces critères. En effet, certaines variables apparaissent dans les deux critères et, de plus, le deuxième est un rapport:

① $$D^2\,(J,C) \;=\; \sum_{i=1}^{6} \left[\, Z_i(J) \;-\; Z_i(C) \,\right]^2 \;\leq\; R^2{}_b(C) \qquad \text{(III-1)}$$

où **J** est la journée potentiellement **analogue**,

C la **journée courante** où l'on veut faire la prévision,

Z_i (J) la valeur de la **ième CP** du champ 700 hPa pour la **journée J**,

R_b(C) le rayon de la boule de proximité.

② $$u^2\,(J,C) \;=\; \frac{D^2\,(J,C)}{R^2\,(J,C)} \;\leq\; 6 \text{ et } R^2\,(J,C) > 0.1 \qquad \text{(III-2)}$$

où **R** est le coefficient de corrélation entre le vecteur V de la journée de référence **C** et celui d'une situation **J** , composé:

- des 6 premières CP du champ de géopotentiel 700 hPa, Z_1 à Z_6,
- des 6 premières CP du champ de géopotentiel 1000 hPa, S_1 à S_6,
- et de la première CP du champ de l'épaisseur E_1.

Aussi avons-nous regardé s'il était possible de simplifier la sélection des analogues en ne gardant qu'un seul critère de distance et sans perdre en qualité de prévision.
En élargissant un peu la sélection, on pourrait remplacer le critère ② par $D^2\,(J,C) \leq 6$, puisque le coefficient de corrélation est toujours inférieur à 1. Et finalement, on pourrait encore simplifier en ne gardant qu'un seul critère de sélection: $D^2\,(J,C) \leq R^2_b$.

Et avec un critère de la sorte ($D^2\,(J,C) \leq R^2_b$. sur le champ de géopotentiel 700 hPa),

- soit, comme dans la méthode de référence, les analogues sont sélectionnés à l'intérieur d'une boule de proximité de rayon R_b(C), fonction de la distance à l'origine du jour C (*méthode 1*),
- soit, la boule de proximité est éliminée et on garde les analogues à l'intérieur d'une boule de rayon fixe R_b quelque soit le jour C (*méthode 2*),
- soit on sélectionne un nombre N constant d'analogues qui correspondent aux journées avec les N plus petites distances (*méthode 3*).

Les deux premières solutions possèdent l'avantage de ne sélectionner que des analogues proches dans l'espace des CP choisies mais avec le risque, pour quelques journées exceptionnelles, d'avoir peu voire pas d'analogues et donc pas de prévision. C'est donc la troisième, avec un nombre fixe d'analogue N, qui nous intéresse le plus. Elle permettrait par la suite, si N est assez grand, d'effectuer une sélection de deuxième niveau sur ces N analogues (cf. chapitre V).

Nous avons tout d'abord vérifié que le fait de modifier de cette manière la sélection des analogues ne diminuaient pas les performances de la méthode de manière rédhibitoire, que ce soit pour une prévision en pluie / non pluie ou en classes de pluie. Les résultats sont donnés dans le tableau III-2, en moyenne pour les 33 groupements français sur la période 1953-1993 (validation croisée) et pour les deux types de prévision:

	Prévision pluie/non pluie IR	Prévision en classes RPS * 100
méthode de référence	**72.2**	**59.7**
méthode 1 (R_b(C))	71.8	60.2
méthode 2 (R_b=2.9)	71.9	59.9
méthode 3 (N=50)	**72.6**	**59.0**

tableau III-2: performances des différentes méthodes
(avec IR qui doit augmenter et RPS diminuer)

- Le fait de ne garder que le critère de proximité avec une boule de proximité variable selon les jours (méthode 1) entraîne une légère perte de performance car on est un peu moins sélectif.

- Pour l'utilisation d'une boule fixe (méthode 2), le rayon optimal, qui minimise le RPS pour une prévision probabiliste et qui maximise l'indice de réussite IR, est compris entre 2.9 et 3 (cf. en annexe III.1.a l'optimisation du rayon de la boule) et les résultats obtenus sont pratiquement équivalents à ceux de la méthode de référence.

- Enfin, si au lieu d'utiliser une boule de proximité fixe ou variable, on choisit de sélectionner un nombre constant d'analogues chaque jour, le nombre optimal d'analogue est compris entre 30 et 70 (cf. annexe III.1.b). Notre choix s'est porté ensuite sur N=50 de manière assez arbitraire encore que ce soit un nombre que l'on retrouve dans la littérature (Woodcock et Woodcock & Keenan, 1980). Il devrait d'ailleurs être fonction de la taille du fichier d'apprentissage disponible (ici 40 ans).

Et c'est cette dernière méthode qui donne les meilleurs résultats.

Toutefois, nous verrons plus loin (§III.3.1) que nous prendrons éventuellement en compte la dispersion des analogues en introduisant une pondération de ceux-ci. Dans ces conditions, le nombre d'analogue retenu perd de l'importance car les meilleurs analogues prennent un poids important tandis que les plus éloignés recevront un poids négligeable. Mais dans l'immédiat, nous

les considérerons comme des candidats équivalents à l'analogie quelle que soit leur distance $D^2(J,C)$.

En résumé, la meilleure prévision est donnée par la sélection de 50 analogues grâce au critère suivant:

$$D^2\,(J,C) \;=\; \sum_{i=1}^{6} \left[\, Z_i\,(J) \;-\; Z_i\,(C) \,\right]^2 \qquad \text{(III-3)}$$

Cependant, on peut se demander dans un premier temps pourquoi seules les 6 premières CP du champ 700 hPa sont utilisées. En effet, il est loin d'être évident que seules celles-ci soient discriminantes pour la pluie.

Aussi peut-on envisager d'optimiser la distance euclidienne de la façon suivante (tout en continuant à sélectionner un nombre constant N = 50 analogues):

$$D^2\,(J,C) \;=\; \sum_{i=1}^{P} \left[\, X_i\,(J) \;-\; X_i\,(C) \,\right]^2 \qquad \text{(III-4)}$$

avec P Composantes Principales X à sélectionner, parmi les CP disponibles, pour optimiser la prévision, c'est-à-dire soit augmenter le score IR pour la prévision en pluie / non pluie, soit diminuer le score RPS pour une prévision probabiliste en classes de pluie.

On pourra, dans un deuxième temps, se poser le problème de savoir si elles doivent toutes être utilisées de la même manière et s'il n'y a pas lieu d'introduire des pondérations différentes suivant les CP (cf. § III.3.2).

III.1.2 La sélection des variables

Il existe de nombreux algorithmes de sélection des variables, plus ou moins performants et plus ou moins coûteux en temps de calcul. Pour notre étude nous avons choisi de faire de la sélection ascendante, car même si la recherche du maximum n'est pas optimale, les temps de calcul restent acceptables.

III.1.2.a La sélection ascendante simple (figure III-2)

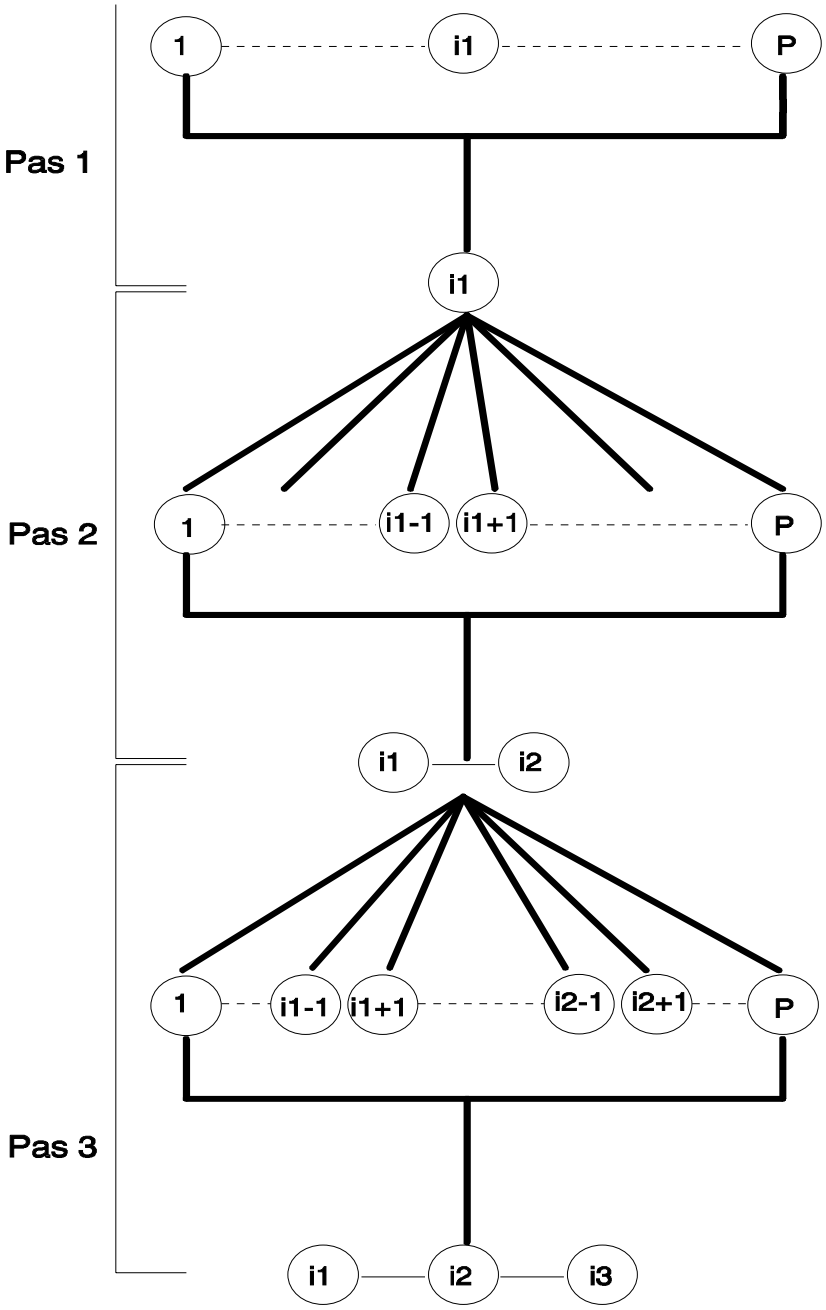


figure III-2: sélection ascendante simple

Pas 1: sélection de la meilleure variable pour extraire les analogues

① Pour chaque journée **C** de 1953 à 1993, N analogues **J** sont sélectionnés grâce à une distance à la journée à prévoir **C** utilisant la seule variable expli cative ou prédicteur X_i:

$$D_{1i}^2(J,C) = \left[X_i(J) - X_i(C)\right]^2 \qquad \text{(III-5)}$$

Les N analogues retenus sont ceux avec les N plus petites distances et la prévision est faite avec leurs précipitations.

② Un critère de performance est calculé pour l'ensemble des 3731 prévisions faites avec la variable X_i.

Cette démarche (①,②) est effectuée successivement pour les P variables disponibles. La variable $\mathbf{i_1}$ sélectionnée est celle dont le critère de performance $\mathbf{Cr(i_1)}$ est le meilleur .

Pas 2: sélection de la meilleure variable, en couple avec i_1, pour extraire les analogues

Le principe est le même qu'au pas 1, mais la sélection des N analogues se fait cette fois grâce à la variable $\mathbf{i_1}$ et une autre variable i = 1 à P, $i \neq i_1$:

$$D_{2i}^2(J,C) = \left[X_{i_1}(J) - X_{i_1}(C)\right]^2 + \left[X_i(J) - X_i(C)\right]^2 \qquad \text{(III-6)}$$

variable déjà retenue variable à tester

La variable $\mathbf{i_2}$ sélectionnée est donc celle qui, en combinaison avec $\mathbf{i_1}$, donne le meilleur critère de performance $\mathbf{Cr(i_1,i_2)}$.

etc.....

Il est bien évident que cette méthode de sélection peut ne pas être optimale mais elle possède l'avantage d'être simple à mettre en oeuvre et assez rapide.

Même la sélection descendante, pourtant peu différente dans son principe (au lieu d'ajouter des variables on en enlève), est beaucoup moins rapide et donc peu intéressante pour nous.

Bien sûr, la méthode optimale de sélection consiste à chercher le meilleur couple, puis le meilleur triplet....Mais il n'était pas envisageable de l'utiliser car elle est beaucoup trop combinatoire et donc coûteuse en temps de calcul.

Cependant, nous avons quand même essayé d'améliorer la méthode ascendante de sélection des variables dans le but de la rendre un peu plus optimale dans sa recherche du maximum.

III.1.2.b Sélection ascendante des K meilleures variables (figure III-3)

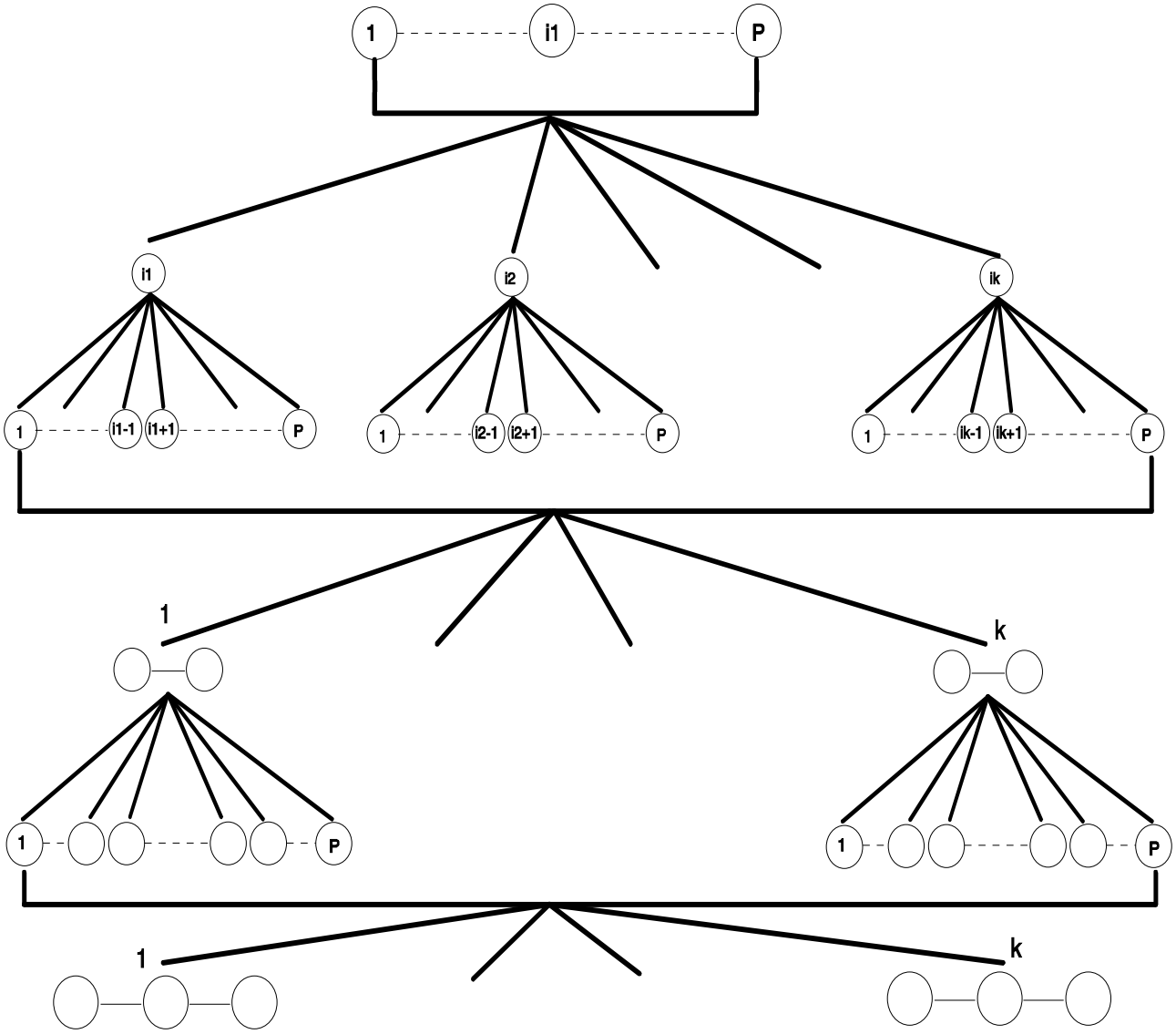


figure III-3: sélection ascendante des K meilleures variables

Pour optimiser la sélection ascendante, on peut envisager de garder non pas une mais **K** meilleures variables à chaque pas. Cela permet d'explorer plus de voies pour trouver le maximum.

Pas 1: c'est le même que pour la sélection ascendante simple mais les K meilleures variables i_1, ... , i_K sont conservées au lieu de la seule meilleure.

Pas 2: le pas 2 de la sélection ascendante simple (cf. § III.1.2.a) est effectué pour toutes les variables i_k, k = 1 à K. C'est-à-dire que l'on teste tous les couples possibles avec i_1, tous ceux avec i_2 etc...jusqu'à tous ceux avec i_K, soient K ∗ (P-1) couples.
Et seuls les K meilleurs couples, sur ces K ∗ (P-1), sont conservés pour passer au pas 3.
Et ainsi de suite...

La valeur K est à optimiser en fonction du temps de calcul, des capacités de la machine et du gain par rapport à K-1. Quelques essais ont montré que **K=2** était un bon compromis. C'est cette valeur qui sera utilisée par la suite.

Finalement, nous avons commencé à travailler sur la sélection ascendante simple (K=1) afin de cerner le problème. Et ce n'est que lorsque nous avons voulu réellement chiffrer les gains obtenus que nous sommes passés à la sélection K=2 car les temps de calculs sont quand même multipliés par 2 environ.

Remarque: nous avons aussi essayé de remettre en cause à chaque pas les variables sélectionnées. Ainsi, après le pas 3 où le triplet (i_1,i_2,i_3) a été sélectionné, la variable i_1 est enlevée et le couple (i_2,i_3) est testé avec toutes les autres variables. Si un triplet meilleur que (i_1,i_2,i_3) est trouvé, il le remplace dans le pas 4, etc.. Mais les gains obtenus étaient trop minimes par rapport à l'augmentation des temps de calcul.

III.2 Résultat de la sélection

Dans ce paragraphe, nous allons présenter les résultats obtenus par les différentes techniques de sélection des Composantes Principales (CP) et par les différentes CP disponibles: quels champs, quelles échéances et quelles CP (calculées par ACP ou ACPP, avec une matrice de corrélation ou de covariance ...). Un grand nombre d'essais a donc dû être effectué pour arrêter d'une part, le choix d'une technique de sélection et d'autre part, le type de CP utiles.

En outre, en raison des temps de calcul relativement importants, cette première série d'essais a été uniquement faite pour la prévision en pluie / non pluie. Les meilleurs résultats seront ensuite appliqués à la prévision probabiliste.

III.2.1 Sélection ascendante simple sur 36 CP (3∗12 CP)

Un premier essai de sélection ascendante simple a été fait en partant de P=36 CP normées calculées par ACP de Processus utilisant la matrice de *covariance* (cf. chapitre II, § II.1.3):

- les 12 premières du champ de géopotentiel 700 hPa: Z_1 à Z_{12},
- les 12 premières du champ de géopotentiel 1000 hPa: S_1 à S_{12},
- les 12 premières de l'épaisseur de la couche 700/1000 hPa: E_1 à E_{12}.

Cette sélection a été faite, bassin par bassin, pour l'ensemble des 33 bassins français et nous nous sommes plus particulièrement penchés sur quelques aspects pour interpréter ces résultats.

III.2.1.a Evolution de l'indice de réussite en fonction du nombre de CP retenues

La courbe de l'indice de réussite IR = f (nombre de CP retenues) a été tracée pour l'ensemble des 33 groupements. Trois exemples sont donnés sur la figure III-4, le reste étant présenté dans l'annexe III.2.

On peut remarquer qu'à l'exception de 3 bassins (Agout-Tarn, Drac et Cure) cette courbe de l'indice de réussite se décompose en 3 parties:

- montée très rapide de IR entre 1 et 3-4 CP retenues,
- montée plus lente jusqu'à 6-8 CP retenues,
- stabilisation au delà.

Quant aux 3 bassins restants, leurs courbes ne se décomposent qu'en 2 parties, le palier étant atteint après une seule montée rapide. Et l'on peut noter que, sur ces 3 bassins au comportement légèrement différent, 2 sont isolés géographiquement par rapport aux autres: Cure et Agout-Tarn (cf. fig. I-2).

Ces quelques remarques nous ont permis de choisir un palier à 8 CP retenues pour les 33 bassins. Les CP sélectionnées ensuite n'apportent pas d'amélioration significative pour la prévision.

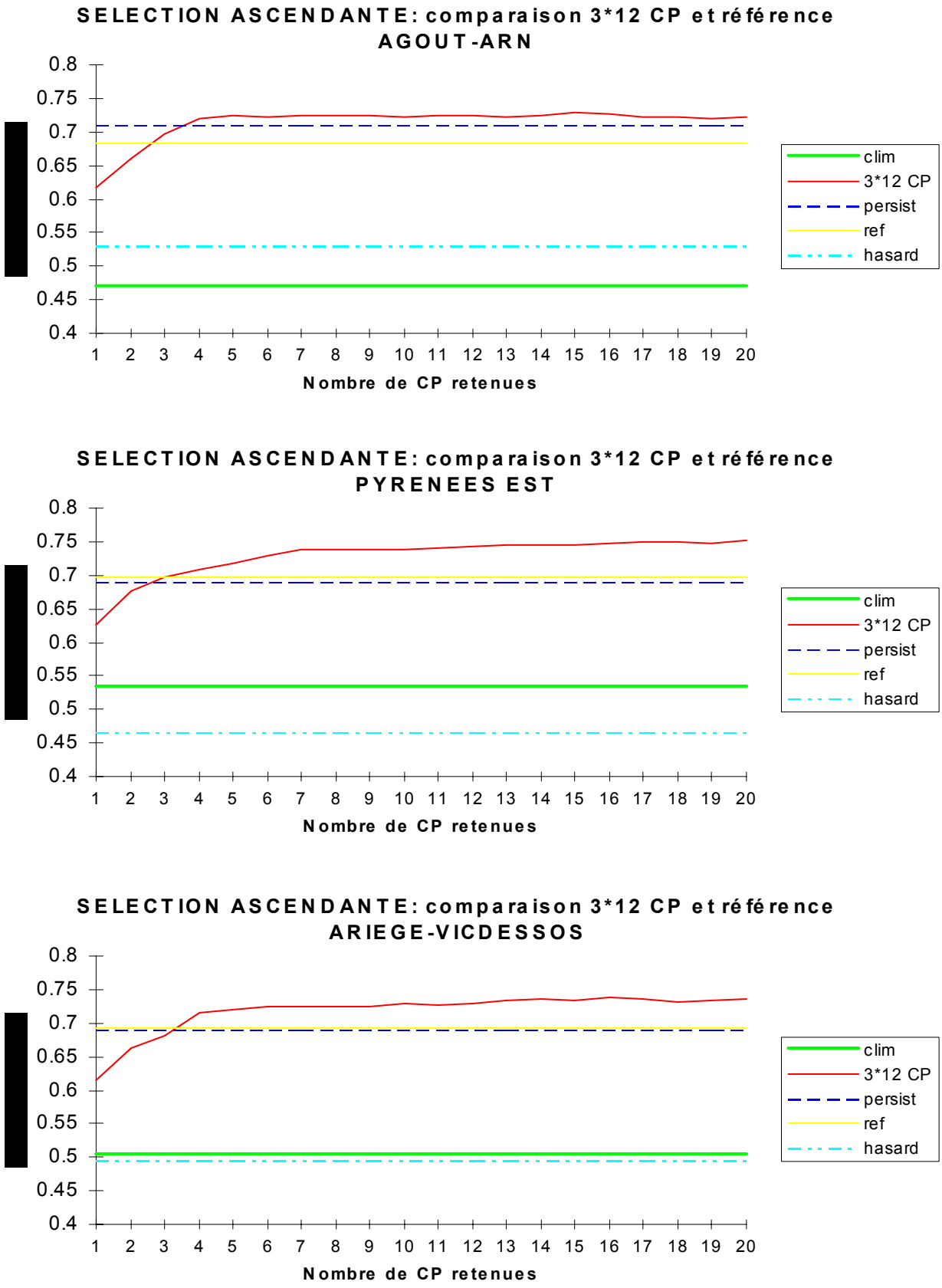


figure III-4: sélection ascendante simple sur 36 CP, IR = f (nombre de CP retenues)

III.2.1.b Comparaison avec les méthodes de référence

Pour tous les groupements, lorsque l'on atteint le palier à 8 CP, l'indice de réussite est meilleur que pour la climatologie, la méthode de référence et le hasard (cf. figure III-4 et le reste en annexe III-2). Il l'est aussi si l'on avait pris les 36 CP disponibles pour faire la sélection des analogues (méthode 36 CP). Enfin, la persistance donne de meilleurs résultats pour 4 groupements uniquement: Vézère-Vienne-Thaurion, Gaves, Ain-Valserine et Arve-Fier.

III.2.1.c Les 8 premières CP retenues (cf. figure III-5)

Une première constatation à faire est que, dans les 8 premières CP sélectionnées, celles de l'épaisseur de la couche 700/1000 hPa sortent très peu, et ce, quelque soit le bassin considéré. En effet, sur les 33 bassins et pour les 8 premières CP retenues, soit sur 264 CP sélectionnées, les CP de l'épaisseur représentent 18.5% de l'ensemble des 264 CP. Pour ce qui est des champs 700 et 1000 hPa, ils correspondent respectivement à 44 et 37.5 %.

Au vu de ces résultats, il est tentant d'éliminer l'épaisseur, qui n'est, en fait, qu'une combinaison linéaire des champs de géopotentiel 700 et 1000 hPa, tout comme les CP E de l'épaisseur sont combinaisons linéaires des CP Z et S des champs de géopotentiel 700 et 1000 hPa. Donc toute l'information contenue dans les CP E de l'épaisseur est incluse dans les CP Z et S.
Cependant, la seule façon de pouvoir enlever les CP de l'épaisseur, c'est de regarder si l'indice de réussite ne se détériore pas si l'on part de 24 CP (Z_1 à Z_{12}, S_1 à S_{12}), au lieu de 36 (Z_1 à Z_{12}, S_1 à S_{12}, E_1 à E_{12}).
Pourtant, avec une régression on arrive à retrouver les 12 CP de l'épaisseur à l'aide des 12 CP 700 et 1000 hPa. Néanmoins cela ne suffit pas car, avec la sélection ascendante, on force la pondération des CP à 0 ou 1. On ne peut donc jamais retrouver exactement les coefficients de l'équation de régression et toute l'information contenue dans les CP de l'épaisseur.

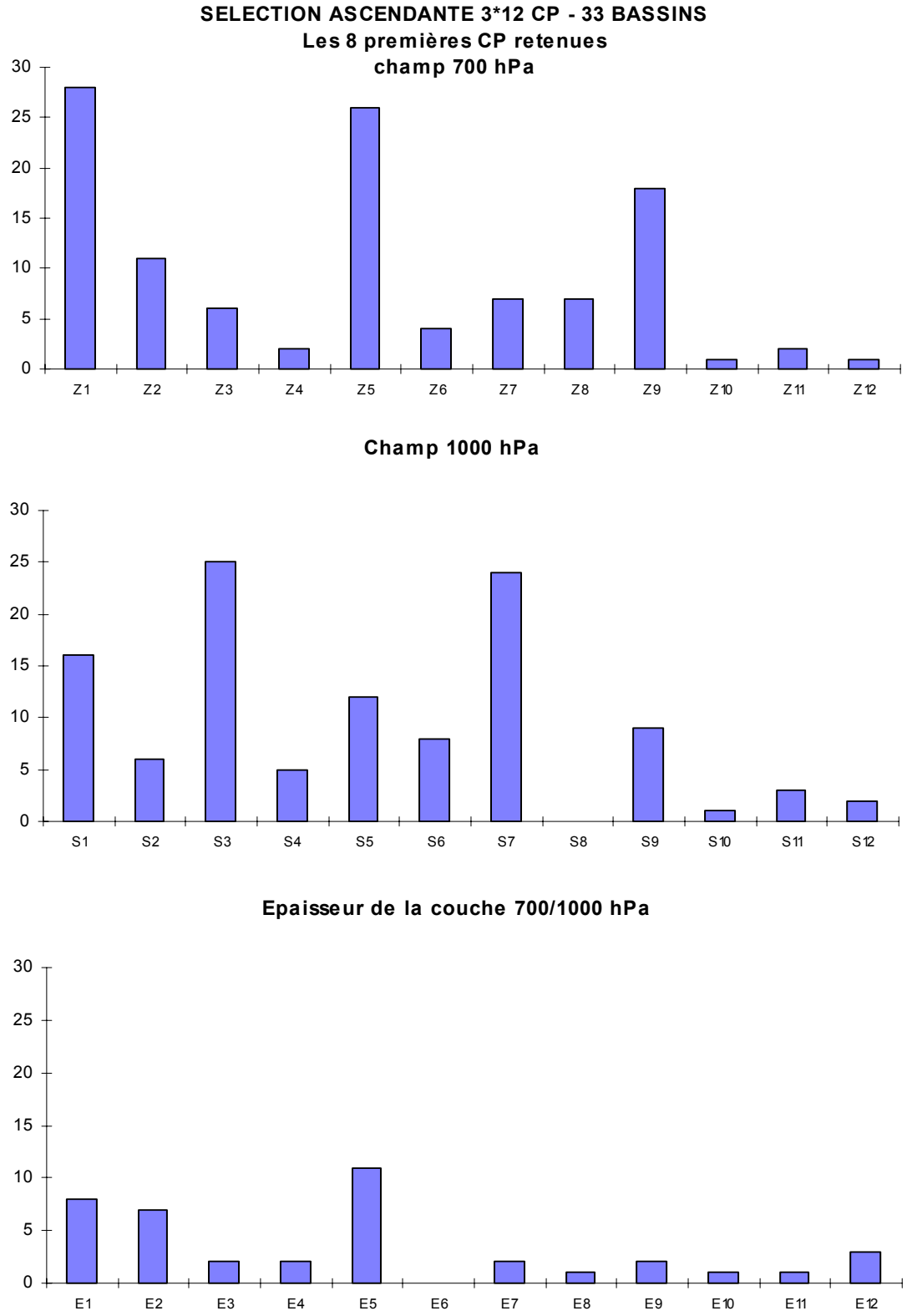


figure III-5: Les 8 premières CP retenues pour les 33 bassins

III.2.2 Elimination de l'épaisseur

Considérons maintenant ce qu'il advient lorsque l'on élimine l'épaisseur et que l'on fait la sélection ascendante sur un ensemble de P=2*12 au lieu de 3*12 Composantes Principales (CP).

III.2.2.a Comparaison 24 / 36 CP

On peut noter une très bonne adéquation entre les 2 courbes de l'indice de réussite (cf. un exemple de courbe en figure III-6). Quant aux résultats obtenus pour 8 CP retenus, ils sont très voisins (cf. annexe III-3) puisque le pourcentage de réussite moyen pour les 33 bassins est de 75.48 si l'on part de 36 CP et de 75.5 sinon. D'ailleurs pour 9 bassins, les 8 premières CP retenues sont les mêmes pour 24 ou 36 CP, l'épaisseur n'y apparaît donc pas. Pour 18, l'indice de réussite est soit meilleur, soit considéré comme équivalent (écart inférieur à 0.3% en valeur absolue). Et seulement pour 6 groupements, l'indice de réussite est plus faible à 24 qu'à 36 CP, mais la baisse n'excède pas 1%.

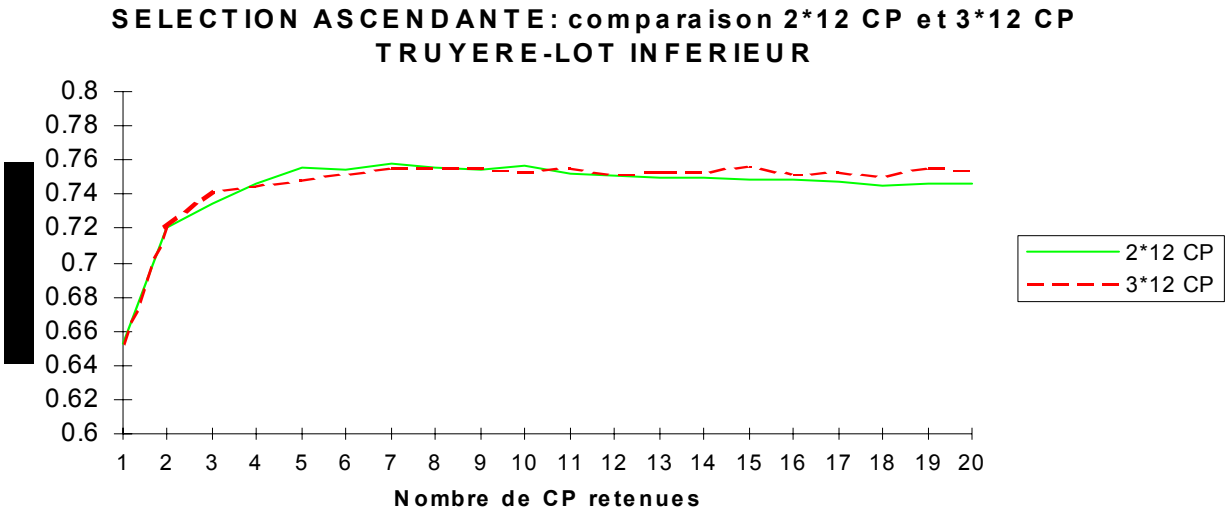


figure III-6: indice de réussite pour 24 et 36 CP

Alors bien sûr, il faut rappeler que ces résultats sont très liés à la méthode de sélection des variables choisie qui n'est pas tout à fait optimale. Cependant on peut considérer que les

performances sont équivalentes. Cela nous permet d'éliminer l'épaisseur sans pour autant nuire à la qualité de la prévision.

III.2.2.b Les 8 premières CP retenues

Nous avons regardé plus en détail les CP retenues pour l'ensemble des 33 bassins (vision globale), puis bassin par bassin (vision locale) pour voir s'il n'existait pas une certaine homogénéité par région.

Vision globale (figure III-7):

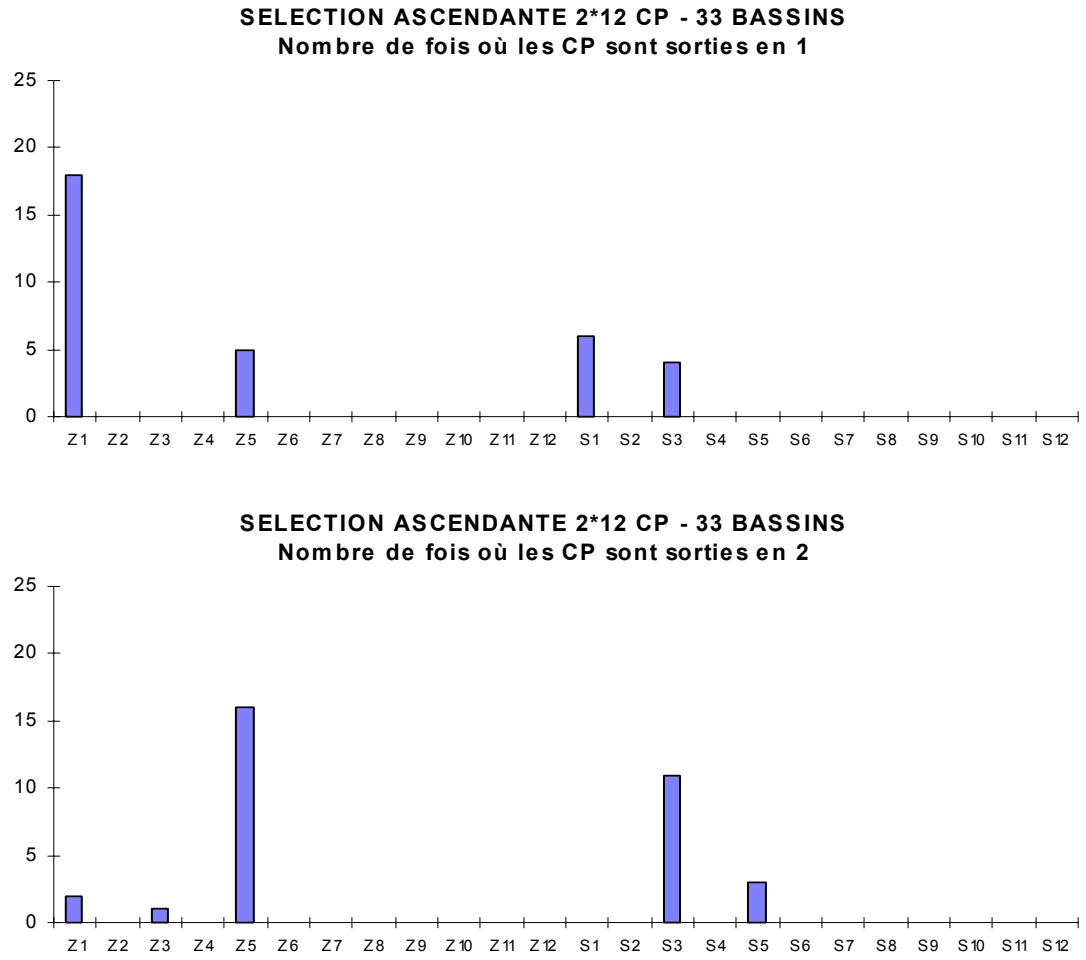


figure III-7: histogramme des CP sorties en 1ère et 2ème position

Sur l'ensemble des 33 bassins seules 4 CP différentes sont sorties en premier, en particulier Z_1 (18/33) et S_1 (6/33) dont la corrélation avec Z_1 est à noter: 0.86 (cf. annexe III-4). Z_5 et S_3 sont sorties de manière plus marginale.

Avec 2 CP retenues par bassin on retrouve, bien sûr, les 4 précédentes (Z_1, Z_5, S_1, S_3) et seulement 2 autres (Z_3 et S_5) dont les corrélations sont aussi très bonnes avec S_3 et Z_5 respectivement: 0.83 et 0.64.

Au delà de 2 CP retenues, le nombre de CP utilisées augmente rapidement (12 pour 3 et 14 pour 4) comme on peut le voir sur les histogrammes portés en annexe III-5.

Vision locale: nous avons dressé des cartes indiquant, à l'emplacement de chaque bassin, la CP sortie aux rangs 1 et 2 (fig. III-8), puis le couple de CP sorties en 1 et 2 (fig. III-9).

a) sorties en 1 *b) sorties en 2*

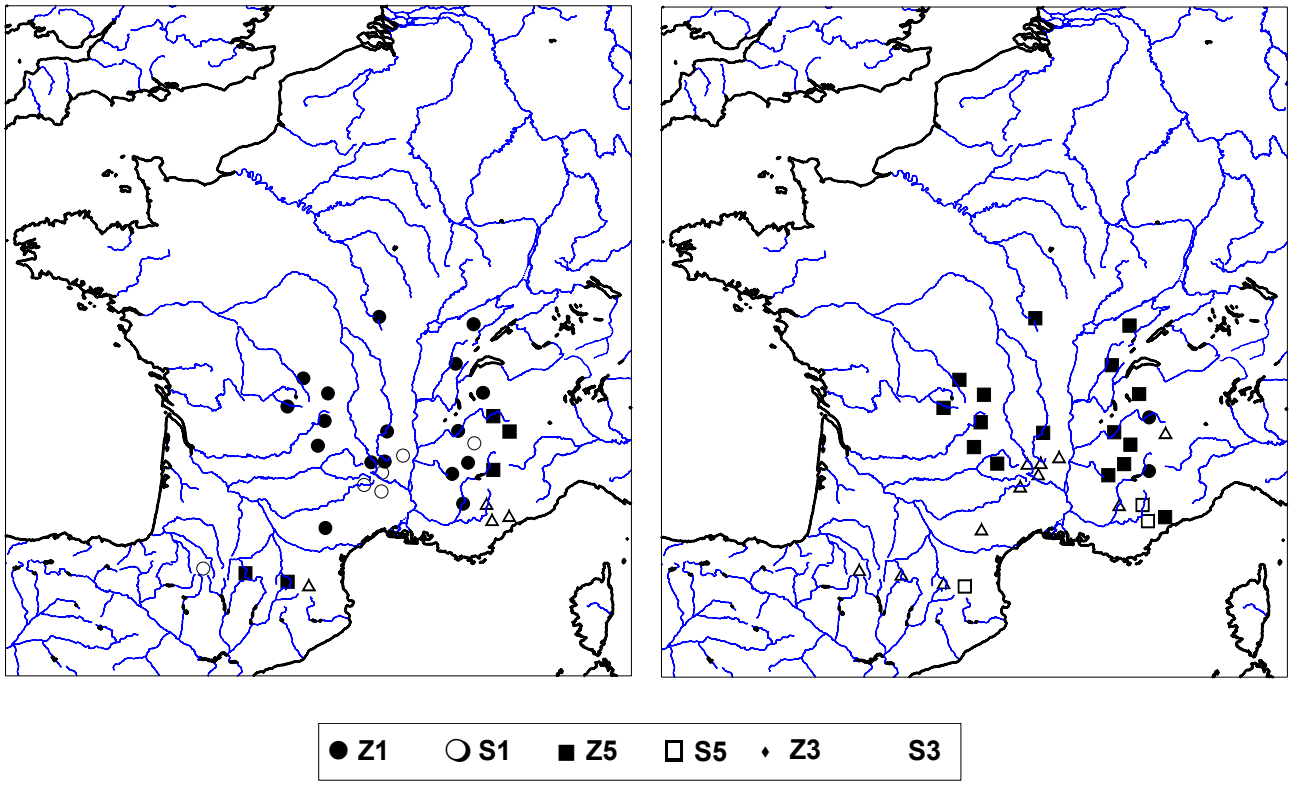


figure III-8: les CP sorties en 1 et 2 pour chaque bassin

a) sorties en 1, b) sorties en 2

1ère CP retenue: dans le Nord, nette prédominance de Z_1. Au centre de la zone d'étude (Cévennes) c'est surtout S_1. Plus au Sud, c'est plutôt S_3 qui apparaît, même si le résultat est moins homogène pour les Pyrénées.
2ème CP retenue: cette fois Z_5 et S_5 prédominent au Nord et à l'Est. Ailleurs et notamment sur toute la zone à influence méditerranéenne, S_3 s'impose.

Si l'on regarde maintenant le couple (1ère - 2ème CP) les résultats sont relativement homogènes par région (cf. fig. III-9). Au Nord et à l'Est, tous les groupements ont sélectionné le couple (Z_1,Z_5) sauf une fois (S_1,Z_5). Dans les Cévennes ce sont les couples (S_1,S_3), (Z_1,S_3) ou (S_1,Z_3) qui ressortent. Enfin, plus au Sud, on trouve le plus souvent les couples (S_3,Z_5) et (S_3,S_5).

Finalement, 3 types de couples apparaissent pour 3 régions:

- au Nord	(Z_1,Z_5) ou (S_1,Z_5)	avec r (Z_1,S_1) = 0.86,
- au Sud	(S_3,Z_5) ou (S_3,S_5)	avec r (Z_5,S_5) = 0.64,
- dans les Cévennes	(S_1,S_3), (Z_1,S_3) ou (S_1,Z_3)	avec r (Z_3,S_3) = -0.83.

Ce résultat est assez cohérent. On peut noter que S_3 et Z_3 sont sélectionnés pour les bassins du pourtour méditerranéen. Or d'après les cartes des cosinus directeurs du § II.1.3.d (fig. II-6), ces 2 CP représentent les flux du Sud, Sud-Est et sont sans doute des marqueurs des « coups » de Sud, Sud-Est, particulièrement violents dans ces régions.

Remarque: on a ensuite regardé les 3ème et 4ème CP sélectionnées. Mais les cartes sont beaucoup moins homogènes et donc beaucoup moins faciles à interpréter aussi n'avons-nous pas été plus loin. On peut cependant noter les couples (S_3,Z_9) ou (Z_3,Z_9), présents pour 1/3 des bassins.

III.2.2.c Vérification du choix de N=50 analogues

Avec la sélection des analogues par une distance euclidienne utilisant les 6 premières CP du champ 700 hPa, nous avions montré que N=50 était un nombre optimal d'analogues à retenir (cf. §III.1.1.b). Mais on peut se demander ce qu'il en est lorsque l'on sélectionne 8 CP à introduire dans ce critère de sélection.

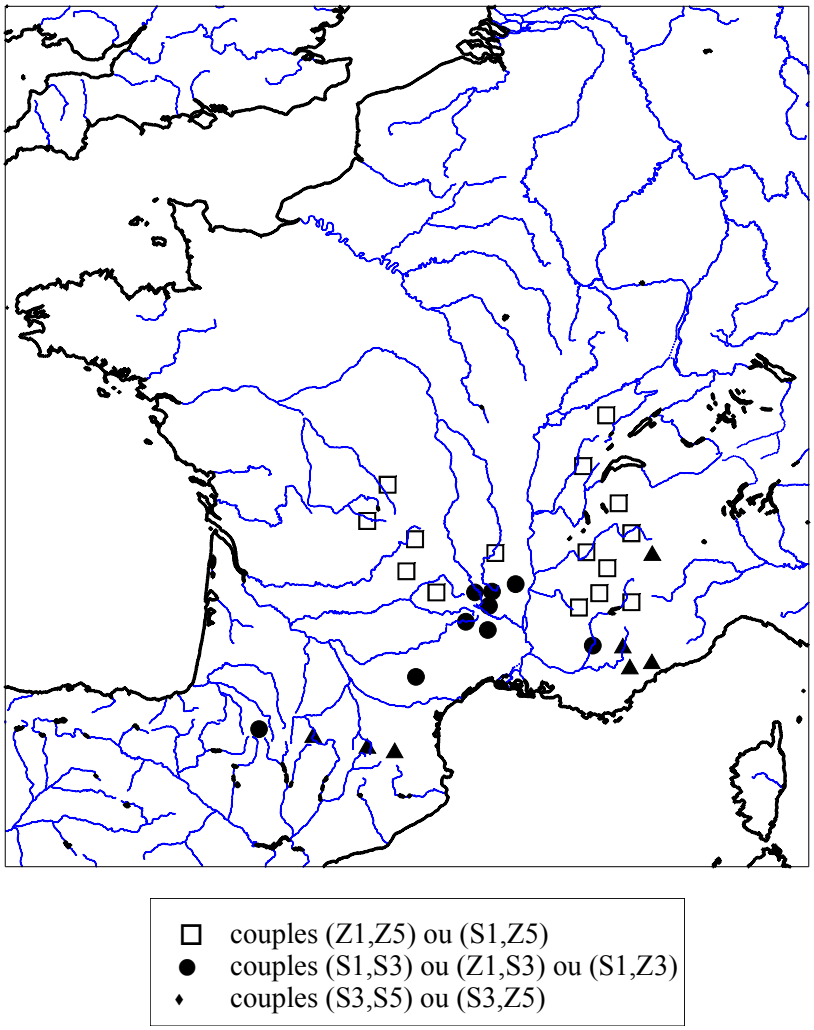


figure III-9: les CP retenues en 1 et 2 pour chaque bassin

Nous avons donc effectué plusieurs sélections ascendantes sur 24 CP avec différents nombres N d'analogues retenus, et ceci pour 6 bassins témoins: Creuse-Cher, Pyrénées Est, Doubs, Isère Moyenne, Var-Tinee-Roya, Chassezac.

Ces 6 groupements ont été choisis pour leur bonne répartition géographique :

- Creuse-Cher au Nord,
- Pyrénées-Est dans les Pyrénées,
- le Doubs pour le Nord-Est,
- l'Isère moyenne dans le Centre-Est,
- Var-Tinee-Roya au Sud-Est,
- et Chassezac dans les Cévennes.

Ils nous serviront de bassins tests dans toute la suite de notre étude. Aussi, nous ne testerons les 33 bassins que si les résultats des 6 bassins témoins sont positifs.

Nous avons ensuite tracé l'indice de réussite pour 8 CP retenues en fonction du nombre d'analogues. Les résultats se trouvent en annexe III-6: l'abscisse représentant le nombre d'analogues est une échelle logarithmique et les points correspondent aux abscisses 1, 2, 3, 5, 8, 12, 18, 25, 36, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500, 600, 700, 800 et 1000.
Dans tous les cas, on remarque que l'optimum se trouve entre 50 et 100 analogues. Donc, l'hypothèse formulée au III-1-2 (N=50) reste acceptable.

III.2.2.d Sélection ascendante K=2 sur 24 CP

Après avoir réduit le nombre de CP utiles à 24 avec un palier à 8 CP retenues, nous avons testé l'algorithme de sélection ascendante K=2, plus coûteux en temps de calcul mais un peu plus optimal dans sa recherche du maximum.
Pour 30 bassins sur les 33, l'indice de réussite pour 8 CP retenues est légèrement supérieur à celui obtenu par la sélection simple (tableau en annexe III-7). Et pour les 3 restants la perte est minime. Aussi avons-nous décidé de conserver cette sélection de variables ascendante K=2 pour la suite de nos essais.

III.2.2.e Conclusion

A l'aide de ces deux premiers essais (sélection sur P=36 puis 24 CP), nous avons pu choisir une méthode de sélection automatique de variables, la **sélection ascendante K=2**. Et nous avons pu **éliminer le champ de l'épaisseur** qui s'est avéré ne pas apporter d'information supplémentaire quant à la prévision d'occurence de pluie.
Nous avons donc abouti à une nouvelle méthode de sélection des analogues, utilisant pour extraire les analogues une distance euclidienne avec 8 CP retenues grâce à une sélection ascendante K=2 sur 24 CP (700 et 1000 hPa): **Méthode S-8CP**.
Les résultats obtenus, bassin par bassin, sont consignés en annexe III-8. Et, en moyenne sur les 33 bassins français sur la période 1953-1993, le pourcentage de réussite est passé de:

- **72.2**% pour la méthode de référence,
- à **75.7**% pour la méthode S-8CP, soit un gain moyen de **3.5**%.

III.2.3 Les différents types de CP

III.2.3.a ACPP de corrélation ou de covariance

Jusque là, les Composantes Principales étaient normées et calculées par une ACP de Processus (cf. § II-.1.3) utilisant la *matrice de covariance* des variables.
Cependant, elles peuvent aussi être obtenues à l'aide de la *matrice de corrélation* et les résultats seront alors différents. Il est aussi possible de ne pas les normer (cf. éq. II-8), ce qui introduit alors une pondération intrinsèque aux CP. Cela nous amène d'ailleurs à penser qu'il pourrait être intéressant, par la suite, de pondérer les variables, et ne pas uniquement les sélectionner.
Il existe donc 4 possibilités pour le calcul des CP:

- ACPP de covariance - CP normées (notre cas jusqu'à présent dans ce chapitre III),
- ACPP de covariance - CP non normées,
- ACPP de corrélation - CP normées,
- ACPP de corrélation - CP non normées.

Il est à noter que le pourcentage de variance expliquée en fonction du nombre de CP varie très peu suivant la méthode utilisée pour calculer les CP (annexe III-9).

Pour les 6 groupements témoins (Creuse-Cher, Pyrénées Est, Doubs, Isère moyenne, Var-Tinee-Roya et Chassezac), les courbes de l'indice de réussite (figure III-10 et annexe III-10) résultant d'une sélection ascendante K=2 en partant de 24 CP, ont été comparées pour les 4 types de CP.

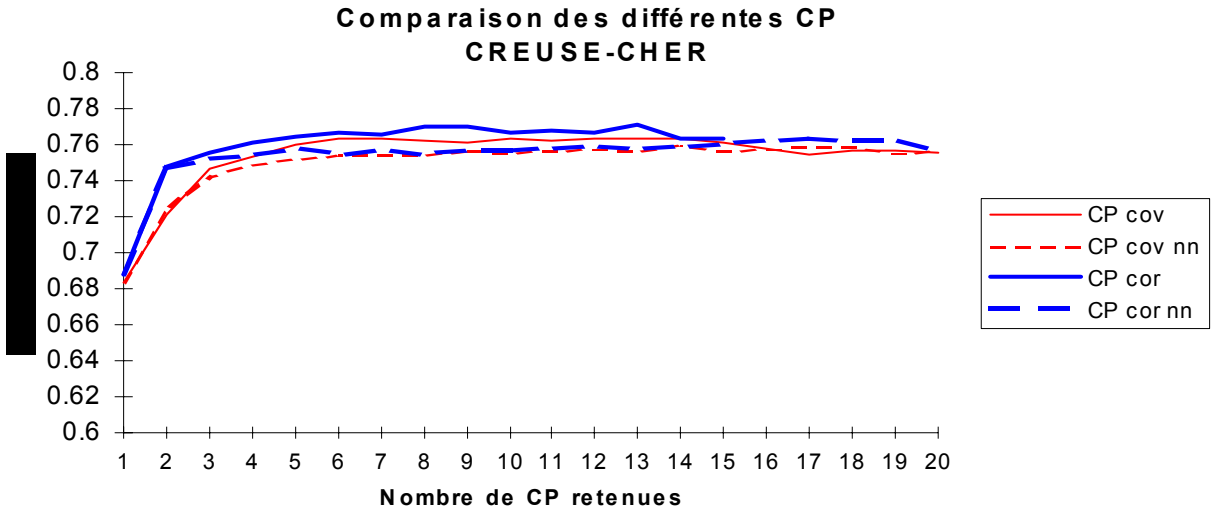


figure III-10: Courbe de l'indice de réussite pour différents types de CP

De manière générale, on remarque que les CP non normées donnent de moins bons résultats que les CP normées (- 1 à 2% pour 8 CP retenues). Les CP calculées par la matrice de corrélation semblent, quant à elles, donner des résultats un peu meilleurs que celles calculées par ACPP de covariance pour 8 CP retenues. Seul le Doubs se comporte un peu différemment. Pour ce bassin, les CP calculées avec la matrice de covariance donnent un meilleur indice de réussite pour 8 CP retenues et les CP non normées calculées avec la matrice de corrélation ont des résultats légèrement supérieurs aux CP normées correspondantes.

Mais finalement, nous retiendrons que les **CP normées** calculées avec la **matrice de corrélation** donnent des résultats un peu meilleurs. De plus, elles possèdent l'avantage de pouvoir y ajouter des variables de nature différente qui, si elles sont centrées réduites, seront alors comparables.

III.2.3.b Application à la méthode de référence

Dans la méthode de référence les CP utilisées sont des CP normées calculées par une ACP classique avec la matrice de corrélation. Nous avons voulu regarder ici ce que donnent les différents types de CP utilisées dans la méthode référence. Ainsi, vous trouverez dans le tableau III-3 les

performances obtenues en moyenne sur les 33 bassins, par la méthode de référence si elle utilise des CP normées calculées par:

- ACP classique,
- ACPP de covariance,
- ACPP de corrélation.

méthode de référence	Prévision pluie/non pluie IR	Prévision en classes RPS * 100
ACP classique	**72.2**	**59.7**
ACPP covariance	72.3	59.5
ACPP corrélation	72.8	59.1

tableau III-3: Performances de la méthode de référence avec différents types de CP

Remarque: nous n'avons pas pu tester les CP non normées car, avec celles-ci, les distances euclidiennes $D^2(J,C)$ deviennent plus grandes et alors les critères de la méthode de référence ne conviennent plus.

Cela nous conforte donc dans notre idée de conserver les **CP calculées par ACPP de corrélation**. De plus, cela indique que l'utilisation des nouvelles CP (provenant de l'ACPP qui tient compte de la topologie du réseau) n'améliore pas uniquement la condensation de l'information; elle intervient de manière positive dans la prévision de pluie.

III.2.3.c Composantes Principales combinées **

Pour l'instant, les CP sont calculées sur les 2 champs séparément. L'analogie se fait donc sur l'un ou sur l'autre. Il pourrait être intéressant de la faire globalement sur les 2 champs par l'intermédiaire de CP orthogonalisées sur l'ensemble des 2 champs (CP globales).

Une ACP classique, utilisant la matrice de corrélation, a été effectuée sur les données à 700 et 1000 hPa des 37 radiosondages. La courbe de variance expliquée en fonction du nombre de CP retenues est présentée en annexe III-11.

Une sélection ascendante K=2 a été effectuée sur les 35 premières CP pour les 6 groupements test. Le choix de 35 CP est basé sur un compromis entre:

- prendre assez de CP pour expliquer un maximum de variance (ici 99%)
- et pas trop pour garder des temps de calcul acceptables.

Mais les résultats, comparés à ceux obtenus par la méthode S-8CP n'étant pas concluants sur les 6 groupements test (annexe III-12), nous avons abandonné cette idée sans même tester l'ACP de Processus.

III.2.4 Prise en compte de l'information à 00 et 24 h

Jusqu'à présent, les CP utilisées pour déterminer la prévision de pluie d'un jour C de 07 h du matin à C+1 à 07 h le lendemain matin, étaient calculées à l'aide des données obtenues à 00 h le jour C. Or, d'après la figure III-11 qui montre les différentes échéances, il semble plus judicieux, pour faire l'analogie, d'utiliser aussi les CP à 24 h, c'est-à-dire celles du lendemain C+1 à 00 h, car cela permettrait de mieux « encadrer » la pluie. Par conséquent, on peut envisager de sélectionner les analogues avec une distance euclidienne de la forme:

$$D^2(J,C) = \sum_{i=1}^{P_1} \left[X_i(J) - X_i(C) \right]^2 + \sum_{i=1}^{P_2} \left[X_i(J+1) - X_i(C+1) \right]^2 \qquad \text{(III-7)}$$

où P_1 représente le nombre de CP à 00 h choisies et P_2 celui des CP à 24 h.

Il pourrait aussi être intéressant d'utiliser les CP de la veille C-1 à 00 h (-24 h) avec celle du jour C car cela pourrait donner une idée de l'évolution temporelle des masses d'air.

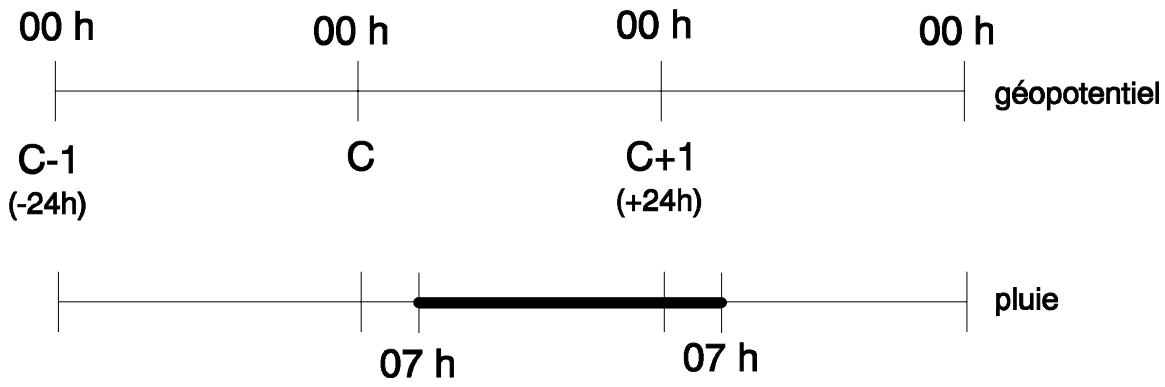


figure III-11: échelle temporelle des données

Nous avons donc effectué 3 essais de sélection de variables K=2 en partant de:

i) 24 CP à 00 h (C) et 24 CP à C-1 (-24h),

ii) 24 CP à 00 h (C) et 24 CP à 24 h (C+1),

iii) 24 CP à 24 h (C+1),

les CP étant encore issues de l'ACPP de covariance pour des raisons chronologiques (l'ACPP de corrélation ayant été faite avant ces premiers essais sur la prise en compte de l'information à 24 h). Avec l'essai *iii)*, nous avons voulu voir ce que donnaient les CP à 24 h seules car elles sont mieux centrées par rapport à la pluie que les données des géopotentiels du jour à 00 h.

Nous avons ensuite comparé ces 3 essais aux résultats obtenus avec la méthode S-8CP (sélection à partir de 24 CP à 00 h) pour nos 6 groupements témoins (figure III-12 et annexe III-13):

- *En rajoutant les données de la veille (essai i*) les résultats obtenus ne sont pas meilleurs. Elles n'apportent donc pas plus d'information pour ce qui concerne la pluie. Elles sont sans doute trop éloignées de la pluie elle-même et ne sont pas, à proprement parler, des variables d'évolution qui pourraient apporter une réelle information d'évolution temporelle de l'atmosphère.

- Par contre *l'introduction des CP du lendemain (essai ii)* apporte un gain non négligeable (de 0.6 à 2% selon les bassins) sur l'indice de réussite pour 8 CP retenues. La pluie est beaucoup mieux encadrée et donc mieux appréhendée. Cependant, il faut noter que les CP à 24 h utilisées dans cet essai sont calculées à partir des observations. Ce sont donc des prévisions parfaites et le gain sera sûrement moindre avec les prévisions en sortie de modèle, même si pour les géopotentiels les prévisions sont relativement fiables (cf. chapitre VI).

- Enfin, *avec les seules CP à 24 h* les résultats sont à peu près équivalents à ceux de l'essai précédent lorsque le palier est atteint. Mais ils sont encore plus sujets à la qualité de la prévision des CP.

C'est donc l'essai avec les **24 CP à 00 h et les 24 CP à 24 h** que nous retiendrons. Cependant, il est à noter que dans ce cas, le palier semble se situer plutôt autour de **12 CP** retenues qu'autour de 8 car un supplément d'information est entré en jeu.

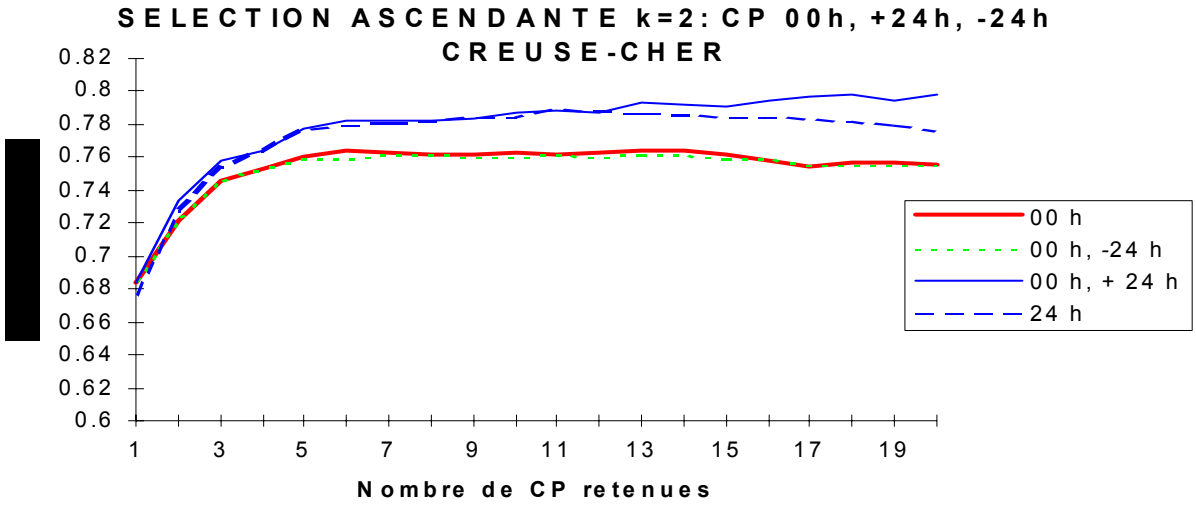


figure III-12: Courbes d'indice de réussite pour les CP à différentes échéances

III.3 Pondération

Dans la suite logique du paragraphe précédent, après nous être penchés sur le problème du choix des Composantes Principales (CP) à introduire dans le critère de sélection des analogues, nous avons envisagé de moduler l'influence de ces CP, et donc de les pondérer. En effet, comme nous le disions au début du paragraphe § III.2, il n'est pas évident que toutes les CP soient informatives de la même manière pour prévoir les pluies.

Mais, avant de passer à la pondération des CP dans le critère de sélection des analogues, nous avons appliqué la même idée à une pondération des analogues, avant de faire le calcul de la prévision de pluie. En effet, un analogue proche doit plus influer sur la prévision qu'un analogue lointain, assez médiocre.

III.3.1 Pondération des analogues

Nous avons voulu inclure ici un paragraphe un peu à part, puisque pour améliorer la prévision, nous allons abandonner la sélection de variables pour nous consacrer aux analogues eux-mêmes.

L'idée est la suivante: les 50 analogues sélectionnés pour la prévision de pluie sont plus ou moins proches de la journée à prévoir C (distance $D^2(J,C)$). Or, normalement, plus ils sont proches - c'est-à-dire plus la distance est faible - meilleurs ils sont. Il faudrait donc donner plus de poids aux analogues les plus proches afin de mieux prévoir la situation pluvieuse qui en découle.

III.3.1.a Méthode de pondération

Nous avons essayé de donner différentes sortes de poids aux analogues:

i) poids de 1 à tous les analogues (ce que nous faisions jusqu'à maintenant),

ii) poids de $1/D^2$,

iii) poids de $1/D$,

où D^2 est la distance euclidienne entre le jour J et la journée test C (notation simplifiée de $D^2(J,C)$). Une pondération de 1 au meilleur analogue et 0 aux autres a même été envisagée mais nous avons montré son inutilité lors de l'optimisation du nombre d'analogues (N=1). Cela aurait apporté un gain significatif si notre analogie était presque parfaite, ce qui est loin d'être le cas.

Les pondérations sont calculées de la manière suivante (cf. tableau III-4), en classant les 50 analogues par ordre décroissant selon leur distance D^2:

	Distance	**Poids (nombre entier)**
analogue 1	D^2_1	p_1
analogue J	D^2_J	p_J
analogue 50	D^2_{50}	$p_{50} = 1$

tableau III-4: pondération des analogues

Un poids minimal de 1 est donné à l'analogue le plus mauvais (n°50). Puis le calcul du poids p_J de l'analogue J se fait de la manière suivante:

pour une pondération en $1/D^2$, $$p_i = n\,int\left(\frac{p_{50} * 1/D_i^2}{1/D_i^2}\right) = n\,int\left(\frac{D_{50}^2}{D_i^2}\right) \quad \text{(III-8)}$$

pour une pondération en $1/D$, $$p_i = n\,int\left(\frac{p_{50} * 1/D_i}{1/D_{50}}\right) = n\,int\left(\frac{D_{50}}{D_i}\right) \quad \text{(III-9)}$$

où « n int » veut dire que l'on prend l'entier le plus proche.

Le nombre final d'analogues n'est plus N = 50 mais la somme des poids des analogues:

$$N' = \sum_{i=1}^{N} p_i \quad \text{(III-10)}$$

Et la prévision en pluie/non pluie est effectuée avec ce nouveau nombre **N'** d'analogues.

III.3.1.b Choix de la pondération

Nous avons effectué une sélection ascendante K=2 sur les 24 CP à 00 h et les 24 à 24 h (encore calculées par ACPP de covariance) pour les 6 bassins témoins et les 3 pondérations énoncées ci-dessus.
Nous avons porté sur un même graphe (cf. figure III-13 et annexe III-14) l'indice de réussite en fonction du nombre de CP retenues pour les 3 différentes pondérations et ceci pour chaque bassin.

Les 2 types de pondération (en $1/D^2$ et $1/D$) donnent des résultats pratiquement similaires et sont un peu meilleurs que l'essai sans pondération (essai i).

Nous avons finalement choisi de conserver la pondération en $1/D$ qui semble être un bon compromis. En effet, par rapport à la pondération de 1, elle possède l'avantage de donner plus de poids aux meilleurs analogues, donc de privilégier l'analogie par rapport à la climatologie. Et par rapport à la pondération en $1/D^2$, cela permet de mieux répartir la pondération.

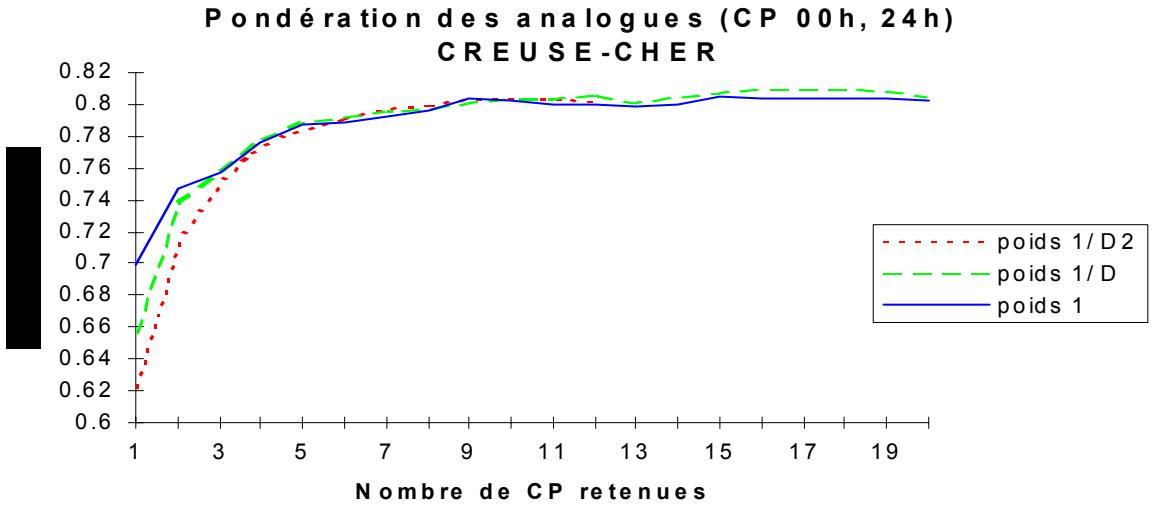


figure III-13: comparaison des pondérations

III.3.1.c Méthode S-12 CP

Avant de passer à la pondération des CP, et pour terminer cet ensemble d'essais concernant les variables et leur sélection ainsi que la pondération des analogues, nous avons lancé une sélection regroupant tous nos meilleurs résultats obtenus jusque-là, à savoir:

- une **sélection ascendante K=2**, lancée bassin par bassin,
- à partir des **12 CP des champs 700 et 1000 hPa à 00 h et à 24 h**,
- les CP étant **normées** et calculées par **ACPP de corrélation**.

Cela nous a permis d'aboutir à une distance euclidienne avec **12 CP** pour extraire les analogues qui sont **pondérés en 1/D** pour le calcul de la prévision.

C'est ce que nous appelerons par la suite la **méthode S-12CP**.

Sur les 33 bassins, l'indice de réussite moyen est passé de:

- **72.2%** pour la méthode de référence
- à **78,6%** (méthode S-12CP) rien qu'en utilisant la même information disponible.

Et si l'on compare les résultats de cette méthode à ceux obtenus avec la seule sélection des variables (méthode S-8CP) le gain moyen est de l'ordre de 3% (cf. tableau en annexe III-15).

Après la sélection des CP, vient naturellement se poser le problème de la pondération de celles-ci. En effet, il n'est pas du tout évident que les CP utiles à la prévision de pluie le soient toutes avec un même poids. C'est ce que nous verrons dans le paragraphe suivant.

III.3.2 Pondération des CP **

Synthèse : Toujours dans le but d'optimiser la sélection des analogues, différentes méthodes de pondération et/ou de sélection des variables ont été mises en place. Cependant, si un gain en prévision a pu être apporté avec la sélection des variables, leur pondération, dans la distance euclidienne n'a pas été retenue (gain inférieur à 1% en moyenne sur les 33 bassins).

III.3.2.a La méthode de pondération

La méthode choisie reprend le principe de la sélection ascendante aussi l'appellerons-nous par la suite "pondération ascendante".
En fait, on procède comme pour la sélection ascendante (cf. § III.2.1) mais au lieu d'introduire une variable avec une pondération de 1 comme c'est le cas en sélection, on va, à chaque pas tester différentes pondérations pour chaque variable (cf. fig. III-14).

Pas 1: pour chaque variable i=1 à P la sélection des analogues se fait grâce à la distance:

$$D^2(J,C) = p \,.\, \left[X_i(J) - X_i(C)\right]^2 \qquad \text{(III-11)}$$

où p est une pondération choisie arbitrairement.
Pour ce premier pas où une seule variable intervient dans la distance, la pondération ne fait que translater la distance D^2 d'une valeur p. Donc les 50 meilleures distances déterminant les analogues restent les mêmes quelque soit la pondération. Par contre le choix de p sera déterminant pour la suite car les résultats diffèreront aux pas suivants.
De la même manière qu'au § III.3.1, l'essai avec le meilleur critère de performance est conservé et noté (variable i_1, pondération p_1).

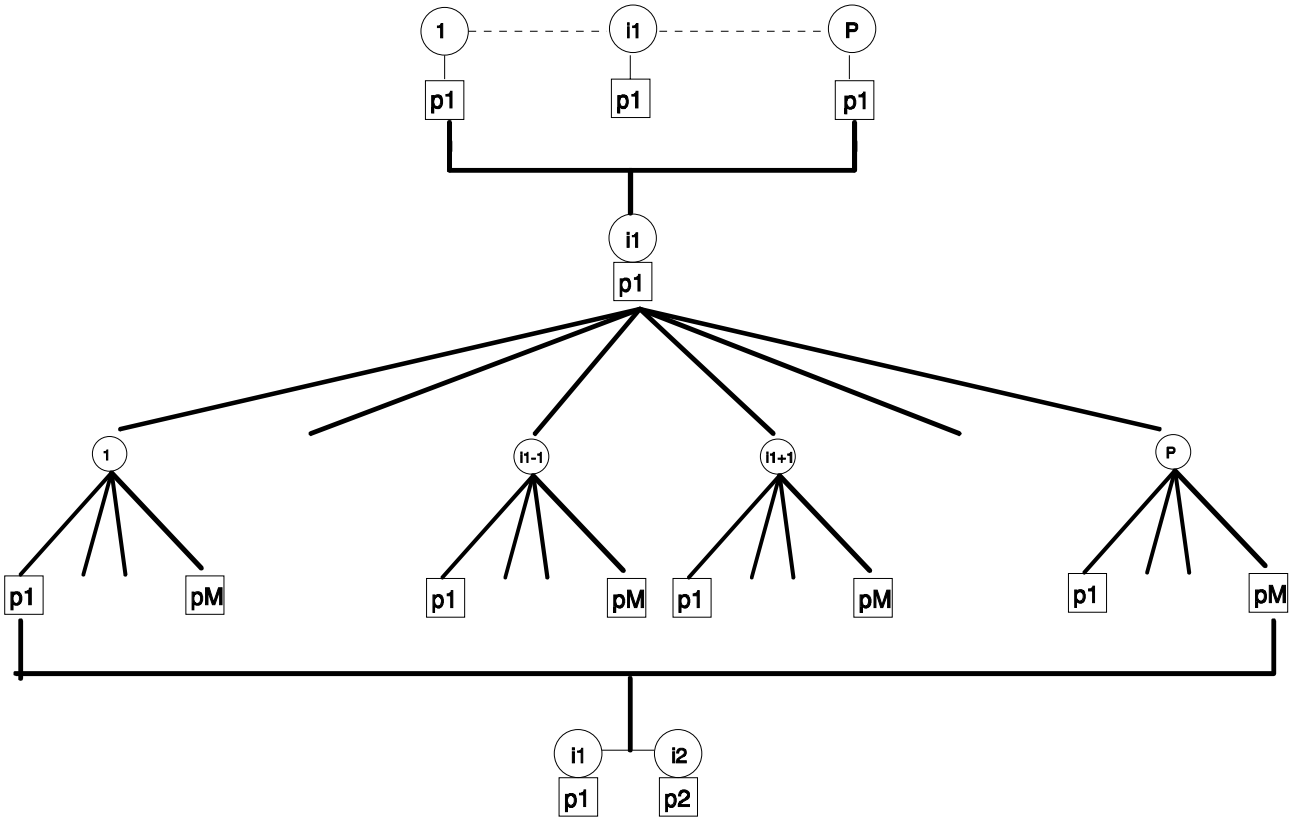


figure III-14: pondération ascendante

Pas 2: pour chaque variable i = 1 à P, $i \neq i_1$, M pondérations différentes sont testées, soit pour m = 1 à M:

$$D^2_{im}(J,C) = p_1 \cdot \left[X_{i_1}(J) - X_{i_1}(C)\right]^2 + p_m \cdot \left[X_i(J) - X_i(C)\right]^2 \qquad \text{(III-12)}$$

variable + pondération choisie au pas 1 — variable + pondération à tester

M x (P-1) essais sont effectués et l'on retient le couple (variable i_2,pondération p_2) qui, associé à (i_1,p_1) donne le meilleur critère de performance.

Les pas suivants se déroulent de la même manière. Les pondérations peuvent être négatives tant que la pondération totale de chaque variable reste positive ou nulle.

Remarques:

① On peut envisager de commencer cette pondération ascendante à partir des résultats obtenus par sélection seule des variables.

Dans ce cas là, on aurait au pas 1, pour i = 1 à P et pour m = 1 à M et avec des pondérations négatives quand cela sera possible:

$$D^2_{im}(J,C) = D^2_o(J,C) + p_m \cdot \left[X_i(J) - X_i(C)\right]^2 \qquad \text{(III-13)}$$

distance initiale — pondération et sélection à tester

avec $$D^2_o(J,C) = \sum_{i=1}^{NCP} \left[X_i(J) - X_i(C)\right]^2 \qquad \text{(III-14)}$$

où les NCP variables sont les CP retenues par sélection.

② Nous avons présenté la méthode de pondération ascendante simple K=1, où à chaque pas un seul couple (variable, pondération) est conservé. Mais, comme pour la sélection ascendante, nous en garderons 2 (K=2).

③ Les pondérations testées seront :

- pour les pas supérieurs à 1: 0.2, 0.4, 0.6,
- et quand cela sera possible -0.2, -0.4 et -0.6.

④ Une méthode de pondération au hasard a été testée mais peu intéressante et très lourde en temps de calcul elle a rapidement été abandonnée.

III.3.2.b Choix de la méthode la plus performante

En raison des temps de calcul élevés et dans le but de choisir une méthode de pondération (avec ou sans distance initiale), nous avons, dans un premier temps, effectué une première série d'essais en utilisant uniquement les 24 CP à 00 h calculées par ACPP de corrélation, sans pondération des analogues.

Pour 3 des 6 groupements test, 4 essais de pondération ont été effectués:

i) pondération ascendante en partant d'une distance initiale nulle ($D^2_O(J,C)=0$) et d'une pondération de 0.2 au pas 1 (essai POND0.2),
0.4 ---------- (essai POND0.4),
0.6 ---------- (essai POND0.6),

ii) pondération ascendante en partant d'une distance initiale $D^2_O(J,C)$ (essai PONDREF), calculée avec les 8 CP retenues par une sélection ascendante K=2 sur les 24 CP à 00 h calculées avec la matrice de corrélation (méthode S-8CP avec les CP de corrélation),

La figure III-15 montre l'évolution de l'indice de réussite en fonction du nombre de pondérations (ou de pas) effectués pour ces 3 bassins. Nous y avons rajouté les résultats de la méthode S-8CP avec les CP de corrélation. Pour cette courbe-là, les 8 premiers pas correspondent aux nombres de CP sélectionnées et les suivants aux pondérations de l'essai PONDREF.

Si les essais i n'apportent pas, dans l'ensemble, d'amélioration significative par rapport à la seule sélection, par contre, l'essai ii, partant des résultats de la sélection, apporte un mieux intéressant. C'est pourquoi nous avons choisi d'explorer plus en détail cette dernière pondération sur l'ensemble des bassins disponibles.

III.3.2.c Application sur les 33 bassins

Nous avons voulu effectuer la pondération ascendante à partir de la meilleure sélection possible, c'est-à-dire la méthode S-12CP.
Un exemple de courbe d'indice de réussite est donné sur la figure III-16: le trait plein correspond à l'indice de réussite en fonction du nombre de CP retenues pour la méthode S-12CP et les pointillés, à partir de 12 CP, représentent l'indice de réussite en fonction du nombre de pondérations effectuées.

Le nombre d'itérations varie de 4 à 10 suivant les bassins mais les gains par rapport à la méthode S-12CP ne sont pas très importants (cf. tableau en annexe III-16); ils sont inférieurs à 1% en moyenne. Nous n'envisageons donc pas de conserver cette idée de pondération, du moins pour la prévision catégorique pluie / non pluie.

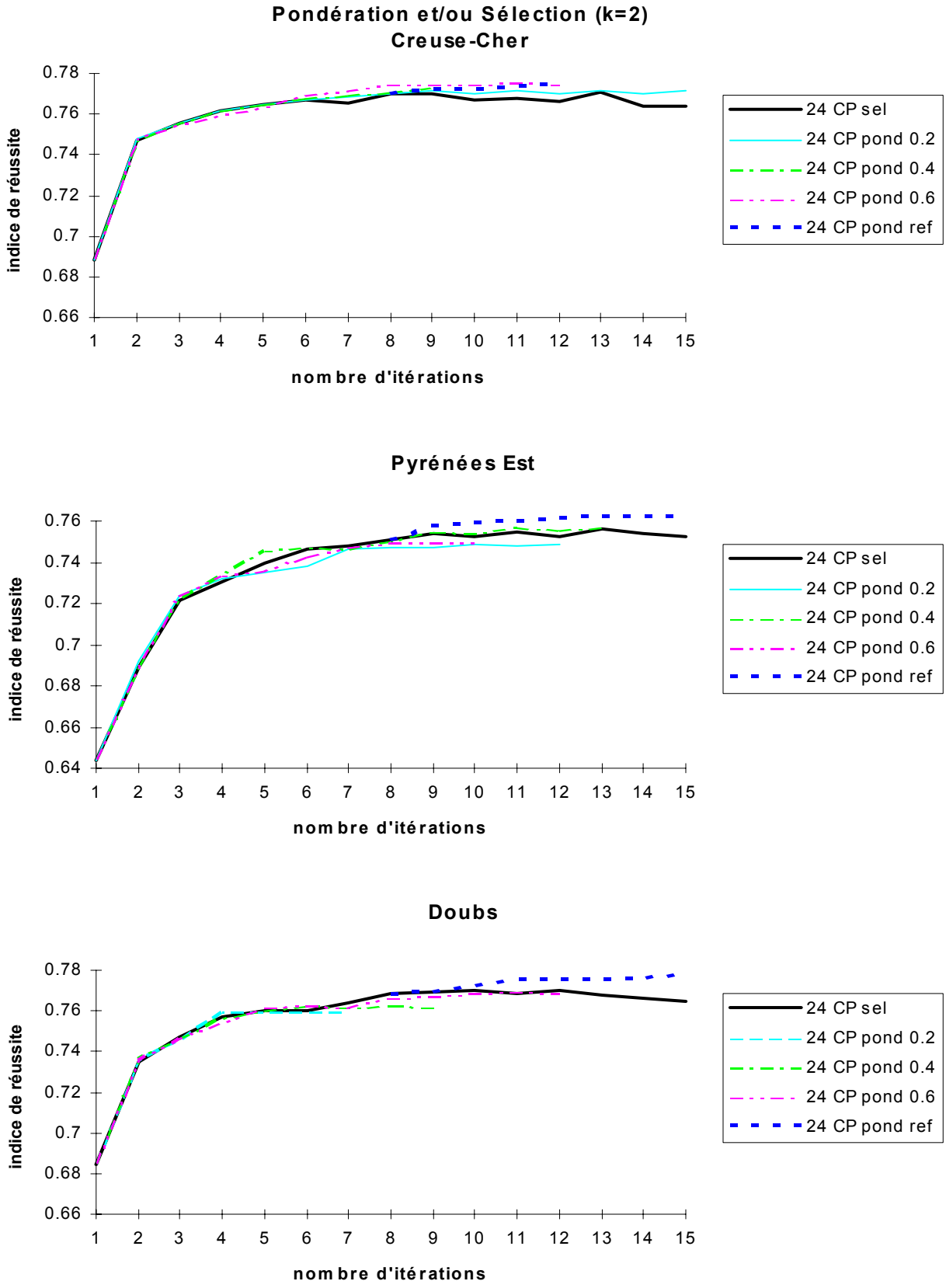


figure III-15: comparaison des différentes pondérations

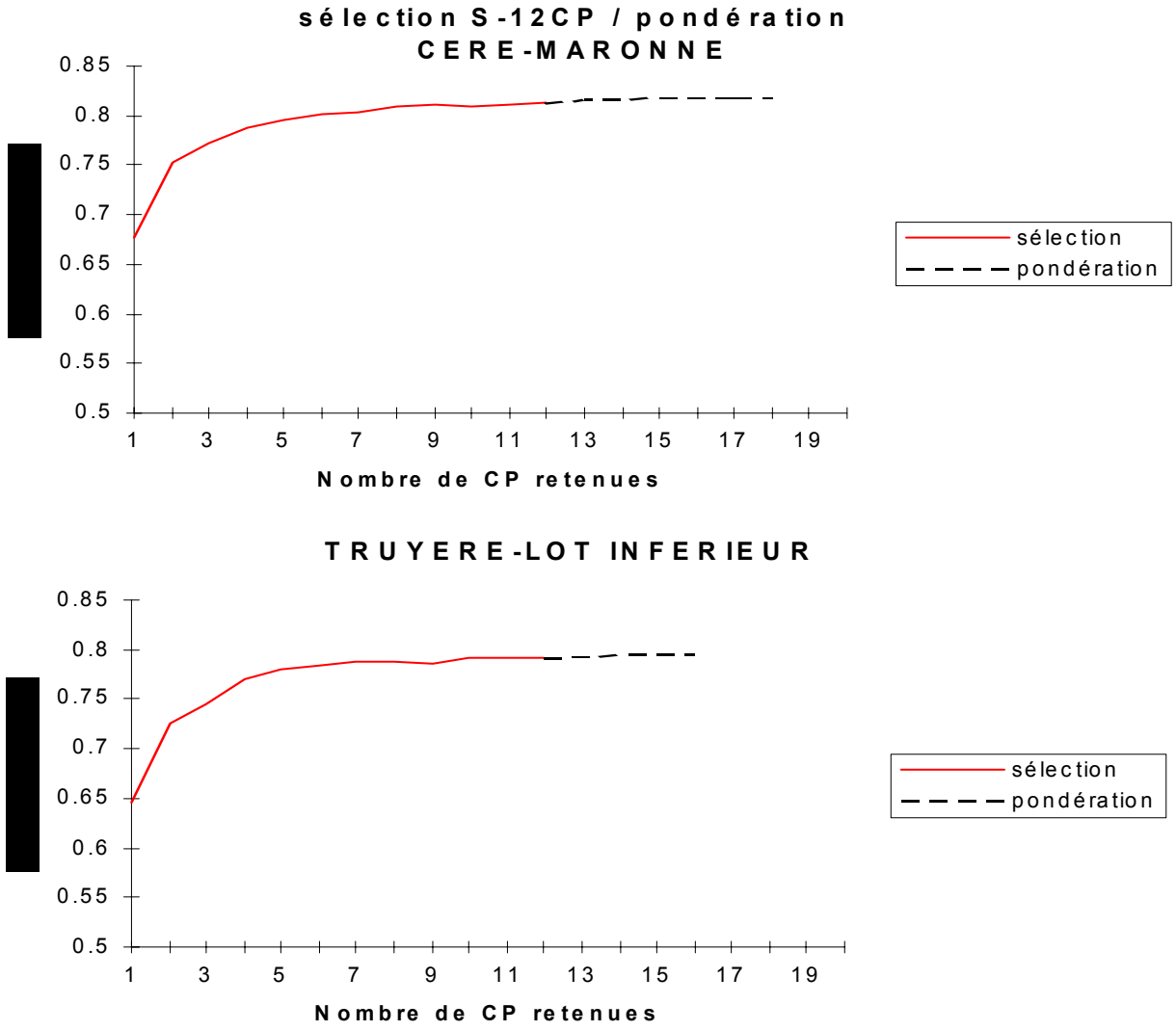


figure III-16: sélection et pondération pour la méthode S-12CP

III.4 Application des techniques précédentes à la prévision probabiliste en classes de pluie

Pour notre étude, nous avons commencé à travailler en pluie / non pluie car il était très simple de mesurer la performance d'une méthode grâce à l'indice de réussite IR (rapport entre le nombre de bonnes prévisions et le nombre de prévisions effectuées). De plus, comme sur l'ensemble des bassins étudiés, environ la moitié des journées sont sans pluie, si l'on arrive à bien prévoir

l'occurrence de pluie, ce n'est pas négligeable. D'ailleurs, dans cette optique, nous sommes arrivés à la **méthode S-12CP** qui permet de prévoir l'occurrence de pluie à presque 80%.
Cependant, cette prévision n'est pas suffisante aussi sommes-nous passés à une prévision plus quantitative: une prévision probabiliste en 8 classes.

III.4.1 Détermination du nombre N d'analogues

Dans un premier temps, nous avons voulu vérifier que, comme pour la prévision pluie / non pluie, le nombre optimal d'analogues à retenir se situe autour de 50. Nous avons donc lancé, en prévision probabiliste, le même essai de sélection qu'au §III.2.2c, à savoir la sélection ascendante K=1 sur 24 CP pour les 6 bassins témoins. Et nous avons pu voir que le nombre N d'analogues à retenir ne dépendait pas du type de prévision (cf. annexe III-17).

III.4.2 Sélection : méthodes S-8CP et S-12CP

Ensuite, nous avons appliqué les sélections ascendantes aboutissant aux méthodes S-8CP (avec les CP calculées par la matrice de corrélation) et S-12CP à la prévision probabiliste sur les 6 groupements témoins, afin de voir si, là aussi, elle réagissait de la même manière que la prévision catégorique (cf. figure III-17 et annexe III-18).

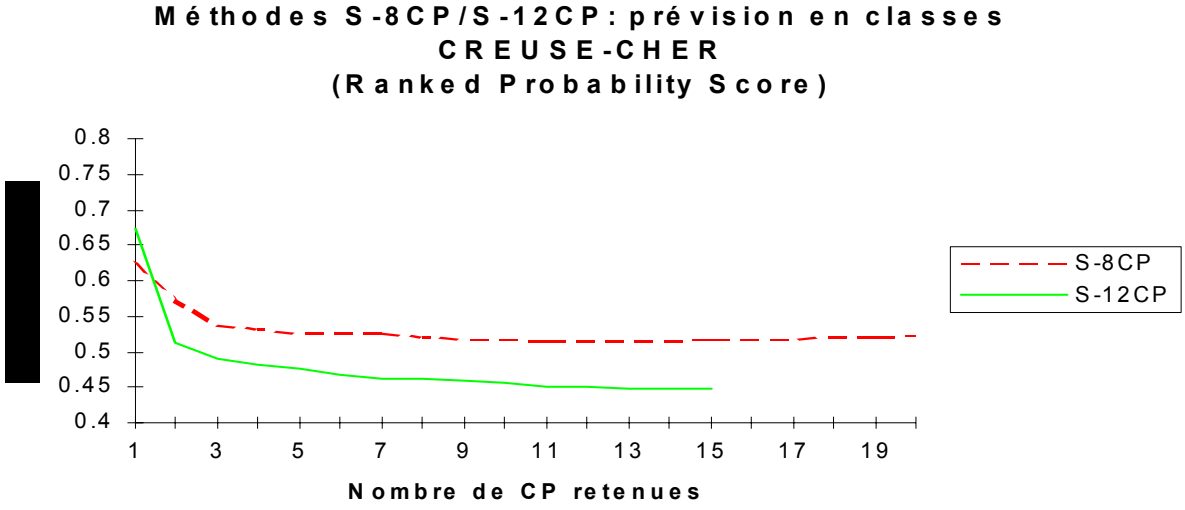


figure III-17: méthodes S-8CP et S-12CP pour la prévision probabiliste

Nous pouvons voir:

- qu'un palier s'établit toujours vers 8 CP pour la méthode S-8CP et un peu plus loin, vers 12 CP, pour la méthode S-12CP,
- qu'à partir de 2 CP retenues, le score est meilleur pour la méthode S-12CP.

III.4.3 Sélection et pondération

Comme la méthode S-12CP apporte un gain non négligeable par rapport à la méthode S-8CP, nous l'avons appliquée à l'ensemble des 33 bassins. Et c'est à la suite de celle-ci que la pondération ascendante K=2 (PONDREF) a été testée. Sur la figure III-18, on trouvera un exemple de la courbe du Ranked Probability Score pour la sélection et la pondération. Et en annexe III-19, sont consignés les résultats pour tous les groupements, à savoir le RPS pour:

- la méthode de référence,
- la méthode S-12CP,
- la pondération après S-12CP ou méthode P-12CP
- et les prévisions de référence: la climatologie, la persistance et le hasard.

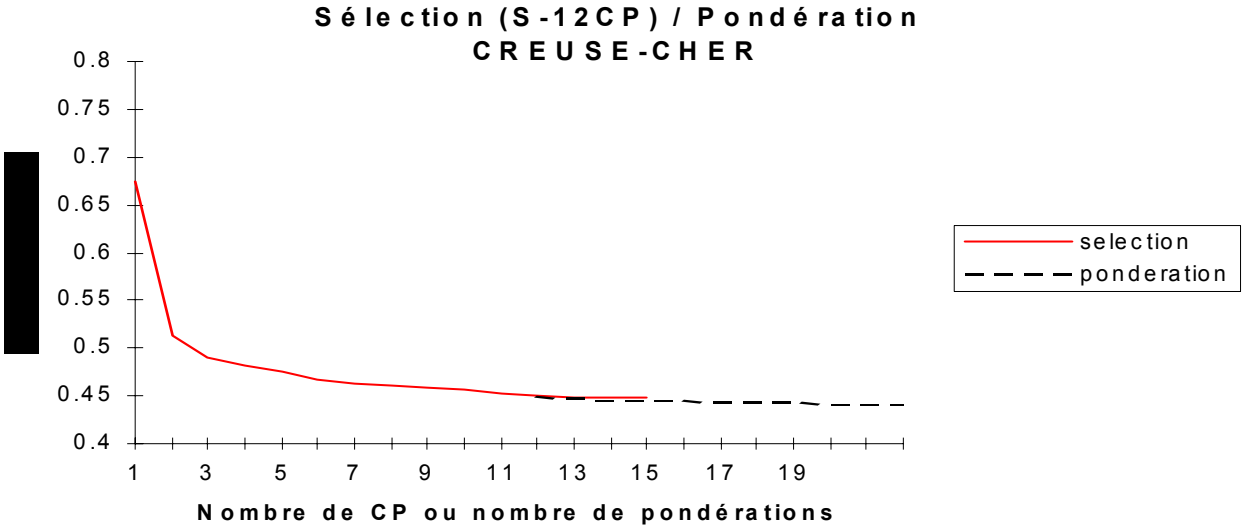


figure III-18: pondération et sélection pour la prévision probabiliste en classes

Cette méthode S-12CP est aussi meilleure que la méthode de référence pour la prévision probabiliste: le RPS a diminué de 0.1 en moyenne sur les 33 bassins. Par contre, tout comme la

prévision catégorique, la pondération des CP s'est révélée beaucoup moins intéressante: baisse du RPS de 0.01 seulement par rapport à S-12CP.

III.4.4 Synthèse des résultats pour les 2 types de prévision

III.4.4.a En terme de score de réussite

Si l'on regarde maintenant la performance moyenne sur les 33 bassins (cf. tableau III-5), on remarque que la pondération des CP, que ce soit en prévision catégorique ou probabiliste, ne donne pas de résultats intéressants par rapport aux gains déjà obtenus. Dans les 2 cas, on en restera donc à la **méthode S-12CP**.

	Prévision pluie/non pluie IR	Prévision en classes RPS * 100
méthode de référence	**72.2**	**59.7**
méthode S-12CP	78.6	49.0
méthode P-12CP	79.5	47.8

tableau III-5: performance des différentes méthodes pour les 2 types de prévision

III.4.2.b En terme de CP retenues

Nous avons comparé les 12 premières CP retenues pour les 2 types de prévision. En effet, il n'est pas évident que pour prévoir l'occurence de pluie (prévision catégorique) ou la quantité de pluie (prévision probabiliste), les mêmes CP soient utiles.

On trouvera en figure III-19, pour les 2 prévisions, les histogrammes donnant le nombre de fois où les différentes CP sont sorties (Z_1 à Z_{12} et S_1 à S_{12} à 00 et 24 h), sur l'ensemble des 12 CP retenues pour les 33 bassins, soit sur 33*12=296 CP.

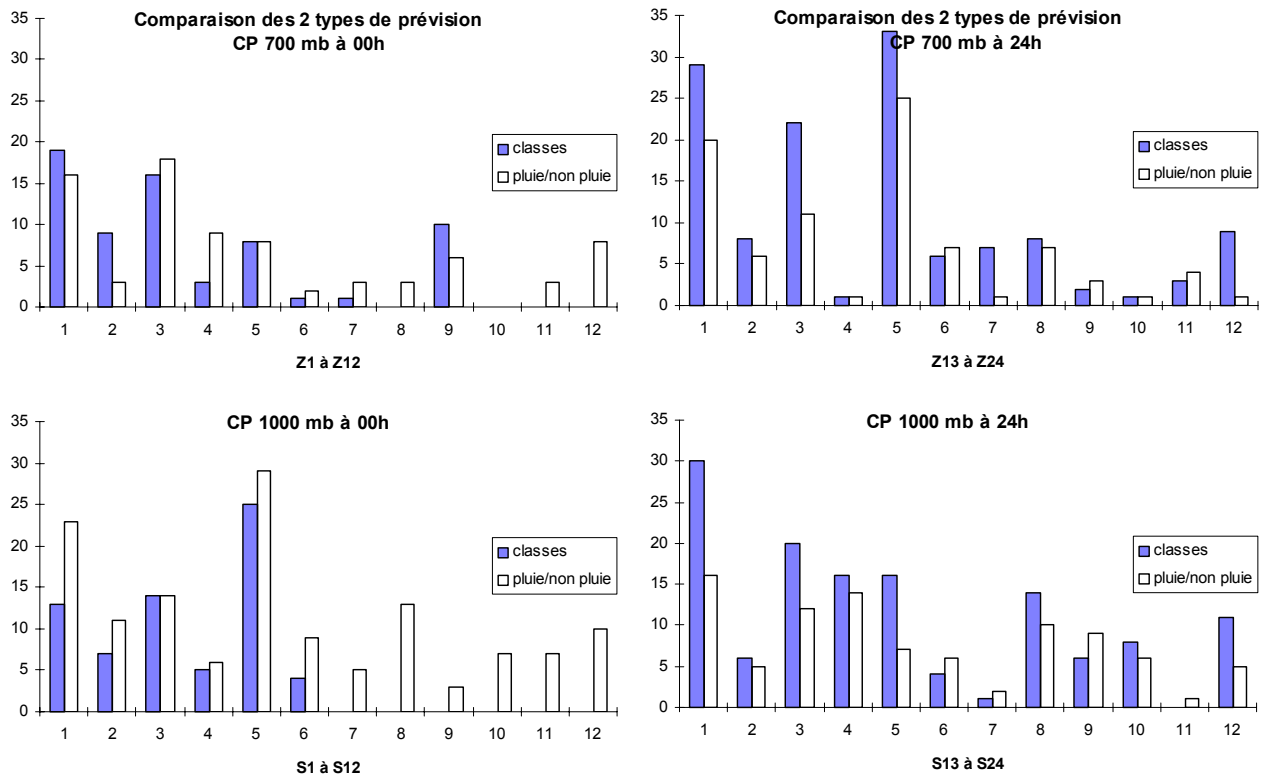


figure III-19: histogramme des 12 CP retenues pour les deux types de prévision

Leur allure est assez similaire pour les 2 prévisions, avec les CP n°1, 3 et 5 plus souvent sélectionnées ainsi que les CP à 24 h.

Nous avons ensuite condensé les résultats dans le tableau III-6: il donne, pour les 2 prévisions, le nombre de fois où les CP S et Z à 00 et 24 h sont sorties en terme de pourcentage. On remarque que, pour les 2 prévisions, les pourcentages sont à peu près équilibrés au niveau des CP du champ 700 hPa (Z) et 1000 hPa (S) sélectionnés: 55/45% et 50/50% en prévision pluie / non pluie et en classes). Par contre, le pourcentage de CP à 00 h et celui des CP à 24 h sélectionnées est plus déséquilibré, surtout pour la prévision probabiliste où les CP à 24 h apparaissent deux fois plus souvent que celles à 00 h (34/66%).

	prévision catégorique	prévision probabiliste
Z_1 à Z_{12} à 00h	20 %	17 %
Z_1 à Z_{12} à 24h	35 %	33 %
S_1 à S_{12} à 00h	22 %	17 %
S_1 à S_{12} à 24h	23 %	33 %
Z et S à 00h / à24h	42 / 58 %	34 / 66 %
Z / S à 00h et 24h	55 / 45 %	50 / 50 %

tableau III-6: % des différentes CP retenues pour les 2 types de prévision

III.5 Conclusion du chapitre III

Dans ce chapitre, nous avons cherché à explorer au maximum les capacités potentielles de l'information disponible sous forme de Composantes Principales (CP). Pour cela, nous avons sélectionné et pondéré les CP afin de mieux extraire les analogues et donc mieux prévoir la pluie.

Si la pondération des CP utilisées dans le critère de sélection des analogues n'a pas apporté les résultats espérés, leur sélection, effectuée bassin par bassin, s'est, elle, révélée intéressante pour la prévision d'occurence de pluie et pour la prévision probabiliste en classes.
Quelques modifications supplémentaires ont permis de gagner encore en performance :

i) l'utilisation des **CP normées,** calculées par **ACP de Processus** et avec la matrice de **corrélation**,

ii) la possibilité de sélectionner les **CP à 00 et 24 h**,

iii) la **pondération des analogues** en 1/D.

Cela nous a permis d'aboutir à une nouvelle méthode de prévision de pluie optimisée bassin par bassin, appelée **méthode S-12CP**. Elle possède comme critère de sélection des analogues:

- une distance euclidienne utilisant 12 Composantes Principales,
- retenues par sélection ascendante K=2,
- sur l'ensemble des 12 premières CP des champs de géopotentiels 700 et 1000 hPa à 00 et 24 h, soit sur 48 CP.

Et la prévision de pluie est calculée en donnant un poids de 1/D aux analogues, D^2 étant la distance euclidienne de l'analogue en question à la journée courante.

En moyenne sur l'ensemble des 33 bassins, pour la prévision d'occurence de pluie, l'indice de réussite est passé de 72.2 pour la méthode référence à 78.6% soit un gain de 6.5%. Quant à la prévision probabiliste en classes de pluie, elle a vu son score RPS (x 100) diminuer de 59.7 à 49, ceci grâce à un travail systématique de sélection des CP les plus adaptées à la sélection des analogues pour prévoir la pluie.

Aussi, sachant que la partie non expliquée de la prévision était :

- pour le score IR de 100 - 72.2 = 27.8,
- pour le score RPS de 0 - 59.7 = - 59.7,

on l'a réduite :

- pour IR de 6.5 / 27.8 = **23 %**
- pour RPS de 10.7 / 59.7 = **18 %.**

Nous avons donc fait le tour des méthodes de sélection et de pondération des différentes CP disponibles. Et finalement, reste une question primordiale:

Pourquoi, dans la sélection des analogues, utilise-t-on comme prédicteurs les CP plutôt que les données brutes?

Cela pouvait s'expliquer il y a 25 ans (Duband, 1974) où les ordinateurs de l'époque rendaient la réduction de l'information absolument indispensable. Il y a 15 ans (Duband, 1981) cela se comprenait encore, les calculateurs n'étaient pas encore aussi puissants que maintenant.

Mais de nos jours, ce problème a considérablement reculé. Par conséquent, on peut envisager de travailler sur des données brutes (ou du moins plus nombreuses car les données que nous possédons ont déjà été filtrées) et non plus condensées.

CHAPITRE IV :

METHODES BASEES

sur les

DONNEES BRUTES ou

INTERPOLEES

Introduction

Jusqu'à présent, l'information contenue dans les champs journaliers de géopotentiel 700 et 1000 hPa était utilisée sous forme de Composantes Principales. Cependant, il pourrait être envisagé de l'utiliser « brute », sous sa forme initiale, à savoir les données journalières aux 37 stations de radiosondage de ces deux champs. C'est ce que l'on appellera les *données brutes* car elles sont censées n'avoir subi aucun traitement. Ce n'est pas tout à fait vrai vu que, de nos jours, un certain nombre de ces radiosondages n'existent plus. Par conséquent, les valeurs en ces stations de radiosondage sont maintenant toutes reconstituées par interpolation des grilles analysées du centre français et du centre européen par Météo-France. Malgré cela, elles seront quand même répertoriées sous le nom de données brutes, par opposition aux Composantes Principales (CP) qui proviennent d'une condensation de l'information contenue dans ces données brutes.

Enfin, il existe une alternative pour homogénéiser la forme des données avec celles qui proviennent des sorties de modèles (grilles) : une interpolation de ces données brutes sous forme de *grilles* de valeurs. Notons que ces grilles diffèrent des grilles analysées par un modèle (par exemple celles du CEPMMT, Centre Européen pour la Prévision Météorologique à Moyen Terme, disponibles depuis 1981 seulement), mais elles possèdent l'avantage de pouvoir être élaborées sur l'ensemble de notre fichier.
Dans ce cas, l'utilisation d'une distance euclidienne pour sélectionner les analogues n'est peut-être pas optimale. C'est pourquoi une certain nombre de critères d'analogie, répertoriés dans la littérature, ont été mis en oeuvre et testés.

Pour terminer, quelques essais comparatifs ont été effectués en utilisant une technique récente en prévision, les réseaux de neurones.

IV.1 La sélection des variables brutes (radiosondages)

Les *données brutes* correspondent aux valeurs journalières à 00 h des champs de géopotentiel 700 et 1000 hPa aux 37 stations de radiosondage (cf. figure I-1).
Avant d'effectuer une sélection sur celles-ci pour déterminer les plus utiles à la sélection des analogues, elles ont été au préalable centrées et réduites afin d'éviter de donner trop de poids aux stations à forte variance et donc pour rendre comparables les différentes stations.

IV.1.1 Sur 6 bassins témoins

Comme le nombre de variables (et les temps de calcul !) augmente considérablement quand on passe de 12 CP à 37 données par champ, un premier essai de sélection des données brutes, appelées aussi données de RadioSondage et plus simplement codées RS, a été fait sur les 6 bassins témoins du chapitre III (Creuse-Cher, Pyrénées Est, Doubs, Isère moyenne, Var-Tinee-Roya et Chassezac) et pour la prévision en pluie / non pluie.

Une sélection ascendante K=2 (les 2 meilleures variables sélectionnées à chaque pas) des RS a été effectuée sur les seules données à 00 h (37 pour le champ 700 hPa et 37 pour le champ 1000 hPa). Et la courbe de l'indice de réussite a été comparée au résultat de la méthode S-8CP, qui utilise seulement les 24 Composantes Principales (calculées par Analyse en Composantes Principales de Processus avec la matrice de corrélation) à 00 h (figure IV-1 et annexe IV-1).

On note, tout d'abord, qu'avec les données brutes (notées RS pour RadioSondage), le palier est aussi atteint aux alentours de 8 variables retenues, comme pour les CP (méthode S-8CP pour les seules données à 00h). En outre, il n'y a pas d'écart flagrant entre les 2 méthodes: si pour 3 bassins l'indice de réussite est meilleur avec les RS, pour les 3 autres (Pyrénées Est, Isère moyenne et Var-Tinee-Roya) il est plutôt moins bon.

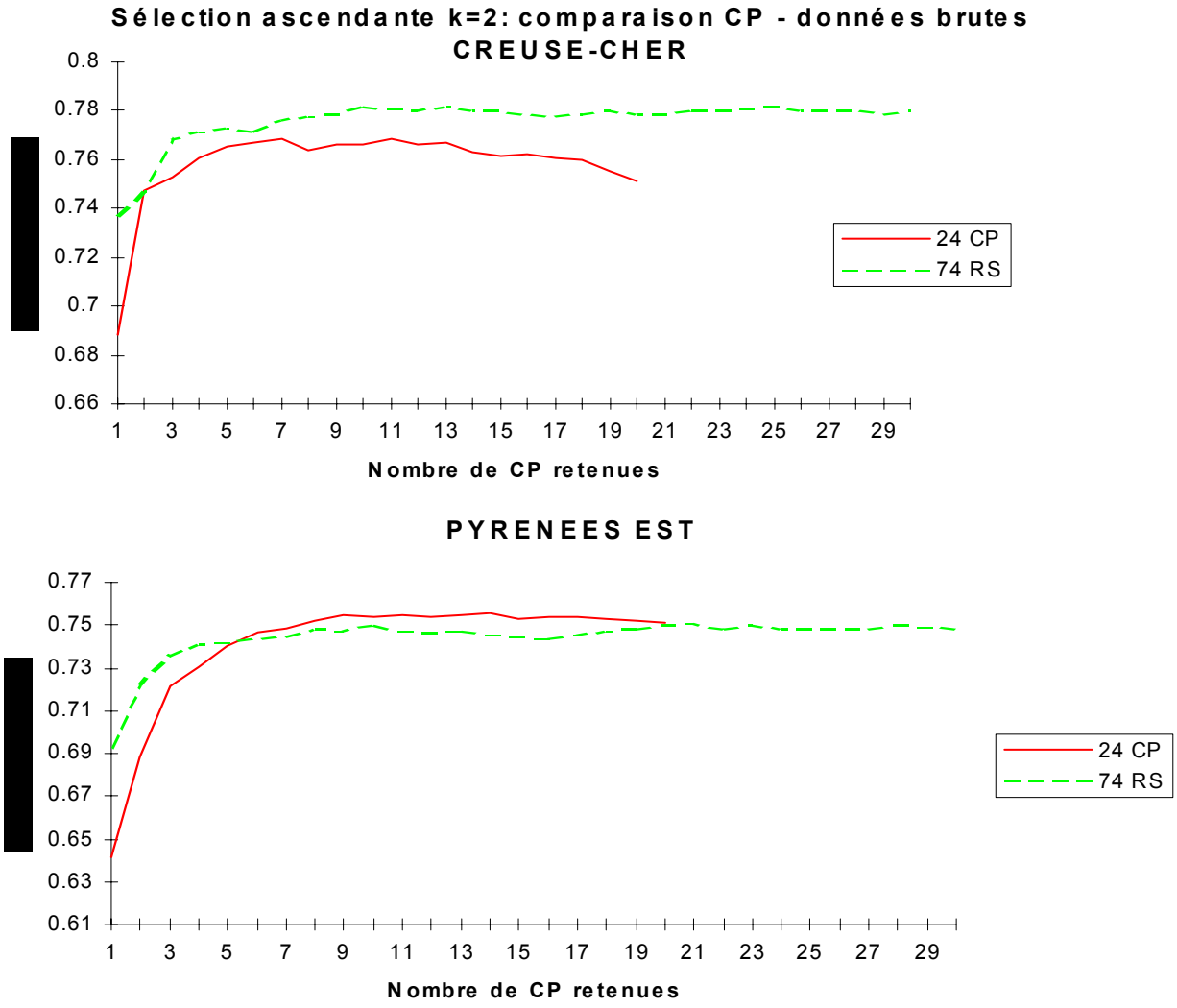


figure IV-1: comparaison CP / RS

Nous avons ensuite regardé les 8 premières stations de RadioSondage (notées RS) sélectionnées. On trouvera sur la figure IV-2 ci-dessous le nombre de fois où elles sont sorties et pour quel champ (700 ou 1000 hPa), ceci pour l'ensemble des 6 bassins (soit 6 x 8 stations):

Seules 21 d'entre elles, sur 37, apparaissent. Et encore, sur ces 21, 9 n'ont été sélectionnées qu'une fois.

Ensuite, nous avons représenté les stations de radiosondage sur la carte de l'Europe par des points plus ou moins gros suivant le nombre de fois où elles ont été utilisées (cf. fig. IV-3).

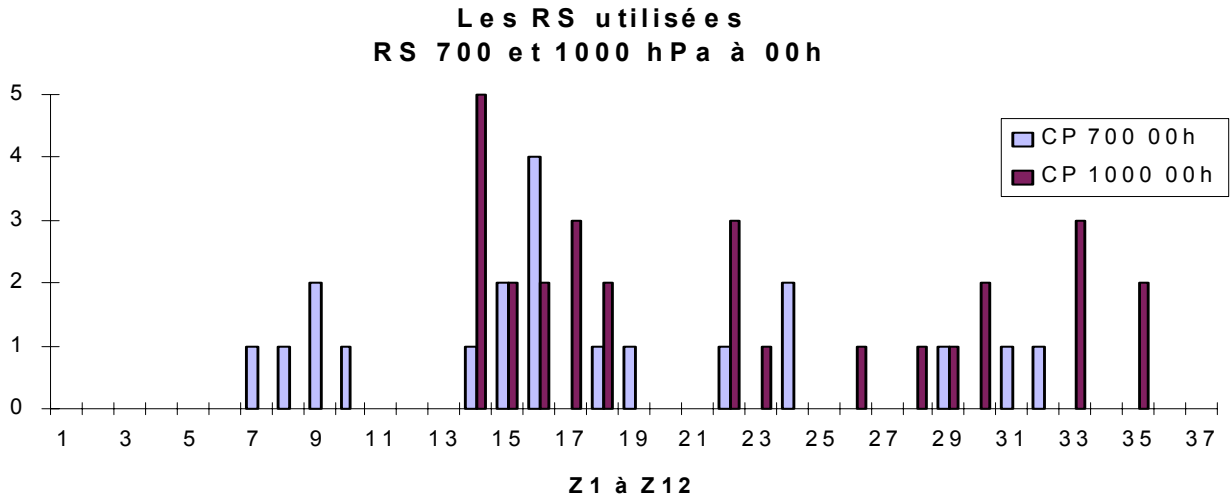


figure IV-2: histogramme des RS sélectionnées

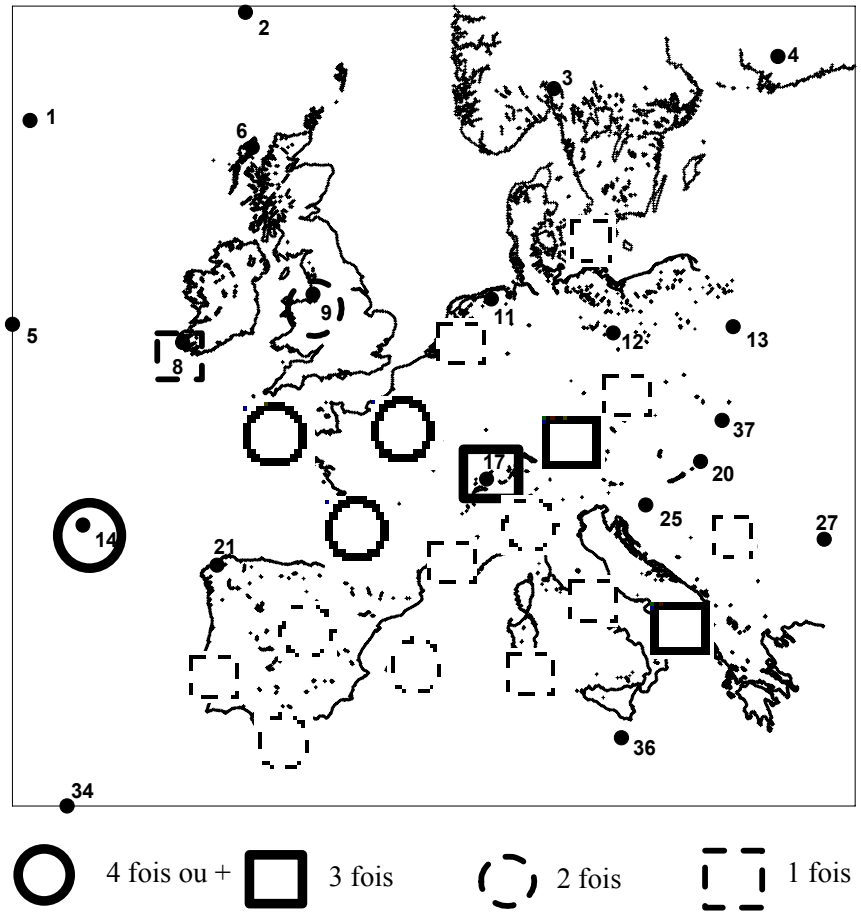


figure IV-3: carte des RS sélectionnées

On remarque que les stations du Nord et de l'Est sont peu, voire pas utilisées. Elles sont d'une part sans doute trop éloignées des bassins étudiés et d'autre part, les perturbations apportant de la pluie sur les groupements viennent surtout :

- de l'Ouest, avec la station 14 comme bon indicateur,
- et de la Méditerranée donc du Sud (coups de Sud pour les bassins méditerranéens et cévenols).

On peut noter que la station 33, au sud-est est assez souvent sélectionnée; elle est sans doute un bon indicateur pour les retours d'Est.

Ces résultats préliminaires intéressants nous ont poussé à continuer dans cette voie. Aussi avons-nous effectué un nouvel essai sur l'ensemble des bassins disponibles et pour les 2 types de prévision.

IV.1.2 Essai sur les 33 bassins

Nous sommes passés directement à notre meilleur essai, celui aboutissant à la méthode S-12CP du chapitre III, en remplaçant juste les Composantes Principales (CP) par les données brutes RS:

- sélection ascendante K=2 des données brutes des 2 champs à 00 et 24 h à introduire dans le critère de sélection des analogues $D^2_{RS}(J,C)$:

$$D^2_{RS}(J,C) = \sum_{i=1}^{N_{RS}} \left[RS_i(J) - RS_i(C) \right]^2 \qquad \text{(IV-1)}$$

avec N_{RS}, le nombre optimal de variables à retenir,

- et pondération des analogues en $1/D_{RS}(J,C)$.

D'après les résultats du paragraphe précédent, et afin de limiter les temps de calcul (environ 12 h par groupement), quelques stations de radiosondage situées au Nord (stations n°1,2,3,4,5) et à l'Est (13,27,37) ont été éliminées.

Quelques courbes de l'indice de réussite de cette méthode, notée **S-12RS**, sont fournies sur la figure IV-4 pour la prévision en pluie / non pluie. Elles sont comparées à celles obtenues par la

méthode S-12CP du chapitre III (avec les CP). Et l'on donne en annexe IV-2, pour les deux types de prévision (pluie / non pluie et probabiliste en classes) les performances des 2 méthodes (avec CP ou données brutes) pour 12 variables retenues, car même avec les données brutes un palier est atteint autour de 12 RS sélectionnées.

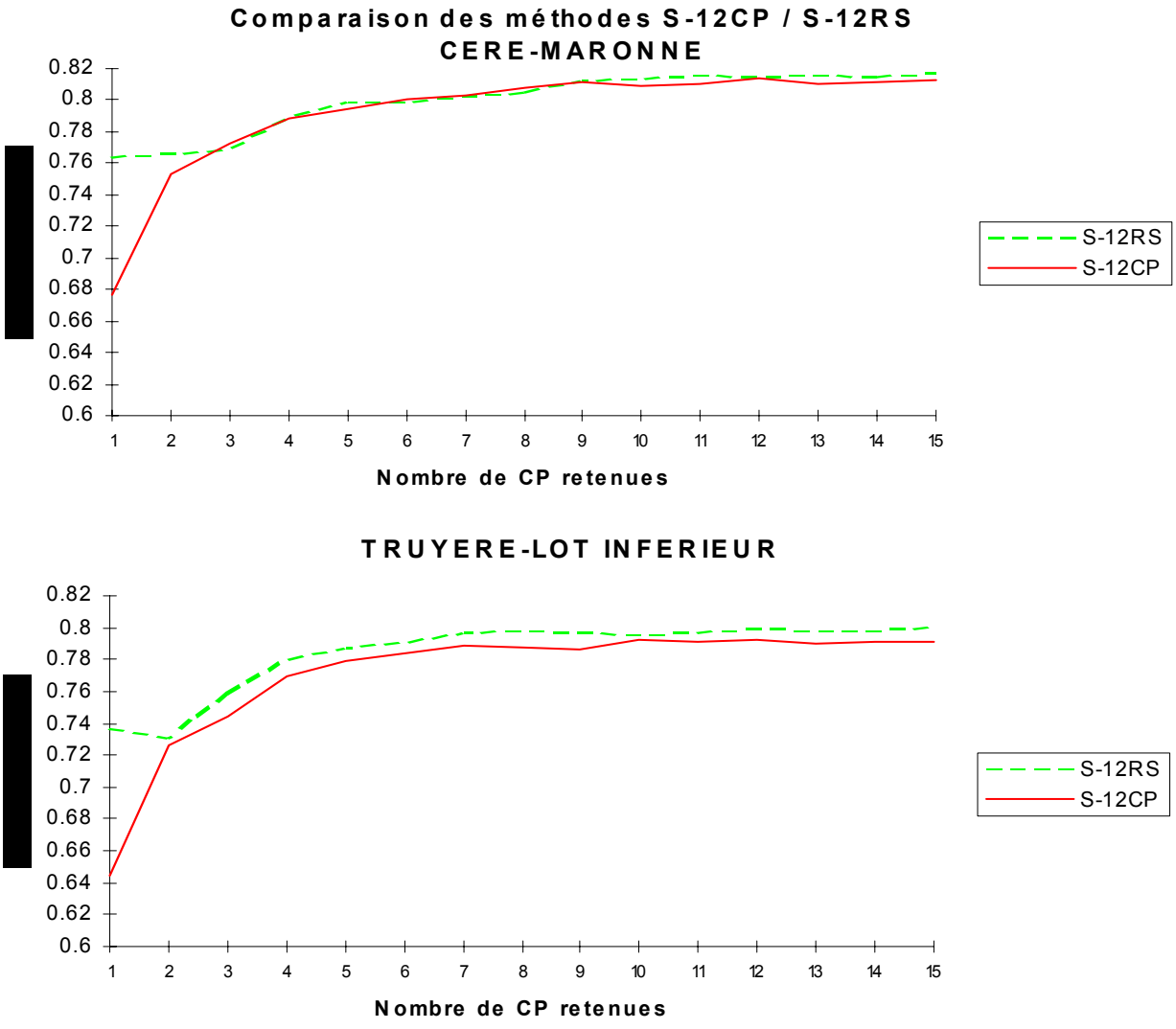


figure IV-4: comparaison des méthodes S-12 RS et S-12CP

Dans l'ensemble les résultats sont intéressants, avec un gain moyen de 0.6% sur l'indice de réussite qui passe de **78.6 à 79.2 %.** Quant au score RPS (multiplié par 100), il baisse de près de deux points (**49.0 à 47.2**). Et si pour la prévision en pluie / non pluie, les performances ont légèrement diminuées avec les données brutes pour 8 bassins, pour la prévision en classes de pluie, l'utilisation des RS donne toujours des résultats supérieurs.

En outre, nous avons vérifié que le fait de sélectionner les RadioSondages (RS) était bénéfique et nécessaire. En effet, en forçant l'utilisation de l'ensemble des 37 RS des 4 champs dans la distance euclidienne (N_{RS} = 37 x 4), les scores sont plutôt moins bons: IR = 75.3 % et RPS = 0.52.

IV.1.3 Les RS sélectionnées

On peut ensuite regarder les résultats en terme de RS retenues.

Dans le tableau IV-1, sont regroupés en pourcentage, le nombre de RS sélectionnées par champ et par échéance sur les 12 RS des 33 bassins, soit sur 33 x 12 = 396 RS.

Il est à comparer à celui obtenu avec la méthode S-12 CP (cf. tableau III-7 du chapitre III).

	prévision pluie/non pluie	prévision en classes
700 hPa à 00 h	22 %	21 %
700 hPa à 24 h	28 %	33 %
1000 hPa à 00 h	26 %	14 %
1000 hPa à 24 h	24 %	32 %
00 h / 24 h	48 / 52 %	35 / 65 %
700 / 1000 hPa	50 / 50 %	54 / 46 %

tableau IV-1: les RS sélectionnés par champ et échéance

On retrouve à peu près le même type de résultat qu'avec la méthode S-12CP, à savoir:

- un partitionnement 700 / 1000 hPa à peu près équilibré,
- les champs à 24 h plus souvent sélectionnés, surtout en prévision probabiliste.

On peut, dans un deuxième temps, se pencher sur les stations de radiosondage les plus sollicitées quels que soient le champ et l'échéance (cf. figure IV-5). Ce sont celles situées:

- autour de la zone d'étude (n°16, 17, 18, 22, 23), pour une analogie de proximité,
- au Sud, Sud-Est (n° 29, 30, 31, 32 et 36), pour les coups de Sud ou retours d'Est pouvant générer d'intenses épisodes pluvieux,
- et un peu à l'Ouest (n°14 et 21), pour la caractérisation des dépressions venant de l'Atlantique.

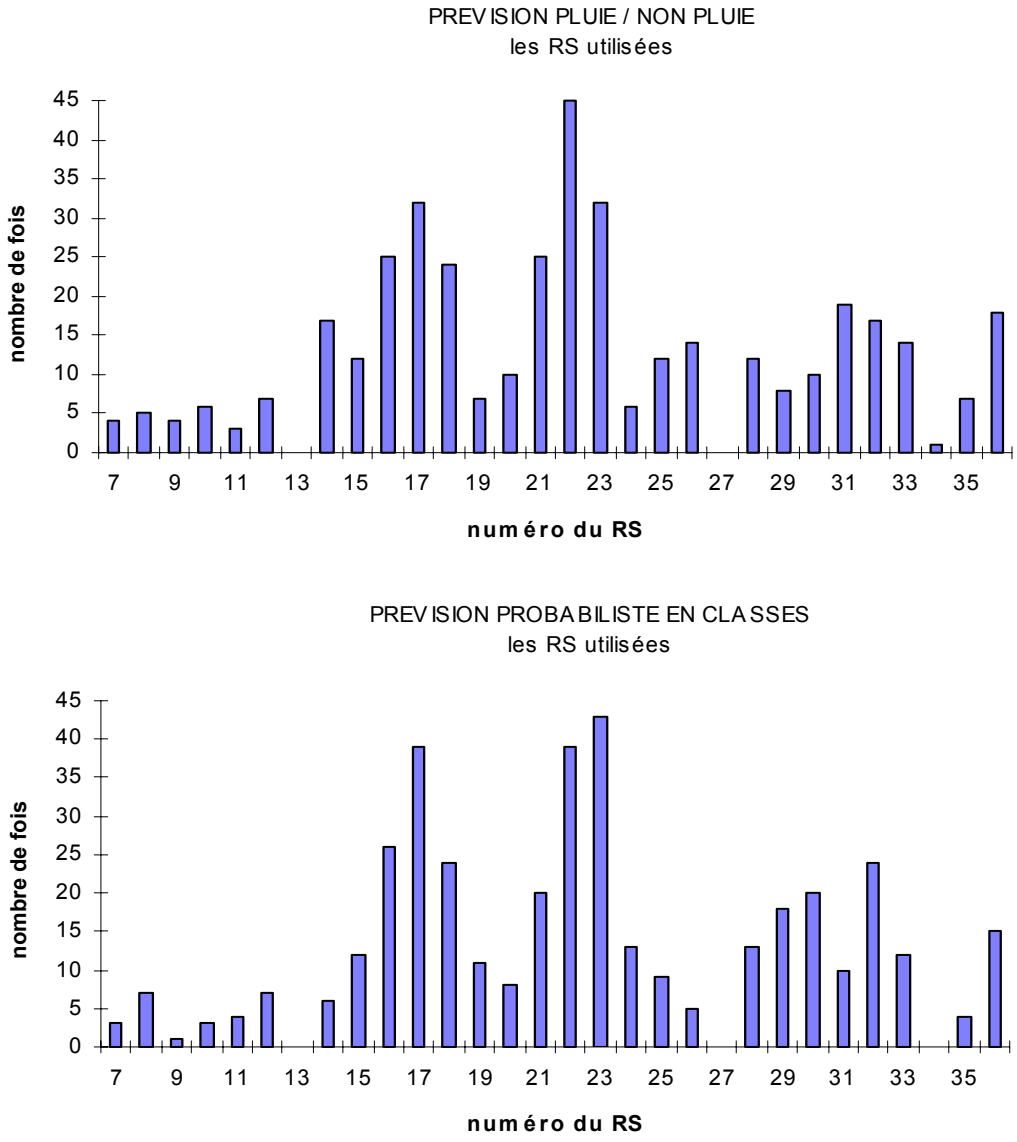


figure IV-5: les stations de radiosondages les plus sollicitées

IV.1.4 Conclusion

Le gain obtenu en utilisant directement les radiosondages n'est pas fortement significatif mais néanmoins, on ne peut pas dire que l'utilisation des CP se justifie encore de nos jours.
Cependant, le fait d'abandonner les CP au profit des données aux stations de radiosondage ne nous satisfait pas complètement. En effet, si l'on a éliminé un prétraitement qui est l'ACP, il reste qu'un certain nombre de ces stations n'existent plus et qu'elles sont reconstituées par interpolation de

points de grille. De plus, en utilisation opérationnelle, il est toujours à craindre qu'une station existante soit défaillante. La remplacer par sa valeur interpolée crée alors une hétérogénéité.

Ainsi, ce qui serait vraiment intéressant, ce serait de travailler directement sur une grille de points assez dense, car:

- à travers le processus d'analyse et d'assimilation, ces données sont critiquées et l'absence éventuelle d'une mesure à une station est aisément compensée,

- la forme des données observées (grille) est alors homogène avec la forme des données prévues fournies par les modèles météorologiques.

- elles permettraient d'harmoniser notre jeu de données : de nouvelles variables disponibles uniquement sous forme de grille (divers champs sur l'Hémisphère Nord) pourraient être testées, en sus des géopotentiels, dans la sélection des analogues (cf. chapitre V).

IV.2 Utilisation de grilles et du score de Teweles-Wobus

IV.2.1 Sélection des données en points de grille *

Comme cela a été mentionné au chapitre II, une grille de 1° x 1° a été reconstituée par interpolation des données de radiosondage grâce à une fonction spline. Elle comporte 19 x 22 soit 418 points (cf. fig. II-12).

En première approche, il a été envisagé de travailler comme avec les Composantes Principales (CP) ou les données brutes aux stations de RadioSondage (RS), c'est-à-dire de faire la sélection des points de grille utiles à l'extraction des analogues un par un.
Cependant, avec une grille de 418 points, cette sélection, même la plus simple possible (ascendante K=1 par exemple) est encore rhédibitoire en temps de calcul. C'est pourquoi nous avons repris la méthode déjà retenue par Martin (1995), à savoir l'utilisation de sous-domaines pour faire l'analogie.

Par conséquent, pour les 6 bassins témoins, 6 tailles de grille ont été testées en prévision pluie / non pluie (cf. figure IV-6). C'est-à-dire que 6 sélections ascendantes par bassin ont été

effectuées, une par taille de grille, à partir des valeurs aux points de la grille choisie des champs de géopotentiel 700 et 1000 hPa à 00 h.

Le premier critère utilisé pour la sélection des analogues est le suivant:

$$D^2_{GR}(J,C) = \sum_{i=1}^{N_{GR}} \left[GR_i(J) - GR_i(C) \right]^2 \qquad \text{(IV-2)}$$

où N_{GR} correspond au nombre de points de grille GR retenus.

Dans le calcul de la prévision avec les analogues, un poids en $1/D_{GR}$ est toujours donné à chaque analogue.

La figure IV-6 présente les grilles choisies:

- *grille n°1:* elle englobe juste les 33 bassins,

- *grilles 2 à 5:* elles augmentent progressivement de chaque côté en se développant 2 fois plus à l'Ouest et au Sud qu'au Nord et à l'Est, car on a vu au §IV.1.1 que les stations du Nord et de l'Est étaient moins sollicitées,

- *grille n°6:* c'est la grille totale avec 19 x 22 points.

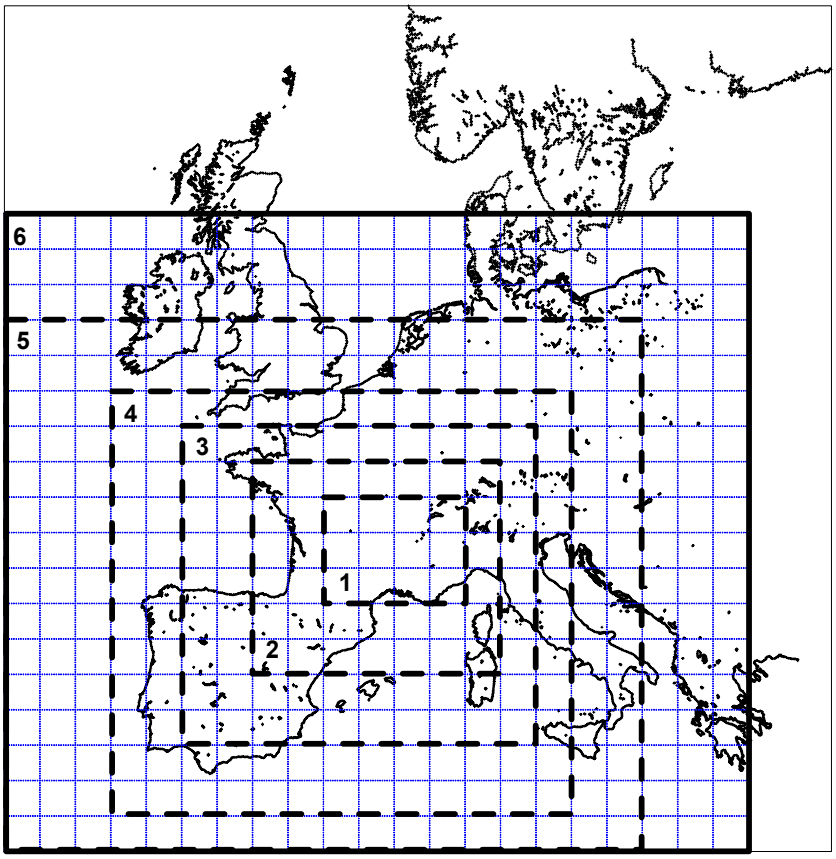


figure IV-6 : les différentes tailles de grille

La méthode de sélection utilisée est une sélection ascendante simple K=1 à cause des temps de calcul qui deviennent vite prohibitifs au fur et à mesure que la taille de la grille augmente. En effet, si pour la grille n°1, 2 heures de calcul sont nécessaires par bassin, pour la grille totale, ce sont près de 40 heures.

Les indices de réussite IR en fonction du nombre de points de grille retenus ont ensuite été tracés pour les différentes grilles. Ils ont été comparés à ceux obtenus par la méthode S-8CP (cf. annexe IV-3). Le tableau IV-2 ci-dessous donne le pourcentage de réussite lorsque le palier est atteint, soit autour de 12 points de grille retenus:

	Creuse-Cher	Pyrénées Est	Doubs	Isère moyen	Var Tinee Roya	Chassezac
grille n°1	75.0	72.2	76.2	76.0	74.2	76.0
grille n°2	76.6	73.3	76.7	77.3	74.9	76.3
grille n°3	77.0	73.0	77.7	78.0	75.0	76.4
grille n°4	***78.2***	74.2	78.3	***78.3***	74.2	77.0
grille n°5	***78.2***	*74.8*	***78.7.***	78.0	***76.5***	***77.5***
grille n°6	*78.2*	*74.8*	***78.7***	77.8	76.4	77.4
S-8CP	**77.0**	**75.1**	**76.82**	**78.29**	**76.7**	**75.7**

tableau IV-2: comparaison des tailles des grilles

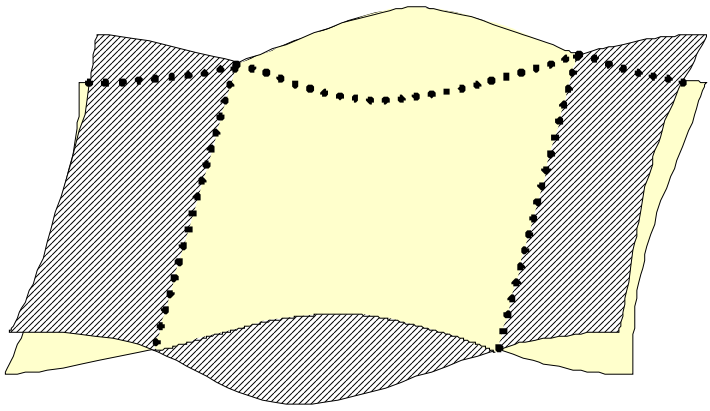
Conclusion :

Les résultats obtenus donnent comme taille optimale de grille la grille n°5. Cela montre que l'analogie doit se faire à une échelle intermédiaire, ni trop grande, ni trop petite : la taille de la zone d'étude, grille n°1, ne convient pas et celle englobant tout le domaine non plus.
A ce sujet, nous retrouvons les résultats obtenus par Van Den Dool en 1989 (Van Den Dool, 1989) qui parle d'analogie à *aire limitée*: il a montré que, pour une prévision à courte échéance, l'analogie à l'intérieur d'un cercle de rayon inférieur à 1000 km suffisait.

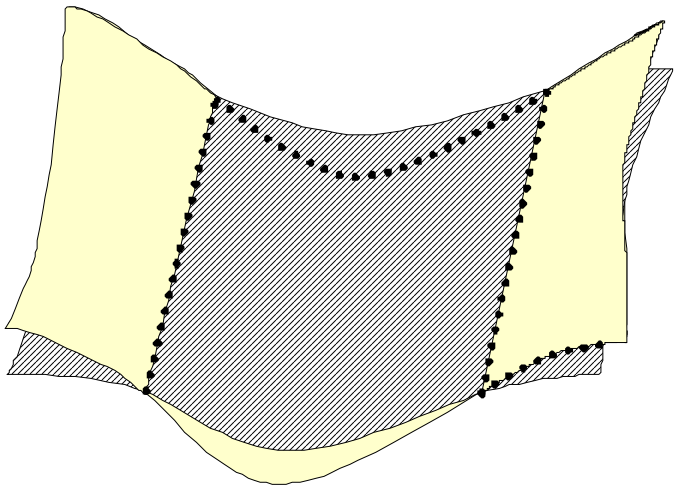
Pour ce qui est de la comparaison avec la méthode S-8CP, les résultats en pluie / non pluie ne sont pas concluants puisque l'utilisation des points de grille donne de meilleurs résultats seulement pour 3 bassins (Creuse-Cher, Doubs et Chassezac). Il faut cependant noter que la

sélection effectuée sur les points de grille est une sélection simple K=1, un peu moins optimale que celle (K=2) de la méthode S-8CP.

Il paraît donc difficile de conclure d'emblée. Et comme les temps de calcul sont très importants, nous n'irons pas plus loin dans cette voie. Néanmoins, une question se posait quant à la pertinence de la distance euclidienne pour mesurer la ressemblance de deux champs connus en une grille de points. Celle-ci s'appuie sur la somme des carrés des écarts entre points de mêmes coordonnées. Si elle traduit bien la distance, elle caractérise assez mal la forme du champ, et deux champs peuvent être déclarés "proches" tout en ondulant quasiment en opposition de phase, ce qui localise les centres d'action en des endroits très différents (cf. figure IV-7). Aussi, d'autres critères de sélection ont-ils été testés.



Formes non ressemblantes



Formes ressemblantes

figure IV-7 : champs proches et différents

IV.2.2 Définition du score de Teweles-Wobus

Nous avons recherché un critère caractérisant mieux l'analogie "globale" de forme que la distance euclidienne. C'est ce que fait le score de Teweles-Wobus TW (Teweles-Wobus, 1954) : historiquement, il a été mis au point dans les années 50 pour juger de la qualité des prévisions des champs de géopotentiel. De plus, il a déjà été proposé dans la littérature pour sélectionner les analogues (Woodcock, 1980; Wilson and Yacowar, 1980...) mais sans que sa mise en œuvre soit détaillée.

Nous avons donc dû revenir à la publication initiale (Teweles-Wobus, 1954) pour comprendre qu'il met l'accent sur la circulation, donc plutôt sur les champs de vents que sur les champs de pression, en considérant en chaque station, non pas le géopotentiel, mais ses deux gradients méridien (Sud-Nord) et zonal (Est-Ouest) :

$$TW = 100 * \frac{\sum |e_G|}{\sum |G_L|} \qquad \text{(IV-3)}$$

où e_G est l'écart entre la différence de pression prévue et observée,

et G_L le maximum entre la différence de pression prévue et observée,

entre 2 stations.

Initialement, la sommation est supposée faite sur tous les couples de stations voisines (dans les directions Est-Ouest et Sud-Nord) d'un réseau le plus régulier possible.

L'utilisation du coefficient $1 / \Sigma |G_L|$ permet de normer et de désaisonnaliser le numérateur, très affecté par les variations saisonnières du champ de pression.

Il mesure, en fait, l'amplitude du gradient de pression entre 2 points voisins, qui est étroitement liée au flux de vent, et c'est cette grandeur qu'il compare entre les 2 cartes.

Si l'on utilise l'indice supérieur $(\ )^i$ pour la direction Est-Ouest et $(\ )^j$ pour la direction Nord-Sud, on obtient, pour le score de Teweles-Wobus, appliqué cette fois sur une grille régulière, la formule suivante IV- 4 (Nieminen, 1982). Pour la rendre plus explicite que la première, il a été convenu de prendre F pour Forecast et A pour Actual.

$$TW = 100 * \frac{\sum_{i,j} \left| \Delta F(i,j)^i - \Delta A(i,j)^i \right| + \sum_{i,j} \left| \Delta F(i,j)^j - \Delta A(i,j)^j \right|}{\sum_{i,j} G_L^i + \sum_{i,j} G_L^j} \qquad \text{(IV-4)}$$

avec (cf. figure IV-8):

- les gradients prévus: $\Delta F(i,j)^i = F(i,j) - F(i+1,j)$

  $\Delta F(i,j)^j = F(i,j) - F(i,j+1)$

- les gradients observés $\Delta A(i,j)^i = A(i,j) - A(i+1,j)$

  $\Delta A(i,j)^j = A(i,j) - A(i,j+1)$

- $G_L^i = \max \left[\left| \Delta F(i,j)^i \right| , \left| \Delta A(i,j)^i \right| \right]$

- $G_L^j = \max \left[\left| \Delta F(i,j)^j \right| , \left| \Delta A(i,j)^j \right| \right]$

Il varie de 0, pour 2 cartes identiques, à 200.

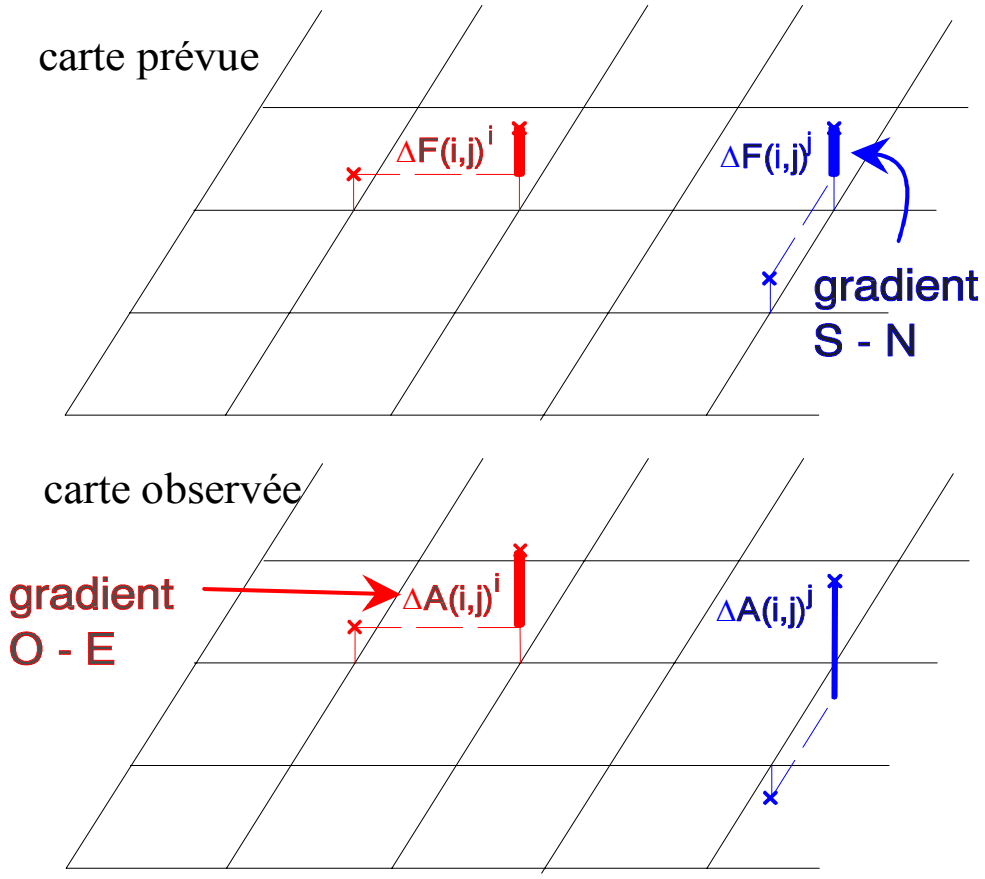


figure IV-8: calcul du score TW de Teweles-Wobus

Pour calculer ce score, Teweles et Wobus avaient sélectionné un réseau de stations représentatives, distantes de 500 km en moyenne. Cette distance inter station est un compromis entre:

- une distance assez grande pour limiter le nombre de station à utiliser, sachant qu'il existe une corrélation spatiale importante,

- et une distance assez petite pour ne pas manquer d'importantes caractéristiques du champ en question.

On retrouvera en figure IV-9, les stations utilisées dans leur travail original. Les gradients de pression sont calculés entre les points reliés par un trait plein (direction Est-Ouest) ou pointillé (direction Nord-Sud). On voit donc que, d'après la construction même du score de Teweles-Wobus, il semble particulièrement indiqué pour des données en points de grille régulièrement répartis.

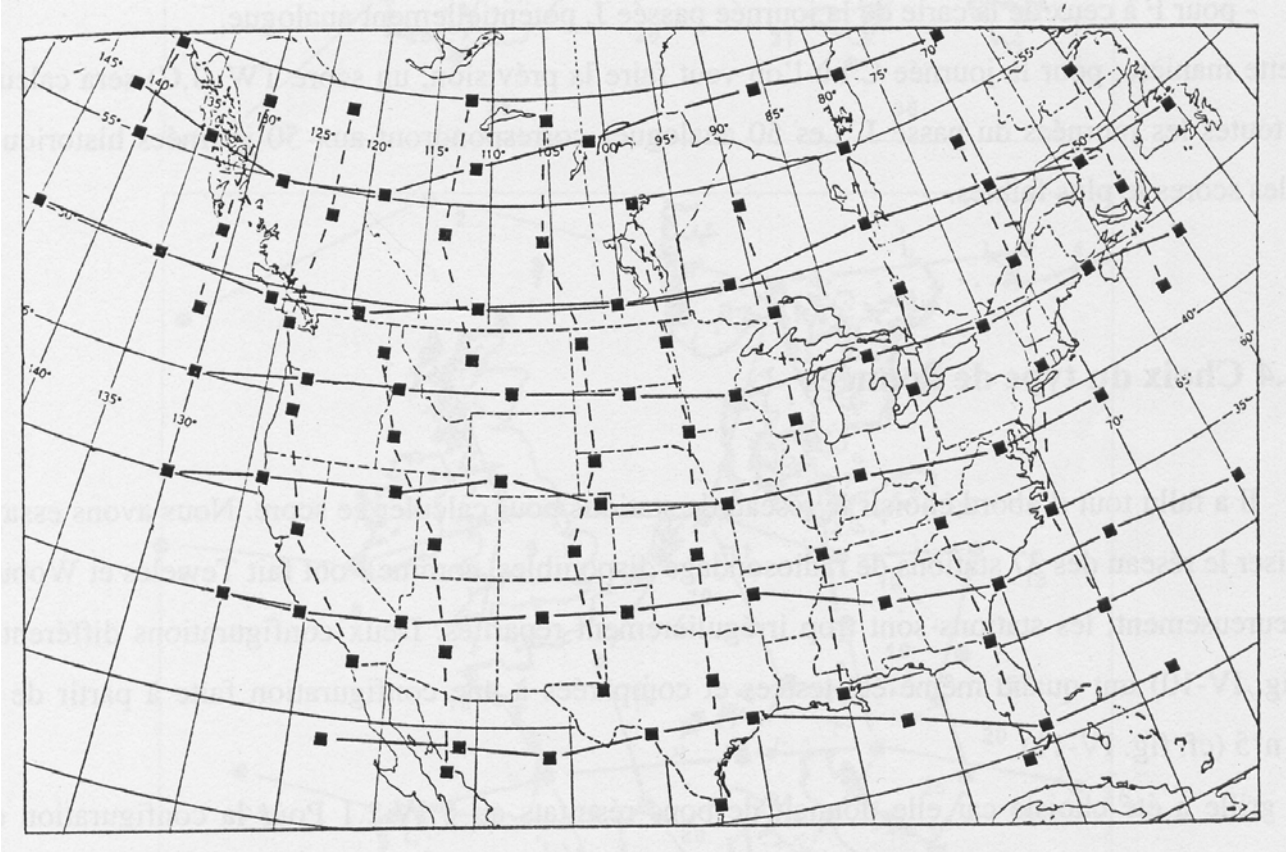


figure IV-9: grille de points utilisée par Teweles-Wobus (1954)

IV.2.3 Utilisation du score TW pour l'analogie

Un des premiers à utiliser le score de Teweles-Wobus a été Woodcock (1980) qui dira : « Ce score, mesurant la similitude des vents géostrophiques entre 2 cartes, est un outil tout à fait adapté pour sélectionner des analogues pour lesquels les éléments du temps sont largement contrôlés par les régimes de vent ».

Au lieu de mesurer la ressemblance entre un champ prévu F et un champ observé A grâce au score de Teweles-Wobus, c'est celle entre le champ de la journée courante C et celui de la journée historique J qui nous intéresse. Aussi, si l'on se réfère à l'équation IV-4, les gradients A et F ne correspondent plus aux gradients prévus et observés mais respectivement:

- pour A aux gradients de la carte de la journée courante C,
- pour F à ceux de la carte de la journée passée J, potentiellement analogue.

De cette manière, pour la journée C où l'on veut faire la prévision, un score TW (J,C) sera calculé pour toutes les journées du passé J. Les 50 analogues correspondront aux 50 journées historiques avec les scores le plus faibles.

IV.2.4 Choix du type de données

Il a fallu tout d'abord choisir le réseau de stations pour calculer ce score. Nous avons essayé d'utiliser le réseau des 37 stations de radiosondage disponibles, comme l'ont fait Teweles et Wobus. Malheureusement, les stations sont trop irrégulièrement réparties. Deux configurations différentes (cf. fig. IV-10) ont quand même été testées et comparées à une configuration faite à partir de la grille n°5 (cf. fig. IV-11).

Cette grille a été choisie car elle donnait de bons résultats au § IV.2.1 Pour la configuration en points de grille, un point sur 2 a été utilisé car les points voisins, distants d'environ 110 km, sont très corrélés entre eux ($R > 0.95$).

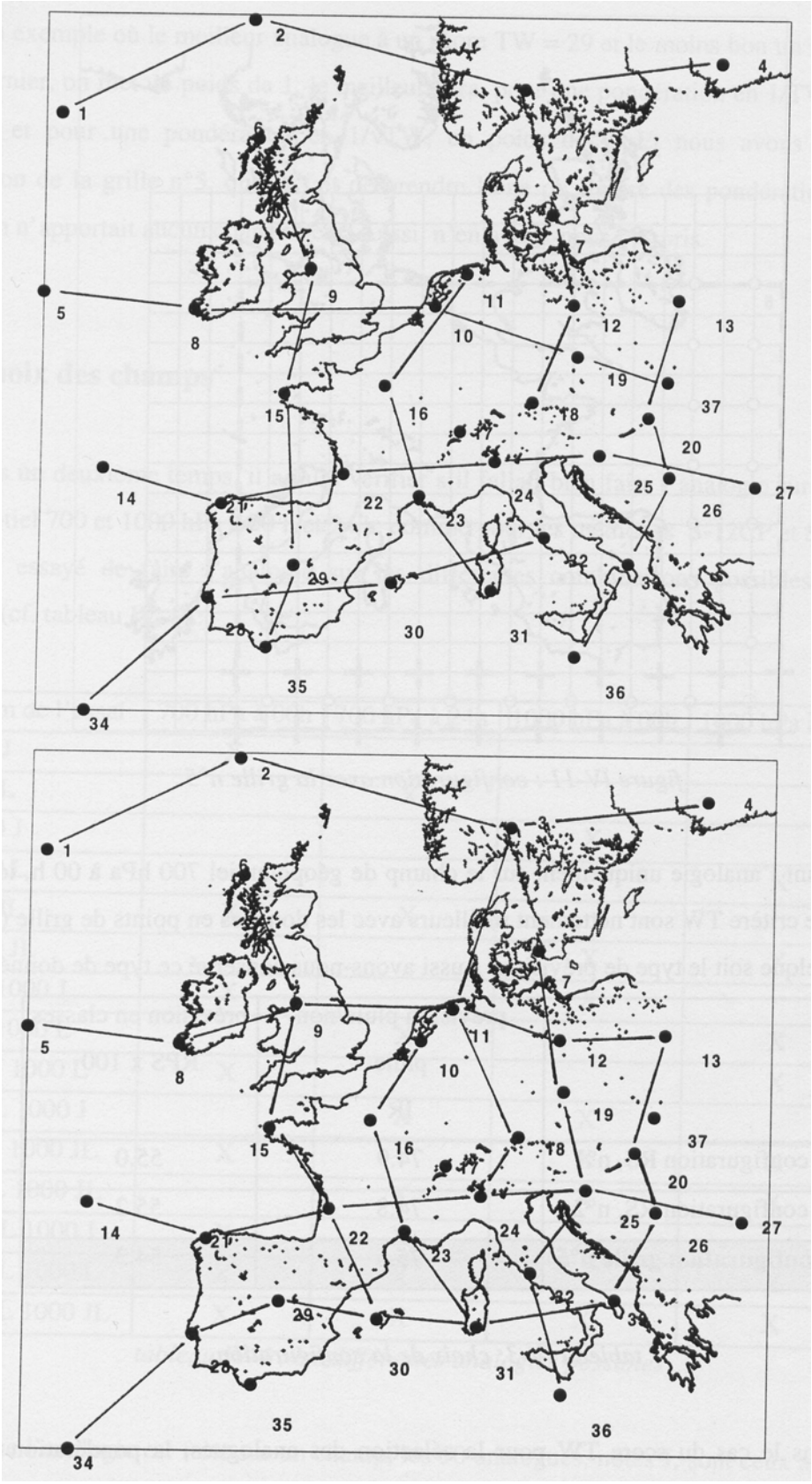


figure IV-10 : les configurations avec les stations de radiosondage

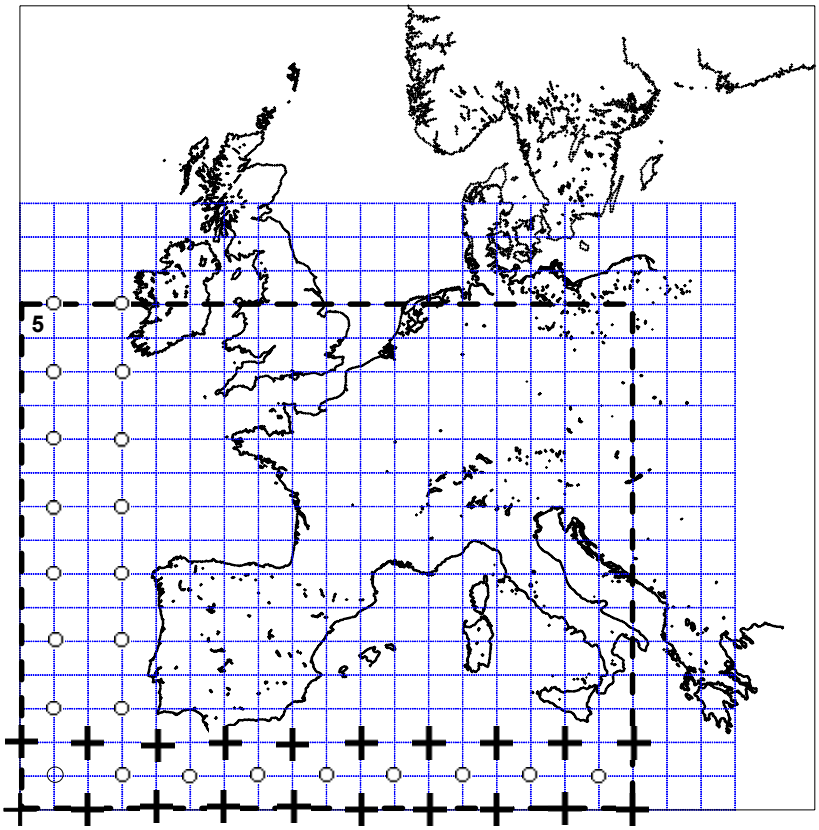


figure IV-11 : configuration avec la grille n°5

En faisant l'analogie uniquement sur le champ de géopotentiel 700hPa à 00h, les résultats obtenus avec ce critère TW sont nettement meilleurs avec les données en points de grille (cf. tableau IV-3), et ce quelque soit le type de prévision. Aussi avons-nous conservé ce type de données.

	prévision pluie/non pluie IR	prévision en classes RPS x 100
configuration RS, n°1	74.9	55.0
configuration RS, n°2	74.8	55.2
configuration grille n°5	**75.6**	**54.3**

tableau IV-3: choix de la configuration

Remarque: dans le cas du score TW pour la sélection des analogues, la pondération de ceux-ci n'apporte rien car, par rapport à la distance euclidienne, le critère TW ne varie pas assez entre le meilleur et le plus mauvais analogue pour avoir des poids significativement différents de 1.

Prenons un exemple où le meilleur analogue à un score TW = 29 et le moins bon un score de 41. Si pour ce dernier, on met un poids de 1, le meilleur aura, pour une pondération en 1/TW, un poids de 1.4 soit 1 et pour une pondération en 1/√TW, un poids de 2. Et nous avons vérifié sur la configuration de la grille n°5, que le fait de prendre l'une ou l'autre des pondérations ou aucune pondération n'apportait aucun changement. Aussi, n'en avons-nous pas pris.

IV.2.5 Choix des champs

Dans un deuxième temps, il a fallu vérifier s'il fallait bien faire l'analogie sur les 2 champs de géopotentiel 700 et 1000 hPa à 00 h et 24 h, comme pour les méthodes S-12CP et S-12RS. Nous avons donc essayé de faire l'analogie sur les différentes combinaisons possibles des champs disponibles (cf. tableau IV-4) :

nom de l'essai	700 hPa à 00h	700 hPa à 24h	1000 hPa à 00h	1000 hPa à 24h
700 J	X			
700 L		X		
1000 J			X	
1000 L				X
700 JL	X	X		
1000 JL			X	X
700 1000 J	X		X	
700 1000 L		X		X
700 J 1000 L	X			X
700 L 1000 J		X	X	
700 J 1000 JL	X		X	X
700 L 1000 JL		X	X	X
700 JL 1000 J	X	X	X	
700 JL 1000 L	X	X		X
700 JL 1000 JL	X	X	X	X

tableau IV-4 : les différentes analogies possibles

Lorsque l'analogie est faite sur un champ, les 50 analogues, notés J, sont ceux avec le score TW(J,C) le plus faible. Dans le cas d'une analogie sur 2 (respectivement 3 et 4) champs, le score

TW est calculé séparément pour les 2 (3, 4) champs et les analogues retenus sont ceux pour lesquels la somme des 2 (3, 4) scores est la plus petite.

Ainsi, si l'on veut faire l'analogie sur les champs 700 et 1000 hPa à 00 et 24 h, le critère d'analogie sera:

$$TW(J,C) = TW_{700}^{00h}(J,C) + TW_{700}^{24h}(J,C) + TW_{1000}^{00h}(J,C) + TW_{1000}^{24h}(J,C) \qquad \text{(IV-5)}$$

Pour les 2 types de prévision et les différents essais, les scores moyens sur les 33 bassins français sont donnés dans le tableau IV-5.

	prévision pluie / non pluie IR	prévision en classes RPS * 100
700 J	75.6	54.3
700 L	77.1	49.8
1000 J	75.9	53.4
1000 L	76.7	50.0
700 1000 J	76.7	52.4
700 1000 L	78.0	48.1
700 JL	78.0	48.6
1000 JL	78.6	47.7
700 J 1000 L	78.7	47.4
700 L 1000 J	79.0	47.3
700 J 1000 JL	78.7	47.8
700 L 1000 JL	**79.1**	**46.5**
700 JL 1000 J	78.6	48.1
700 JL 1000 L	78.6	47.1
700 JL 1000 JL	**79.1**	**46.7**

tableau IV-5 : performances des différentes analogies

Quelques remarques peuvent être faites:

- l'utilisation d'un seul champ pour faire l'analogie ne suffit pas,

- dans tous les cas, les performances sont meilleures quand on utilise les champs à 24 h plutôt qu'à 00 h,

- l'utilisation des 2 champs 700 et 1000 hPa apportent une légère amélioration,

- les meilleurs résultats en prévision pluie / non pluie sont obtenus avec les 4 champs. Par contre, en prévision par classes, seulement 3 champs sont nécessaires pour avoir les meilleures performances. Néanmoins, l'utilisation des 4 champs donne un score pratiquement équivalent.

Aussi, nous retiendrons que, même en changeant de critère de sélection des analogues, l'analogie doit continuer à se faire sur **les 2 champs de géopotentiel 700 et 1000 hPa, aux échéances 00 et 24 h**. Et dans ce cas, l'utilisation du score de Teweles-Wobus donne de meilleurs résultats que la méthode S-12CP, et ce, pour les 2 types de prévision.
Quant à la méthode S-12RS utilisant les données brutes, elle est moins intéressante pour la prévision en classes et pratiquement équivalente en pluie / non pluie (cf. tableau IV-6).

	prévision pluie / non pluie IR	prévision en classes RPS * 100
méthode S-12CP	78.6	49.0
méthode S-12RS	79.2	47.2
700 JL 1000 JL	79.1	46.7

tableau IV-6 : comparaison des différentes méthodes

L'utilisation de ce score TW semble donc très prometteur. Cependant, si nous avons choisi d'abord le type de données (grille) et ensuite les champs sur lesquels faire l'analogie (champs de géopotentiel 700 et 1000 hPa à 00 et 24 h), il pourrait être intéressant d'optimiser et la taille et la localisation de la grille, que nous avons, jusqu'à maintenant, choisie de manière assez arbitraire.

IV.2.6 Optimisation de la grille

IV.2.6.a Taille

Pour optimiser la taille de la grille, les 6 tailles de grille du §IV.1.2 ont été reprises. Dans le tableau IV-7 sont consignées les performances de chacune des grilles, pour les 2 types de prévision,

sachant que l'analogie est faite, pour tout ce qui suivra, sur les champs de géopotentiel 700 et 1000 à 00 et 24 h.

Une grille de taille intermédiaire (grille n°3) suffit donc, et avec celle-ci, les performances dépassent nettement celles des méthodes utilisant et les CP et les données brutes. Toutefois, on doit pouvoir encore faire mieux en jouant sur la localisation de cette grille.

	prévision pluie / non pluie IR	prévision en classes RPS * 100
grille n°1	75.5	51.6
grille n°2	78.3	47.7
grille n°3	**80.1**	**45.0**
grille n°4	79.5	45.5
grille n°5	79.1	46.7
grille n°6	78.4	48.1
méthode S-12CP	**78.6**	**49.0**
méthode S-12RS	**79.2**	**47.2**

tableau IV-7 : optimisation de la taille de la grille pour le critère TW

IV.2.6.b Localisation

Cette grille peut être déplacée vers l'Est ou vers l'Ouest, ou encore vers le Sud ou le Nord. Cependant, un test systématique de toutes les positions possibles vers l'Est et l'Ouest de la grille n°3 n'a pas apporté de gain supplémentaire (cf. annexe IV-4). On a ensuite essayé de déplacer la grille jusqu'à 3 points de grille vers le Sud et le Nord, tout en la bougeant vers l'Est et l'Ouest mais cela non plus n'a rien donné de mieux que la grille n°3 initiale (cf. annexe IV-4).

Globalement sur les 33 bassins français, la grille n°3 est donc la plus performante. On verra dans le chapitre VI que cela peut être affiné pour des sous-ensembles de bassins, et en particulier pour les bassins italiens et espagnols.

IV.2.7 Conclusion et interprétations complémentaires du score TW

L'utilisation du score de Teweles-Wobus, après optimisation de la grille, a donc apporté un gain appréciable pour la sélection des analogues. Cependant, ce gain peut provenir de différentes raisons :

i) Tout d'abord, avec ce score, l'analogie est effectuée, en fait, non pas sur le champ de géopotentiel mais sur son gradient. On travaille donc sur les champs « différenciés », c'est-à-dire en pratique, sur les écarts entre 2 points. La démarche est alors un peu analogue à celle qui conduit, en analyse objective, à privilégier le variogramme (ou la "fonction de structure") sur la fonction de corrélation.

L'utilisation de ces écarts peut, à elle seule, expliquer les gains en prévision. Cela converge alors vers les résultats de Martin (1995) qui a montré que la distance euclidienne appliquée aux champs de gradient de géopotentiel donnait de meilleurs résultats qu'appliquée directement aux champs de géopotentiel.

ii) Outre le fait que ce score utilise les gradients, des différences peuvent provenir de la formulation du score lui-même, par rapport à la distance euclidienne, pour expliquer la hausse des performances.

Ainsi, il s'agit d'un critère plus « doux » que la distance euclidienne dans la mesure où il travaille en valeur absolue des écarts, sans les « exacerber » par une élévation au carré.

De plus, on a calculé le score de Teweles-Wobus et la distance euclidienne pour 3 exemples de champs à une dimension, très schématisés (cf. figure IV-12). On remarque que, pour le score TW, plus les champs sont ressemblants, meilleur est le score qui tend vers 0. Il est d'ailleurs égal à 0 pour 2 champs identiques, même décalés en ordonnée. Par contre, la plus mauvaise valeur (200) est obtenue pour 2 champs en opposition de phase. Ce comportement diffère de la distance euclidienne: celle-ci pénalise beaucoup les champs très ressemblants en forme mais relativement éloignés (cf. cas 3) et privilégie, au contraire, des champs moins similaires en forme mais plus voisins en distance (cf. cas 2).

Enfin, sans que nous l'ayons testé nous-mêmes, Nieminen (1982) montre qu'il est relativement peu sensible à la maille utilisée (tant que celle-ci reste raisonnablement petite par rapport à la "longueur d'onde" de la circulation atmosphérique). Il cite comme exemple que le passage d'une grille de 3x6° en latitude-longitude à 3x3° ne modifie pas les résultats. Dans notre

cas, la grille interpolée à 1.5x1.5° est utilisée sur une base de 3x3°. Nous bénéficions donc probablement de la robustesse citée par Nieminen.

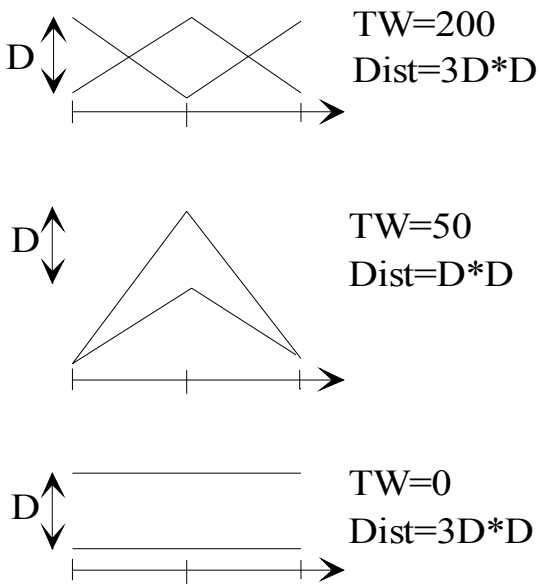


figure IV-12 : distance euclidienne D^2 et score TW pour des champs très schématisés

En dehors de ces considérations, l'analogie peut être faite sur un champ particulier (ex: 700 hPa) ou sur plusieurs. Dans nos essais, il est apparu qu'elle devait être faite sur l'ensemble des champs disponibles à savoir les champs de géopotentiel **700 et 1000 hPa à 00 et 24 h** (cf. équation IV-5). C'est une nouvelle méthode de sélection des analogues que nous appellerons **méthode TW-GR.**

Dans une perspective ultérieure, on peut même imaginer de pondérer différemment (mais globalement) chaque champ. Cependant, nos tentatives précédentes de pondération (cf. Chapitre III, § III.3.2) ne nous y ont pas encouragés.

Cependant, on trouve dans la littérature d'autres façons de sélectionner des analogues, que nous avons aussi voulu tester de manière à être le plus exhaustif possible.

IV.3 Autres critères de sélection des analogues **

Synthèse : On rencontre dans la littérature différentes façons de sélectionner les analogues, en dehors de la distance euclidienne et du score de Teweles-Wobus. Nous les avons testés afin d'être le plus exhaustif possible, mais sans succès.

En fait, on rencontre principalement dans la littérature 3 grands types de critère de sélection des analogues:

- la distance euclidienne utilisée par Duband (1970), Barnett et Preisendorfer (1978)...,
- le score de Teweles-Wobus (Woodcock, 1980; Keenan et Woodcock, 1979...),
- le coefficient de corrélation (Yacowar, 1980; Bergen and Harnack, 1982 etc.),

ainsi que des combinaisons de ceux-ci, à savoir:

- soit l'utilisation de 2 critères l'un après l'autre comme Navarre en 1980 qui utilise une distance euclidienne pour sélectionner de manière objective les analogues et un coefficient de corrélation pour avoir une appréciation plus subjective de la qualité des analogues ou encore Duband (1970) avec son critère de corrélation (cf. § I.2.2.b),
- soit l'utilisation d'un critère de sélection combinant 2 des 3 précédents comme l'a fait Yacowar en 1980.

IV.3.1 Utilisation d'un critère seul

Il a été montré précédemment que, pour optimiser l'utilisation de la distance euclidienne, il s'avérait nécessaire de sélectionner les variables (Composantes Principales ou données aux stations de RadioSondages) à y introduire (cf. chapitre III et § IV.1.1).
Néanmoins, l'utilisation du score de Teweles-Wobus pour extraire les analogues, notamment après optimisation de la grille, augmente encore les performances de la prévision.

Ensuite, nous avons essayé de prendre un coefficient de corrélation R pour extraire les analogues. Mais, que ce soit en faisant la corrélation :

- sur des CP ou des RS,

- sur un champ ou plusieurs,

et même en pondérant les analogues en $1/R^2$, les résultats sont nettement inférieurs à tout ce que l'on a pu obtenir jusqu'à maintenant. Donc, si l'on doit utiliser un seul critère, c'est la distance euclidienne ou le score de Teweles-Wobus TW.

IV.3.2 Combinaison de critères

Pour prévoir la pluie à 12 heures, Verret (1985) sélectionne ses analogues à l'aide du critère suivant, que nous noterons par la suite TWR:

$$TWR = \left[\left(1 - R_1\right)^2 + \left(1 - R_5\right)^2 + \left(1 - \frac{TW_1}{100}\right)^2 + \left(\frac{TW_5}{100}\right)^2 \right]^{1/2} \qquad \text{(IV-6)}$$

avec - les coefficients de corrélation entre la journée courante C et la journée historique J:

R_1, pour le champ de géopotentiel 1000 hPa,

R_5, pour le champ de géopotentiel 500 hPa,

- les scores de Teweles-Wobus toujours entre la journée à prévoir C et la journée J pour les 2 champs: TW_1 et TW_5.

Là encore, si l'on utilise les 2 champs 700 et 1000 hPa à 24 h plutôt qu'à 00 h, les résultats, en prévision pluie / non pluie, sont meilleurs :

- 74.6 % pour 00 h,
- 76.2 % pour 24 h.

Néanmoins, s'ils sont nettement meilleurs que ceux obtenus avec le critère de corrélation utilisé seul pour sélectionner les analogues, ils sont toujours inférieurs aux performances des 2 autres critères utilisés seuls:

- 79.2 % avec la distance euclidienne,
- 80.1 % avec le score TW de Teweles-Wobus.

Aussi, n'avons-nous pas poursuivi dans cette voie. Nous avons préféré tester une idée parue récemment dans la littérature (Harnack *et. al.*, 1986 ; Barnston and Livezey, 1986 ; Van Den Dool, 1987), les anti-analogues.

IV.3.3 Les anti-analogues

Le principe a été exposé au chapitre 1, § I.3.2. Il s'agit de différencier, dans le fichier de journées historiques, les journées potentiellement *analogues* de celles qui sont *anti-analogues*. Cela se fait grâce au signe du produit scalaire entre le vecteur d'état climatique - représentant la situation en question - de la journée courante C et celui de la journée historique J: s'il est positif, on est en présence d'un analogue potentiel et s'il est négatif, c'est un anti-analogue potentiel. Pour ce dernier, son vecteur d'état climatique est alors multiplié par (-1) et il redevient un analogue positif, tout en restant répertorié comme un anti-analogue ou analogue négatif.
Ensuite, on recherche les meilleurs analogues, qui peuvent être soit positifs, soit négatifs, grâce à un critère prédéterminé. Et, pour le calcul de la prévision, on considère pour l'analogue négatif l'inverse de ce qu'il annonce.

Dans notre cas, nous avons testé l'utilisation des anti-analogues sur:

i) la méthode S-12CP avec

- comme vecteur d'état climatique les 12 CP de la méthode S-12CP, le produit scalaire entre le jour C et le jour J étant:

$$PS\ (J,C)\ =\ \sum_{i=1}^{12}\left[\mathbf{X_i}(J)\times\ \mathbf{X_i}(C)\right] \qquad \text{(IV-7)}$$

- et comme critère de sélection des 50 analogues, la distance euclidienne de la méthode S-12CP,

ii) la méthode TW-GR avec

- comme vecteur d'état climatique les valeurs des champs 700 et 1000 hPa à 00 et 24 h, aux points de la grille n°3,

- et comme critère de sélection le score TW sur ces mêmes champs (cf. éq. IV-5).

Ensuite, la prévision pluie / non pluie est déterminée comme d'habitude, en utilisant le nombre d'analogues sans pluie et la climatologie, mais en attribuant aux journées *sans pluie* répertoriées comme anti-analogues, *pluie* et inversement.

Les résultats consignés dans le tableau IV-8 étant moins bons avec les anti-analogues, cette idée a été abandonnée. En outre, ce principe d'anti-analogue est difficilement applicable à une

prévision en 8 classes de pluie. Si on peut encore envisager que l'opposé de la classe 8 (>100 mm) est la classe 1 (0 mm), que peut-on dire de l'opposé de la classe 4 par exemple ?

	analogues	analogues/anti-analogues
S-12CP	78.6	74.2
TW-GR	**80.1**	**79.3**

tableau IV-8 : analogues et anti-analogues

IV.4 Utilisation des réseaux de neurones **

Synthèse : L'utilisation d'un réseau de neurones nous a semblé attractive par sa capacité à traiter les non-linéarités. Malheureusement, son application directe à la pluie journalière en mm s'est révélée décevante et celle pour prévoir l'occurrence de pluie a donné des résultats comparables à ceux de la méthode des analogues TW-GR. Aussi n'avons-nous pas poursuivi dans cette voie. Cependant, des perspectives de combinaison des 2 méthodes pourraient être envisagées, comme un système expert.

Même si ce paragraphe est un peu à part, puisque l'on va abandonner un temps la prévision par une technique d'analogues, il nous a semblé indispensable de regarder ce que pouvaient apporter une technique récente : *les réseaux de neurones*.

En effet, ceux-ci sont de plus en plus utilisés de nos jours pour, d'une manière générale, résoudre des problèmes avec un grand nombre d'exemples, mais sans règles claires. En particulier, ils ont trouvé un vaste champ d'application dans la modélisation et la prévision des séries chronologiques (Dimopoulos et. al., 1996; Grégoris, 1996; Durot, 1996 etc...), qui nous intéresse principalement. Ceci est dû à leur capacité d'interpoler des données liées entre elles par des relations non-linéaires.

IV.4.1 Présentation des modèles neuronaux

Une présentation succincte sera faite ici mais pour plus de détails, on peut se reporter aux thèses de Bengio (1991), Fessant (1995) ou encore au livre « Des réseaux de neurones » (Davalo et Naïm, 1993) pour une théorie plus approfondie.

Les réseaux de neurones sont une tentative de modélisation mathématique du principe de fonctionnement du cerveau constitué de cellules nerveuses appelées *neurones* et reliées entre elles par des *synapses*.

Un réseau de neurones artificiel est donc inspiré d'un réseau neuronal biologique, constitué de cellules décisionnelles, interconnectées suivant différentes architectures et fonctionnant en parallèle. Par analogie au cerveau, ces cellules sont appelées des neurones et leurs connexions des synapses.

Les réseaux sont caractérisés par i) leur architecture, ii) leur mode de propagation et iii) leur phase d'apprentissage.

i) Pour ce qui est de leur architecture, nous nous intéresserons aux réseaux à couches, principalement utilisés en prévision, en traitement de l'image et de l'information. Ils sont aussi appelés Perceptron Multi-Couches, PMC (Rumelhart *et al*., 1986).

ii) D'une manière simple, on peut considérer le modèle neuronal à couche comme un modèle non linéaire de régression, permettant de relier un vecteur d'entrée U à un vecteur de sortie S.

La transformation S = F (U) est représentée par un réseau de neurones à couches comportant une ou plusieurs *couches cachées* (cf. figure IV-13) où F correspond au *mode de propagation* d'un noeud à l'autre.

Pour les J neurones de la couche cachée :

$$H_j = f\left[\sum_{i=1}^{P} w_{ij} U_j\right] \qquad 1 \le j \le J \qquad \text{(IV-8)}$$

avec U_i état du neurone d'entrée i (composante du vecteur d'entrée U),

H_j sortie du neurone j de la couche cachée,

w_{ij} pondération associée à la connexion (i,j) entre le neurone i de la couche d'entrée et le neurone j de la couche cachée,

f fonction de transfert ou règle de transition des neurones pour calculer leur sortie.

Pour les K neurones de la couche de sortie :

$$S_k = f\left[\sum_{j=1}^{J} w_{jk} H_j \right] \qquad 1 \le k \le K \qquad \text{(IV-9)}$$

avec S_k état du neurone k de sortie (composante k du vecteur de sortie S).

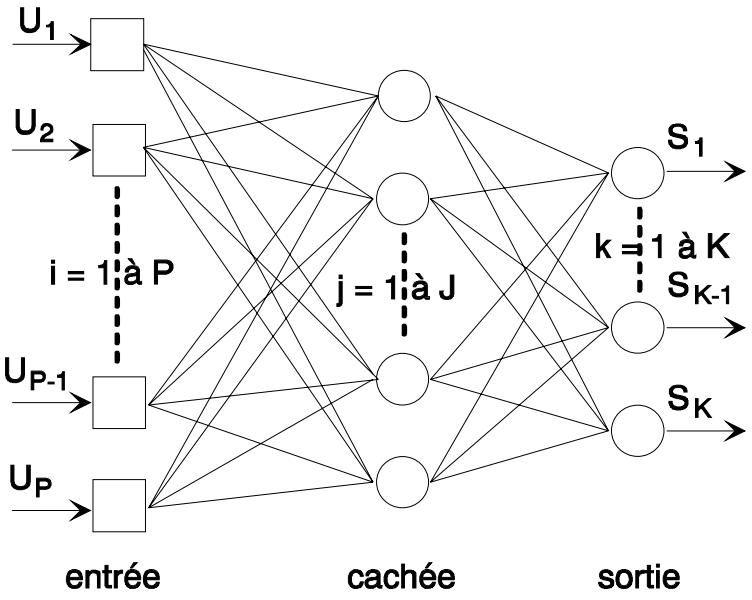


figure IV-13 : Schématisation d'un réseau à une couche cachée

La fonction de transfert a été choisie de forme **sigmoïde**, ce qui introduit la non-linéarité dans le réseau de neurone :

$$f(z) = \frac{1}{1 + e^{-z}} \qquad \text{(IV-10)}$$

Remarque : la structure des connexions et l'architecture du réseau peuvent être beaucoup plus compliquées mais pour notre étude on a simplement utilisé des connexions en couche (Perceptron Multi-Couche) à une couche cachée.

iii) Enfin, comme tout modèle de régression, le modèle neuronal possède des paramètres d'ajustement :

- le nombre de couches cachées (ici pris égal à 1, mais qui dépasse rarement 2 ou 3 car alors le système à résoudre devient trop complexe),

- le nombre J de noeuds dans la couche cachée,

- les poids w_{ij} et w_{jk}.

J étant choisi arbitrairement (de 2 à 6 en général pour ne pas alourdir le système à résoudre en terme de paramètres), il reste à déterminer les poids.

Pour cela, on minimise un critère quadratique (somme des carrés des erreurs de prédiction entre la sortie S connue et celle calculée par le réseau de neurones s) par rétropropagation. C'est la *phase d'apprentissage*, pendant laquelle le réseau apprend à reconnaître certaines informations en comparant les sorties établies par le réseau et celles désirées. Cela nécessite un échantillon d'apprentissage où sont connues l'entrée et la sortie.

Ensuite, le réseau de neurones est près pour faire de la prévision si on lui donne un vecteur d'entrée U.

Cependant, on peut remarquer que le traitement des incertitudes et notamment de l'échantillonnage est assez peu formalisé. Il n'est pas facile d'affecter à une valeur numérique unique (la sortie du réseau) un intervalle de confiance.

Remarque : on cherchera à utiliser le moins de paramètres, de couches et de poids possibles en limitant le critère d'apprentissage à quelques itérations pour obtenir un calage correct et un modèle robuste.

IV.4.2 Utilisation du réseau de neurones pour la prévision des pluies

Quelques essais ont été effectués sur le logiciel NNFIT créé par Cloutier, Grandjean et Thibault à l'université de Laval Sainte-Foy (Québec) au Canada, et avec l'aide de F. Fessant, chercheur post-doc au laboratoire LIFIA de Grenoble.

Les données d'entrée sont les 48 Composantes Principales des champs 700 et 1000 hPa (12 à 00 h et 12 à 24 h pour les deux champs) et celles de sortie la pluie sur le groupement 27 des Cévennes, la Loire supérieure.

Le fichier d'apprentissage est constitué des 2/3 du fichier historique 1953-1993, soit 2500 journées et le fichier test pour la prévision du 1/3 restant (les 1231 suivantes).

Dans un premier temps, la sortie était la pluie en mm observée sur le groupement considéré. Mais les résultats obtenus, et en apprentissage, et en test, se sont révélés inintéressants.

Aussi avons-nous remplacé la sortie par l'occurrence de pluie : 0 pour non pluie et 1 pour pluie. Et comme le réseau de neurones donne une sortie continue, on l'a arrondie à 0 quand elle était comprise entre 0 et 0.5 et à 1, entre 0.5 et 1. Cette fois, sur le fichier test les résultats ont été de l'ordre de ceux obtenus avec la méthode TW-GR sur ce même échantillon :

78 % avec le réseau de neurones,

78,2 % avec la méthode TW-GR.

Ces performances équivalentes ne nous ont pas poussé à continuer dans cette voie. En effet, nous avons préféré nous concentrer sur la prévision par analogues à partir du moment où l'utilisation du réseau de neurones n'apportait pas de gain significatif. Cependant il ne fallait pas ignorer une approche candidate sans la tester.

Néanmoins, en terme de perspectives, il pourrait être intéressant d'envisager un système expert couplant les deux méthodes, avec

- un système décisionnel à construire dans le cas où les 2 méthodes donneraient des prévisions différentes,

- et une affectation d'un degré de confiance à la prévision.

0n aurait aussi pu mettre en sortie multiples des sous-ensembles de bassins (voire la totalité) pour bénéficier d'emblée de la covariation spatiale du risque pluviométrique. Cependant, avec le même fichier d'entrée, le nombre de poids w à ajuster rendrait le modèle beaucoup moins robuste.

IV.5 Conclusion du chapitre IV

Si l'utilisation des données brutes aux stations de RadioSondages (RS), au lieu des Composantes Principales (CP), n'apporte pas un gain considérable en terme de prévision de pluie (+ 0.6 % en pluie / non pluie et - 1.8 en prévision par classes - RPS x 100 -), elle met cependant en doute l'intérêt des CP quand la condensation des données n'est pas indispensable.

Pour travailler sur des données les plus comparables possible à ce qui se fait actuellement, à savoir des valeurs en points de grille, une grille de 1° x 1° a été reconstituée par interpolation des données de radiosondage.

Et l'utilisation d'un nouveau critère de sélection des analogues, le score de Teweles-Wobus (1954) mesurant la ressemblance entre 2 cartes, a donné des résultats relativement intéressants après optimisation de la grille à utiliser:

- la grille optimale est de taille régionale (intermédiaire), légèrement décalée vers le sud-ouest par rapport à la zone d'étude pour la période d'automne considérée. En effet, en automne, la majorité des pluies proviennent soit de la Méditerranée, soit des grands systèmes dépressionnaires de l'Atlantique.

- les performances sont un peu supérieures à celles obtenues par les variables orthogonalisées (méthode S-12CP) et les variables brutes (S-12RS), ce qui est probablement dû à une meilleure adéquation du critère d'analogie (cf. tableau IV-9):

	prévision pluie / non pluie IR	prévision en classes RPS * 100
référence	72.2	59.7
méthode S-12CP	78.6	49.0
méthode S-12RS	79.2	47.2
méthode TW-GR	80.1	45.05

tableau IV-9 : les différentes méthodes

Finalement, rien qu'en travaillant sur les données disponibles à ce jour dans le fichier EDF initial, cela nous a permis d'augmenter les performances de prévision de la méthode par analogue de près de 8 % en pluie / non pluie, mais aussi en prévision probabiliste puisque le RPS, qui peut varier entre 0 et 2, a baissé de près de 0.15 soit un gain de 7.5%.

Nos deux conclusions fortes sont donc :

- **qu'il fallait améliorer les critères d'analogie**, et qu'il faut désormais implanter en mode opérationnel ces versions améliorées,

- qu'il serait opportun aujourd'hui de travailler sur **des données sous forme de grilles**.

Pourtant, celles que nous avons utilisées ne sont sans doute pas optimales. En effet, nous les avons élaborées « brutalement », par application d'une technique d'interpolation « aveugle » (fonction spline) sur les radiosondages.

On peut penser que des grilles analysées, filtrées par des modèles qui forcent à respecter certaines contraintes physiques, sont de meilleure qualité. Cependant, au moment de la réalisation de ce travail, de telles grilles n'étaient disponibles que de 1981 à 1995 au CEPMMT. Depuis, le NCEP et le NCAR (National Centers for Environmental Prediction et le National Center for Atmospheric Research) ont réanalysé et homogénéisé toutes les données possibles depuis 1955, à tous les niveaux standards. C'est probablement la source de données à envisager en priorité aujourd'hui.

Mais, pour l'instant, c'est l'apport de nouvelles variables sur lesquelles faire l'analogie, tant au niveau synoptique qu'an niveau local, qui nous a semblé pouvoir être bénéfique. C'est ce que nous verrons au chapitre V.

CHAPITRE V :

INTRODUCTION et ELABORATION

de NOUVELLES VARIABLES

Introduction

Après avoir tenté d'exploiter au maximum les données disponibles sous leurs différentes formes (brutes, grilles, CP), nous avons cherché à introduire de l'information nouvelle pour sélectionner les analogues, tout d'abord par l'intermédiaire de variables de grande échelle (différents champs de géopotentiel, température, humidité).

Nous avons ensuite tenté d'introduire de l'information locale (données d'un radiosondage) pour affiner l'analogie.

V.1 Variables synoptiques *

V.1.1 Présentation des données

Les données, récupérées auprès de Météo-France, proviennent de l'archive HEMIS. Ce sont des données analysées de

- ***géopotentiel*** à 1000, 700 et 500 hPa (Z1, Z7 et Z5),

- ***température*** à 850 et 500 hPa (T85 et T5),

- *et d'**humidité relative*** à 900, 700 et 500 hPa (U9, U7 et U5),

en points de grille à 00 et 12 TU avec quelques données manquantes. Elles proviennent de 3 origines différentes:

- archives américaines du N.M.C. (National Meteorological Center) du 01/01/1963 au 06/01/1970,

- archives françaises du S.C.E.M. (Service Central d'Exploitation Météorologique) du 02/04/1969 au 30/11/1980,

- archives européennes du 01/11/1980 au 31/12/1985.

Leur homogénéité est donc douteuse.

La période d'archivage va du 01/01/1963 au 31/12/1985 pour le géopotentiel et la température avec une exception pour la température à 1000 hPa qui commence le 02/04/1969. Quant à l'humidité, elle débute le 14/01/1975.

Le domaine couvert représente une partie de la grille octogonale N.M.C. qui couvre l'hémisphère nord pour les latitudes supérieures à 20° nord (cf. figure V-1).

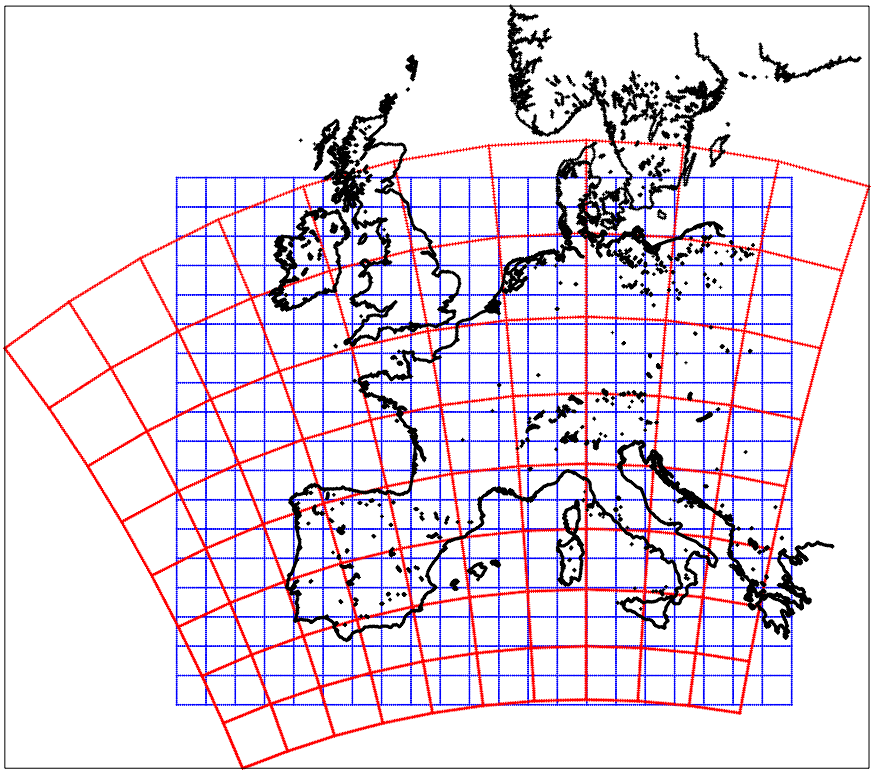


figure V-1: domaine géographique des données de l'archive HEMIS

Un contrôle de qualité succinct des données a été effectué par Météo-France jusqu'au 30/11/1980 mais il ne tient pas compte des ruptures d'homogénéité qui peuvent exister dans l'évolution chronologique des données. Aussi, avons-nous comparé les moyennes mensuelles des différents champs à 00 et 12 h pour les 3 types d'archives. Des résultats identiques ayant été trouvés à 00 et 12 h, seules les courbes à 00 h ont été portées en annexe V-1. Un seul problème a été décelé pour l'humidité à 900 hPa (cf. figure V-2): cela nous a permis de nous rendre compte que l'archive européenne était composée de données d'humidité à 850 hPa au lieu de 900.

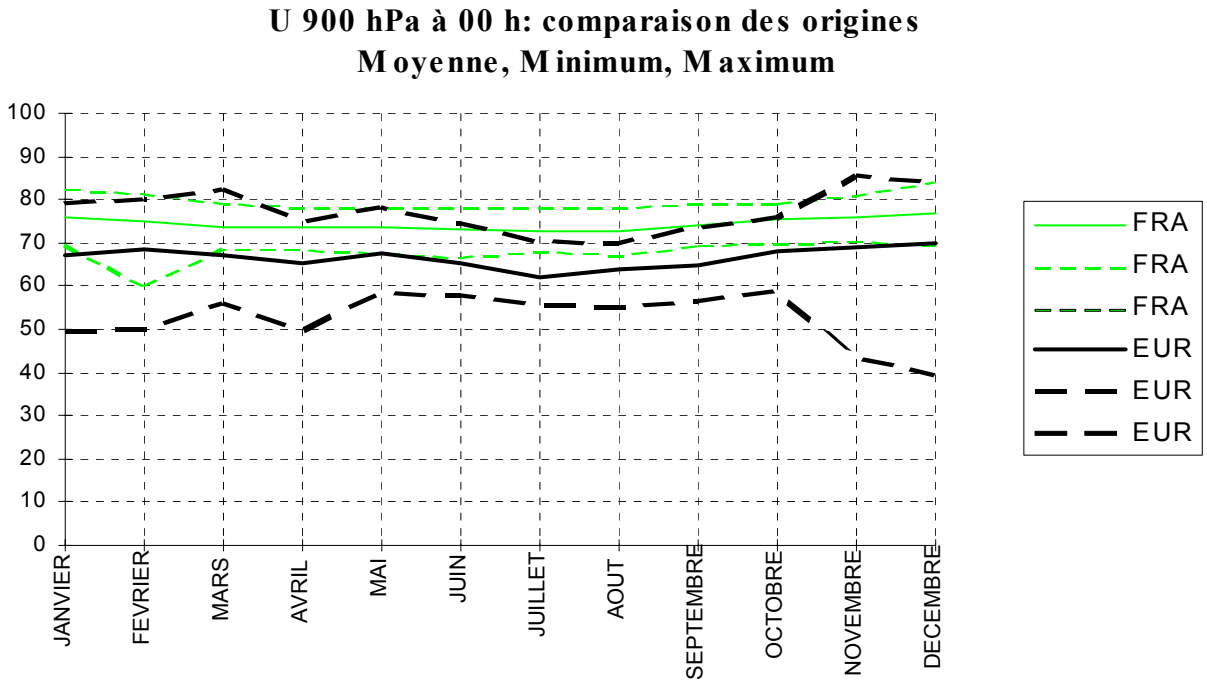


figure V-2: moyennes mensuelles pour l'humidité à 900 hPa à 00 h

Enfin, Météo-France nous a signalé que :

- la qualité des données de géopotentiel et de température était relativement correcte (tableau en annexe V-2),

- celle des champs d'humidité l'était beaucoup moins, en particulier jusqu'en octobre 1980 (cf. annexe V-2). De plus, leur correction, ou du moins l'estimation de leur qualité est rendue difficile par la grande variabilité spatiale et temporelle de ce paramètre.

- les géopotentiels 1000 hPa de la période américaine ont été reconstitués à partir des géopotentiels à 500 hPa et de l'épaisseur de la couche 1000 / 500 hPa.

D'un point de vue statistique, afin de mieux connaître le type de données, nous avons pu faire différentes remarques sur les champs moyens d'automne, puisque c'est cette période que nous étudions (cf. annexe V-3):

- *sur l'autocorrélation temporelle (annexe V-3a) :* celle des champs de température est bonne et diminue relativement lentement: R passe de 0.85 à 12 h à 0.5 à 96 h. Elle diminue beaucoup plus rapidement pour les champs de géopotentiel puisque pour celui à 1000 hPa, si elle

est de R=0.85 à 12 h, elle est déjà inférieure à 0.5 à 36 h. Par contre elle est mauvaise dès 12 h pour l'humidité (R< 0.5).

- *sur les effets saisonniers :* la température, comme nous avons pu le constater sur les graphes de l'annexe V-1, possède un effet saisonnier important qui n'existe pas pour l'humidité. Pour les géopotentiels, s'il apparaît pour les couches élevées (500 et 700 hPa), il est inexistant pour les basses couches (1000 hPa).

- *sur l'effet diurne (annexe V-3b) :* il est très important pour la température, moindre pour les géopotentiels et nul pour l'humidité.

V.1.2 Utilisation de ces champs

Tout d'abord, ces champs ont été utilisés seuls dans le critère de Teweles-Wobus pour sélectionner les analogues. Puis, comme cela a été fait au chapitre IV (cf. § IV.2.5), différentes combinaisons de champs ont été testées pour la prévision probabiliste en classes:

V.1.2.a 1 Champ, 1 Echéance (00, 12, 24 ou 36 h)

Le critère de sélection S des analogues est un simple critère de Teweles-Wobus pour un champ C et une échéance E, $S = TW_C^E$.

La figure V-3 présente les résultats pour la période 1975-1985, où l'on trouve en ordonnée le Ranked Probability Score ou RPS (qui doit être le plus faible possible pour améliorer la qualité des prévisions) en fonction du champ et de l'échéance considérée, avec comme référence la méthode TW-GR sur cette même période. Pour la période 1963-1985, se reporter à l'annexe V-4a.

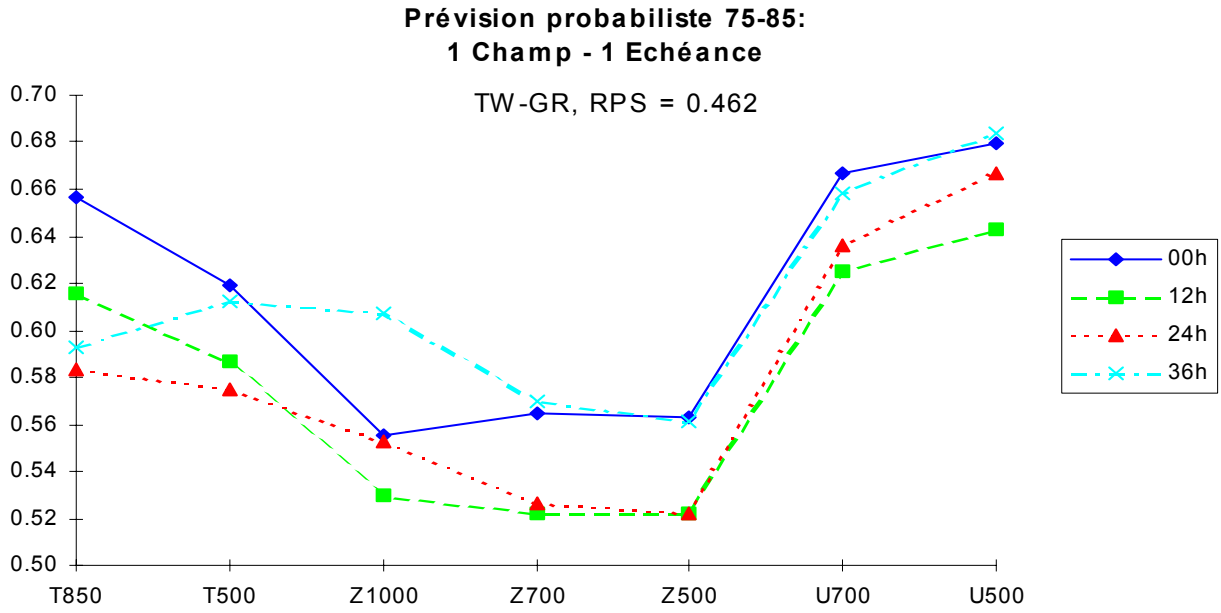


figure V-3 : un Champ, une Echeance, période 1975-1985

On peut rapidement noter que :

- l'humidité utilisée seule n'est pas intéressante pour prévoir la pluie,

- les meilleures performances sont obtenues, pour les deux périodes, avec les champs de géopotentiel à 12 ou 24 h mais on est loin des performances de la méthode TW-GR.

Aussi, avons-nous cherché à faire l'analogie sur un plus grand nombre de champs.

Remarque: la méthode TW-GR voit son score RPS augmenter si l'on diminue trop le fichier historique. Si le passage de 41 à 23 ans n'a pas apporté de grand changement (respectivement 45.1 et 45.0) on peut penser qu'un seuil d'une vingtaine d'années est suffisant. Par contre, lorsque l'on arrive à 11 ans de données, RPS = 46.2

V.1.2.b 1 Champ, 2 Echéances

Cette fois le critère de sélection est la somme de 2 critères de Teweles-Wobus avec un même champ et 2 échéances E_1 et E_2 :

$$S = TW_C^{E_1} + TW_C^{E_2} \qquad \text{(V-1)}$$

On obtient les meilleures performances (cf. figure V-4 et annexe V-4b) avec les champs de géopotentiel 700 ou 500 hPa aux échéances 12 et 24 h:

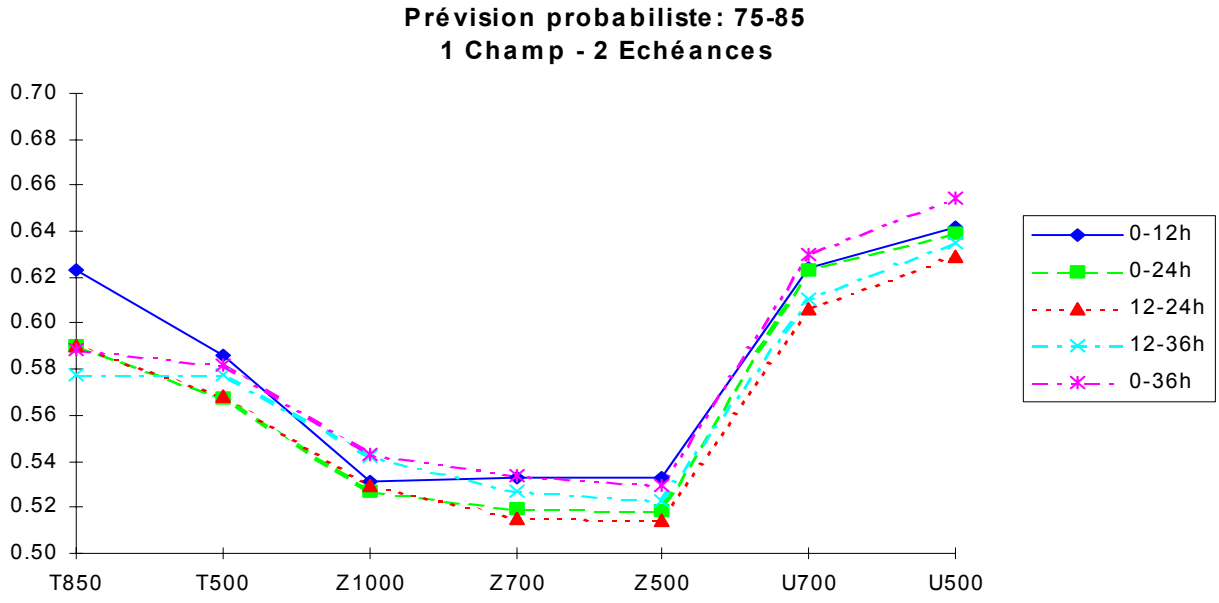


figure V-4: un Champ, 2 Echéances, période 1975-1985

V.1.2.c 2 Champs, 1 Echéance par champ

Le critère de sélection est pratiquement le même que précédemment mais cette fois avec 2 champs C_1 et C_2 et une seule échéance E :

$$S = TW_{C_1}^{E} + TW_{C_2}^{E} \qquad \text{(V-2)}$$

Tous les couples possibles de champs ont été testés pour les échéances 0, 12 et 24 h (cf. figure V-5 et annexe V-4c), les meilleurs étant (Z1-Z7) et surtout (Z1-Z5) à 12 h. D'ailleurs, à ce sujet, on se rappelle qu'aux débuts de la méthode, il y avait eu une petite hésitation quant au choix entre la 500 et la 700 hPa pour faire l'analogie (cf. chapitre II, § II.1.1).

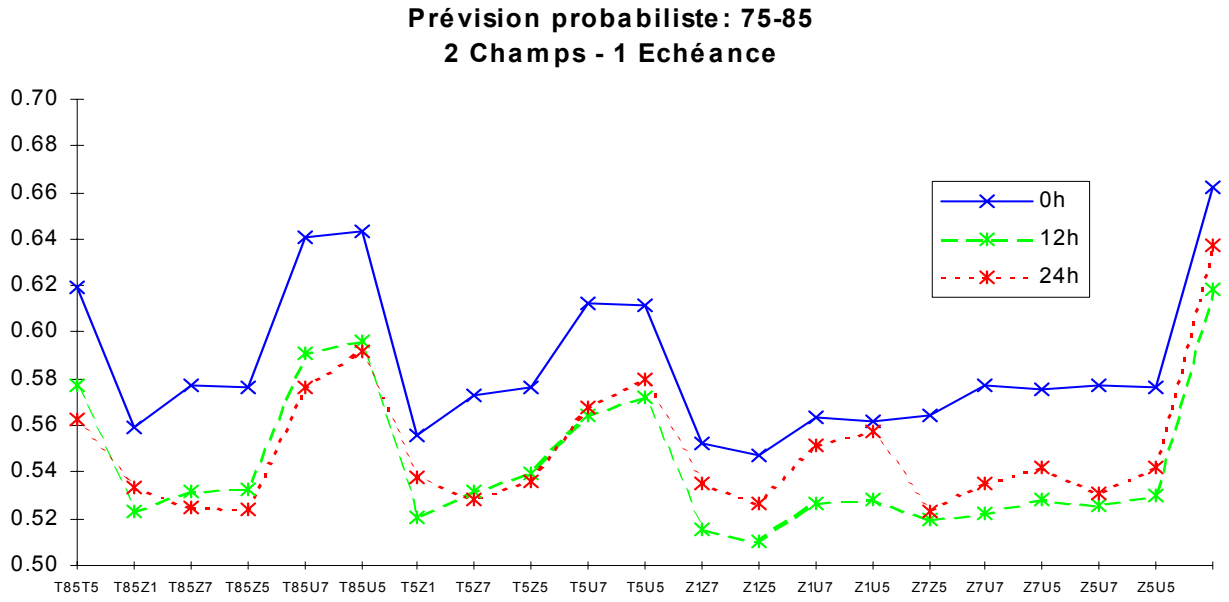


figure V-5: 2 Champs, 1 Echéance, période 1975-1985

Si l'on compare les performances obtenues à celles du cas précédent (1 champ, 2 échéances), il en ressort qu'il vaut mieux utiliser deux champs à une échéance que le contraire. Il faut donc privilégier l'apport d'information nouvelle, moins corrélée à celle du premier champ que ne peut l'être un même champ à une autre échéance.

V.1.2.d 2 Champs, 2 Echéances par champ

Nous sommes ensuite passés à la configuration ayant abouti à la méthode TW-GR: un critère de sélection des analogues composé de 4 scores de Teweles-Wobus avec 2 champs C_1 et C_2 et 2 échéances E_1 et E_2 par champ.

$$S = TW_{C_1}^{E_1} + TW_{C_1}^{E_2} + TW_{C_2}^{E_1} + TW_{C_2}^{E_2} \quad \text{(V-3)}$$

Là encore, les meilleurs résultats sont obtenus avec les champs de géopotentiel et en particulier 1000 et 500 hPa à 12 et 24 h (cf. figure V-6 et annexe V-4d).

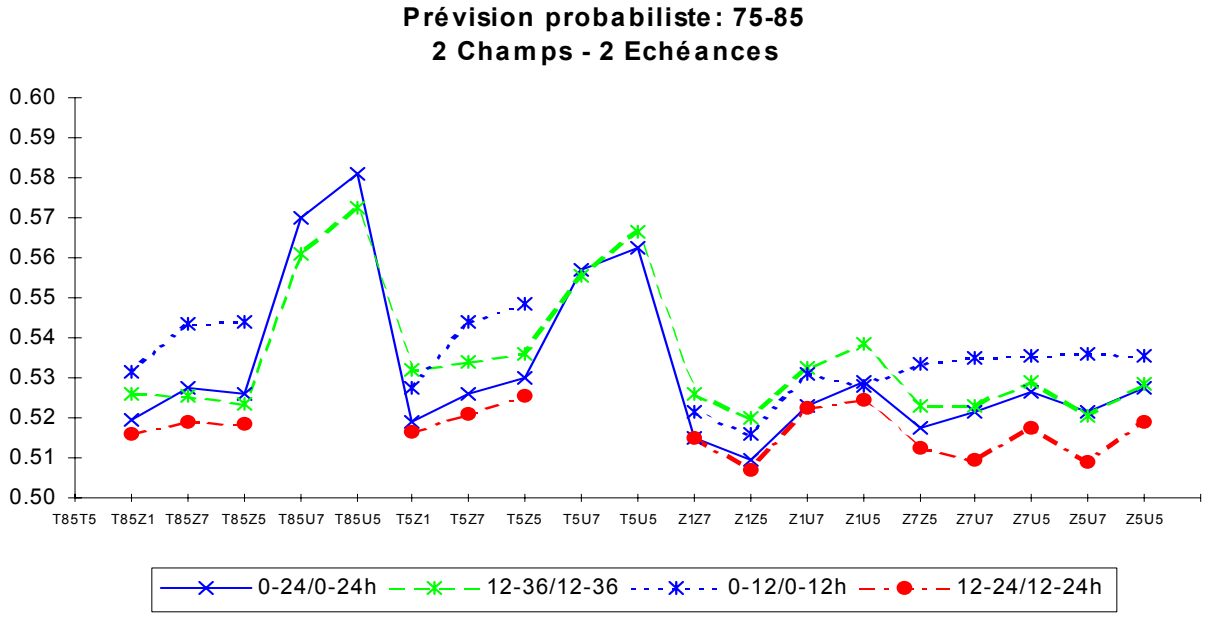


figure V-6: 2 Champs, 2 Echéances, période 1975-1985

Cependant, malgré toutes les combinaisons envisagées, les meilleures performances restent nettement en-deçà de la méthode TW-GR. En particulier, les champs de température et d'humidité n'ont pas l'air d'apporter de l'information pertinente, et surtout l'humidité. Pourtant celle-ci semble primordiale pour la génération des pluies : il paraît en effet évident que deux masses d'air analogues ne vont pas réagir de la même façon si leur humidité est différente. Aussi, avons-nous voulu l'utiliser, non pas à la place d'une variable, mais en plus des couples les plus intéressants.

V.1.2.e 3 Champs, 2 Echéances

Pour aller plus loin que la méthode TW-GR, un troisième champ a été rajouté, avec les 2 échéances 12 h et 24 h puisque ce sont les meilleures.

Seuls les triplets associant les meilleurs couples du cas précédent (Z1-Z5, Z1-Z7 et Z5-Z7) et une autre variable, en particulier l'humidité, ont été testés. Et cette fois, sur la période 1975-1985 (cf. figure V-7), c'est l'humidité à 700 hPa qui ressort, en combinaison avec Z1-Z5 ou même Z5-Z7. Par contre, pour 1963-1985 (cf. annexe V-4e), le meilleur essai est obtenu avec la température à 850 hPa et Z1-Z5.

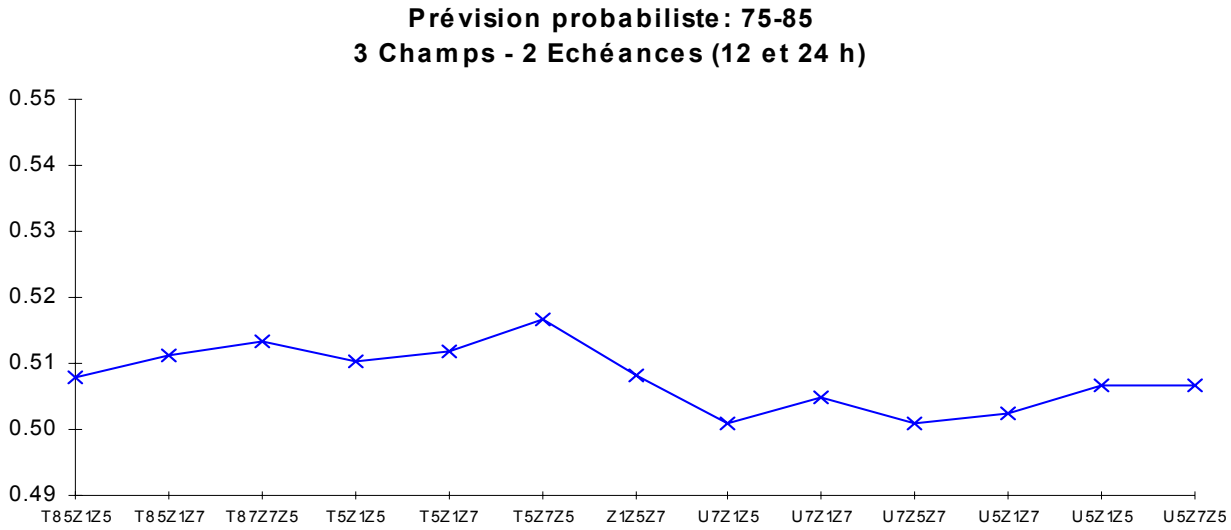


figure V-7: 3 Champs, 2 Echéances, période 1975-1985

V.1.3 Conclusion

Pour synthétiser tous ces essais, le meilleur de chaque cas a été conservé et placé sur le graphe de la figure V-8, ceci pour les 2 périodes.

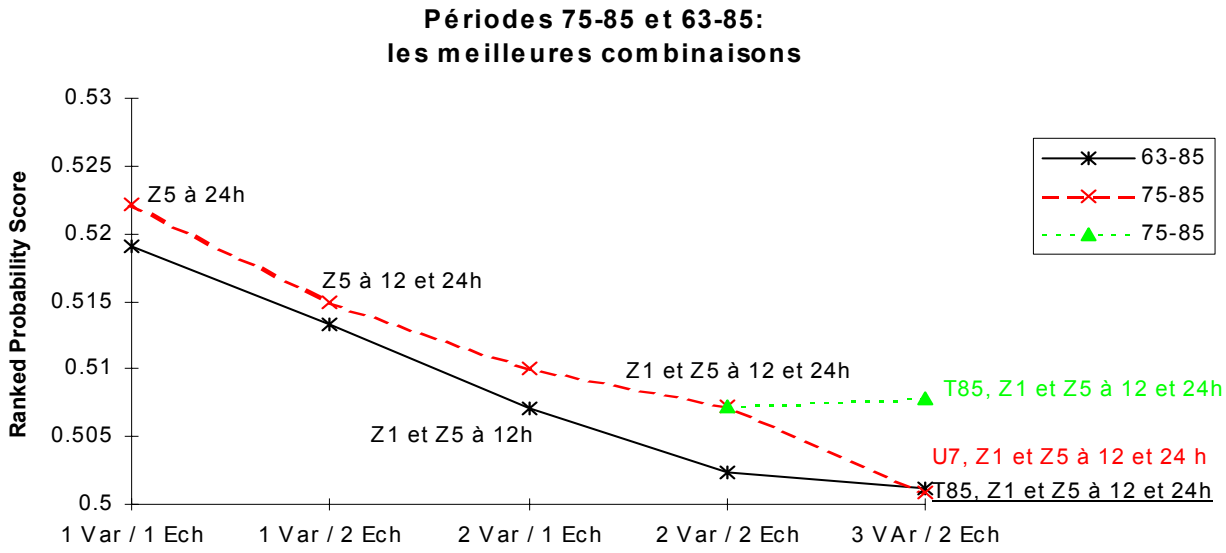


figure V-8: synthèse des résultats, périodes 1975-1985 et 1963-1985

Certains enseignements peuvent en être tirés :

i) Tout d'abord les résultats restent toujours inférieurs à ceux de la méthode TW-GR, dont la grille est sans doute mieux adaptée à la zone d'étude (cf. figure V-1) mais aussi dont le fichier historique est le plus long et donc le plus riche.

ii) Cependant, les résultats étant un peu meilleurs avec les champs de géopotentiel 1000 et 500 hPa qu'avec ceux à 1000 et 700, il pourrait être envisagé de remplacer le géopotentiel 700 hPa par celui à 500 hPa sur la même grille dans la méthode TW-GR.

iii) Enfin, l'introduction d'une information sur l'état hygrométrique des masses d'air semble nécessaire pour améliorer la prévision. En effet, sur la période 63-85 où l'on ne possède pas de données d'humidité, le fait d'ajouter une troisième variable ne permet pas d'augmenter les performances qui plafonnent. Par contre, sur 75-85, l'introduction de l'humidité compense la diminution de la taille du fichier historique qui est pourtant loin d'être négligeable (divisé par 2 !), puisque le score devient égal à celui obtenu sur la période 63-85, sans l'humidité.

On peut d'ailleurs penser que si l'on avait pu utiliser l'humidité en basses couches (900 hPa), les résultats auraient été encore meilleurs. Nous l'avons vérifié sur les 6 années (1975-1980) de données disponibles pour l'humidité à 900 hPa (U9) (cf. tableau V-1).

Triplets de variables sur 75-80	Prévision probabiliste RPS * 100
Z1 - Z5 - U5	53.1
Z1 - Z5 - U7	52.5
Z1 - Z5 - U9	51.7
TW-GR (75-80)	**48.7**

tableau V-1 : période 75-80 avec U9

On se rapproche donc du résultat obtenu par la méthode TW-GR pour cette période**.** Cependant, les gains obtenus ici en ajoutant une information hygrométrique sur les masses d'air ne compensent pas ceux perdus par la diminution du fichier historique.

Néanmoins, on dispose maintenant d'une perspective intéressante pour récupérer de telles données sur une plus longue période avec les fichiers NCEP/NCAR, réanalysés par un même modèle sur la période 1955-1995.

V.2 Variables à caractère local

Un problème soulevé par l'approche par analogues concerne le nombre de niveaux qu'elle doit comporter. A priori, on postule que deux situations synoptiques semblables ont des chances de générer des effets locaux semblables. Pourtant, ces effets locaux dépendent à la fois de conditions synoptiques (circulation générale) mais aussi de conditions locales, d'où l'idée d'introduire une *analogie au niveau local*. En effet, si l'occurrence des précipitations et leur intensité sont liées à la situation atmosphérique globale à l'échelle du continent, elles sont aussi en rapport avec des paramètres locaux très variables dans le temps et l'espace, paramètres qui sont mal représentés par les premières Composantes Principales (CP) des champs météorologiques ou par les points de grille.

Idéalement, l'analogie devrait être globale, c'est-à-dire concerner toutes les variables informatives concernant le prédictand considéré, la pluie en divers bassins. Par « toutes les variables », il faut entendre à toutes les échelles et en respectant leur chaîne de causes à effets. Cela n'est bien sûr pas possible dans la mesure où le chaînage est imparfaitement connu et où certaines variables nécessaires à sa description sont indisponibles.

Comme cette analogie ne saurait être globale, nous l'avons hiérarchisée en deux niveaux: le niveau synoptique et le niveau local. L'idée directrice serait donc de faire une sélection des analogues en deux temps:

i) analogie synoptique*:* sélection de N analogues grâce à un critère de sélection utilisant des variables synoptiques (cf. méthodes S-12CP, S-12RS, TW-GR),

ii) analogie locale: sur ces N analogues, un deuxième critère de sélection est activé, avec des informations à caractère local, pour en extraire $N_2 < N$.

Cela a déjà été testé de manière ponctuelle par Mandon (1985) et Vallée (1986) qui ont fabriqué des critères de sélection avec des données du radiosondage de Nîmes (vent et humidité en particulier), critères activés uniquement lorsque le modèle prévoit un risque de fortes pluies sur les Cévennes (cf. chapitre I, § I.4).

Cependant, leurs résultats ont été testés sur un ensemble de quelques situations météorologiques bien particulières, les « épisodes cévenols ». Aussi, ayant récupéré les données du radiosondage de Nîmes de 1954 à 1983, nous avons voulu les utiliser de manière plus systématique.

V.2.1 Le radiosondage de Nîmes

V.2.1.a Les données brutes

Pour cette étude nous disposons des données journalières du radiosondage de Nîmes à 00 et 12 heures TU pour la période 1954-1983, soit 30 ans. Il contient, pour différents niveaux de pression, des informations de

- *géopotentiel:* altitude en mètres géopotentiels à laquelle on retrouve la pression considérée,

- *température de l'air*,

- *température du point de rosée*, température à laquelle une particule d'air humide devient saturée par refroidissement isobare et à pression de vapeur constante,

- *humidité relative* (en pourcentage),

- *rapport de mélange à saturation*, qui représente la masse maximale de vapeur d'eau qu'il est possible d'associer à l'unité de masse d'air sec dans une particule,

- *direction et force du vent*.

Quelques détails supplémentaires sont donnés, en annexe V-6a, sur ces variables.

Pour les données de vent, les relevés ne sont disponibles que jusqu'en 1964, date à laquelle un fichier séparé (le CRV, compte-rendu du vent) a été créé pour ces données. De plus, de trop nombreuses données manquantes n'ont pas permis de les utiliser de manière systématique.

Pour notre étude nous avons retenu les données aux niveaux standards 1000, 950, 900, 850, 800, 700, 600, 500, 400 et 300 hPa, c'est-à-dire aussi bien des niveaux situés dans les basses couches de l'atmosphère que des niveaux élevés (jusqu'à 300 hPa) de manière à découvrir le niveau le plus informatif pour les pluies extrêmes. Pour les mêmes raisons nous traiterons les données à 00 et 12 h afin de juger de l'heure la plus informative.

Quant au choix du radiosondage de Nîmes, il a été imposé par sa disponibilité mais aussi par sa position géographique au sud des Cévennes. Il est bien placé pour voir venir les coups de sud qui produisent les fameux « épisodes cévenols ». Les conditions favorables à ces pluies intenses, détaillées par Mandon (1985), ont montré que la station de Nîmes semble être toute indiquée pour fournir des renseignements sur l'état hygrométrique et thermique des masses d'air qui abordent les Cévennes.

V.2.1.b Les différentes index : indices d'instabilité et autres paramètres**

synthèse : Ces différents index seront utilisés de la même façon que les données brutes, mais des résultats moins intéressants seront obtenus (cf. § V.2.4).

Dans un deuxième temps, un certain nombre d'index ont été choisis et calculés par l'équipe de Physique Atmosphérique du Pr. Carmen Llasat de l'Université de Barcelone, dans le cadre d'un programme d'action intégrée PICASSO. Ils sont censés être plus corrélés avec les fortes pluies et devraient donc être plus informatifs que les données brutes pour les pluies et en particulier les pluies extrêmes. En effet, pour chacun il existe une échelle numérique établissant l'équivalence entre la valeur de l'indice et la probabilité de phénomènes violents convectifs.

Les index utilisés sont présentés dans le tableau V-2, inspiré de celui-ci de Sénési et Thépenier (1993) qui les utilisent pour la prévision d'orage à des échéances plus courtes. Ces index peuvent être distingués en trois types (cf. Sénési & Thépenier, 1993) :

i) les indices d'instabilité définis par divers auteurs. C'est le cas de :

* *l'indice de Showalter SI* (Showalter, 1953)
* *le Lifted Index ou indice de Galway LI* (Galway, 1956)
* *l'indice K* (Georges, 1960)
* *l'indice « Total Totals »* (Miller, 1967)
* *l'indice SWEAT* (Bidner, 1970 et Miller et. al., 1971)
* *l'instabilité potentielle humide Iph* (David & Smith, 1971).

Nom	Formule	Commentaires
Indice de Showalter	$SI = T_{500} - Tp_{500}$	différence entre la température à 500 hPa du point d'état (T_{500}) et celle de la particule du niveau 850 hPa soulevée adiabatiquement jusqu'à son niveau de condensation puis pseudo-adiabatiquement jusqu'à 500 hPa (Tp_{500})
Lifted Index ou indice de Galway	$LI = T_{500} - Tpx_{500}$	comme SI mais calculé en utilisant la particule sol ayant la température maximale de la journée (Tpx_{500}) et le rapport de mélange moyen des 1000 premiers mètres
Indice K	$K = T_{850} + Td_{850}$ $-T_{700} + Td_{700} - T_{500}$	Td : température du point de rosée
Indice Total Totals	$TTI = T_{850} - T_{500}$ $+ Td_{850} - T_{500}$	
Indice SWEAT	$SWEAT = 12*Td_{850} + 20*(tt-49) + 2*f_{850} + f_{500} + 125*(S+0.2)$ $S = \sin(dd_{500} - dd_{850})$	Severe WEAther Threat: combinaison linéaire du cisaillement, de l'instabilité et de la force du vent en basses couches
Instabilité potentielle humide	$Iph = \theta'_{w850} - \theta'_{w500}$	différence entre la température potentielle humide θ'_w des niveaux 850 et 500 hPa
Cisaillement vertical moyen	$U = \lvert\lvert V_{sol\text{-}500} - V_{sol\text{-}6000} \rvert\rvert$	différence entre les vents moyens des couches sol-500 m et sol-6000 m
CAPE (cf. annexe V-6b)	$g \int_{NCL}^{NE} \frac{\theta_p - \theta_a}{\theta_a} dz$	Energie potentielle convective disponible θ_p, θ_a températures. potentielles de la particule et de l'environnement ; NE niveau d'équilibre ; NCL niveau de convection pour chauffage, $g=9.8\ m/s^2$
Nombre de Richardson	$R = 2CAPE/U^2$	rapport sans dimension de stabilité statique et de cisaillement vertical moyen U
Masse d'eau précipitable	$MAP = \bar{r}.\Delta p / g$	masse d'eau maximale que la couche d'épaisseur Δp est capable de transformer en pluie (r rapport de mélange moyen en g/kg, $g=9.8\ m/s^2$)
Vitesse maximale verticale	$Wmax = (2\ CAPE)^{0.5}$	
Isotherme 0°C	P, Alt	Pression et ALTitude à 0°C

tableau V-2 : les différents index

ii) des paramètres liés à la théorie du soulèvement de la particule :

* *l'énergie potentielle convective disponible ou CAPE* pour Convective Available Potential Energy (Moncrieff & Miller, 1976)

* *le nombre de Richardson Ri*

* *la vitesse maximale verticale* en m/s.

iii) des paramètres de base :

* *le cisaillement entre 500 et 6000 m*

* *l'isotherme 0°C* (*pression* et *altitude*)

- *la masse d'eau précipitable* (kg/m^2).

Remarque : dans le tableau V-2, θ'_w est la température potentielle humide, Lfc le niveau de convection libre (cf. annexe V-6a), T_v la température virtuelle de la particule soulevée et T_v celle de l'environnement.

V.2.2 Méthodologie

Une analogie locale a été mise en place dans un premier temps après celle, synoptique, de la méthode S-12CP. Mais comme cette dernière a été calée, pour chaque bassin, sur la période 1953-1993, il a fallu la valider sur la période 1954-1983 pour laquelle les données de Nîmes sont disponibles. C'est-à-dire que, pour chaque bassin, la distance euclidienne optimisée sur la période 1953-1993 par sélection ascendante des CP a été reprise pour faire la sélection des analogues et la prévision sur la période 1954-1983. La comparaison des performances pour les 2 périodes est donnée en annexe V-7.

Des résultats à peu près équivalents ont été obtenus voire légérement meilleurs sur la période 1954-1983, en particulier grâce à 4 bassins (Vézère-Vienne-Thaurion, Gaves, Ain-Valserine et Arve-Fier) qui ont vu leurs scores augmenter de manière significative, que ce soit en prévision d'occurence de pluie ou en prévision probabiliste en classes. Ce résultat est un peu surprenant car on se serait plutôt

attendu à une perte de performance due à la diminution de la taille du fichier historique. On peut néanmoins penser que cela vient du fait que la sélection ascendante effectuée sur les CP à introduire dans le critère de sélection des analogues n'est pas optimale ce qui peut entraîner un biais.

Nous avons ensuite greffé à la suite de cette sélection d'analogues une analogie locale. Ainsi, pour chaque jour test (jour C) de 1954 à 1983, 50 analogues sortent grâce à la distance euclidienne de la méthode S-12CP. Sur ces 50 analogues, une nouvelle sélection est effectuée, utilisant l'information locale par l'intermédiaire d'une autre distance euclidienne de la forme:

$$D^2(J,C) = \sum_{i=1}^{Pl} \left[Vl_i(C) - Vl_i(J) \right]^2 \qquad (V\text{-}4)$$

avec Pl variables locales Vl.

Et de la même manière qu'au chapitre III, une sélection ascendante est effectuée mais cette fois sur les variables locales, le critère d'analogie synoptique étant fixé une fois pour toute (celui de S-12CP dans l'immédiat).

Remarque: il a fallu gérer le problème des données manquantes dans le radiosondage de Nîmes. Notre choix a été le suivant:

- si la journée courante C n'a pas de données de Nîmes, la prévision est faite avec les 50 analogues de la méthode S-12CP (pas de second niveau),

- si c'est la journée analogue J, nous avons choisi de la conserver quand même.

V.2.3 Utilisation des données brutes de Nîmes

Nous avons commencé par utiliser comme variables d'intérêt local les données brutes du radiosondage de Nîmes que nous rappelons ici:

- la température **T**,

- la température du point de rosée **Td**,

- l'humidité **H**,

- le rapport de mélange à saturation **R**,

- l'altitude **ALT**,

en chacun des 10 niveaux de pression standards choisis. Nous y avons rajouté l'écart entre T et Td (**T-Td**), autre bon indice de l'humidité de la masse d'air.

Cela fait donc 6 x 10 = 60 variables à introduire dans la sélection ascendante. Elles ont été au préalable centrées et réduites afin de les rendre comparables avec des poids équivalents.

Remarque: tous les graphes sont donnés pour la prévision probabiliste en classes mais des résultats similaires ont été obtenus avec la prévision d'occurence de pluie. Aussi, les différentes conclusions, mais ci ce n'est pas explicite, sont valables pour les 2 types de prévision.

V.2.3.a Choix du nombre d'analogues N_2

L'optimisation du nombre $N_2 < N$ analogues à retenir par l'analogie locale s'est faite sur 9 bassins témoins: les 6 du chapitre III (Creuse-Cher, Pyrénées Est, Doubs, Isère moyenne, Var-Tinee-Roya et Chassezac) plus 3 autres, situés autour de Nîmes (Haut Tarn-Haut Lot, Agout-Tarn et Durance moyenne) pour lesquels l'information locale de Nîmes devrait être intéressante.

Deux exemples de résultat graphique sont donnés sur la figure V-9 et le reste en annexe V-7. Le trait horizontal représente le score de la seule analogie synoptique (méthode S-12CP sur la période 1954-1983). Quant aux autres, ils montrent l'évolution des scores en fonction du nombre de variables locales introduites dans le critère d'analogie locale (cf. équation V-4) lorsque le nombre d'analogue N_2 vaut 10, 20, 30 ou 40.

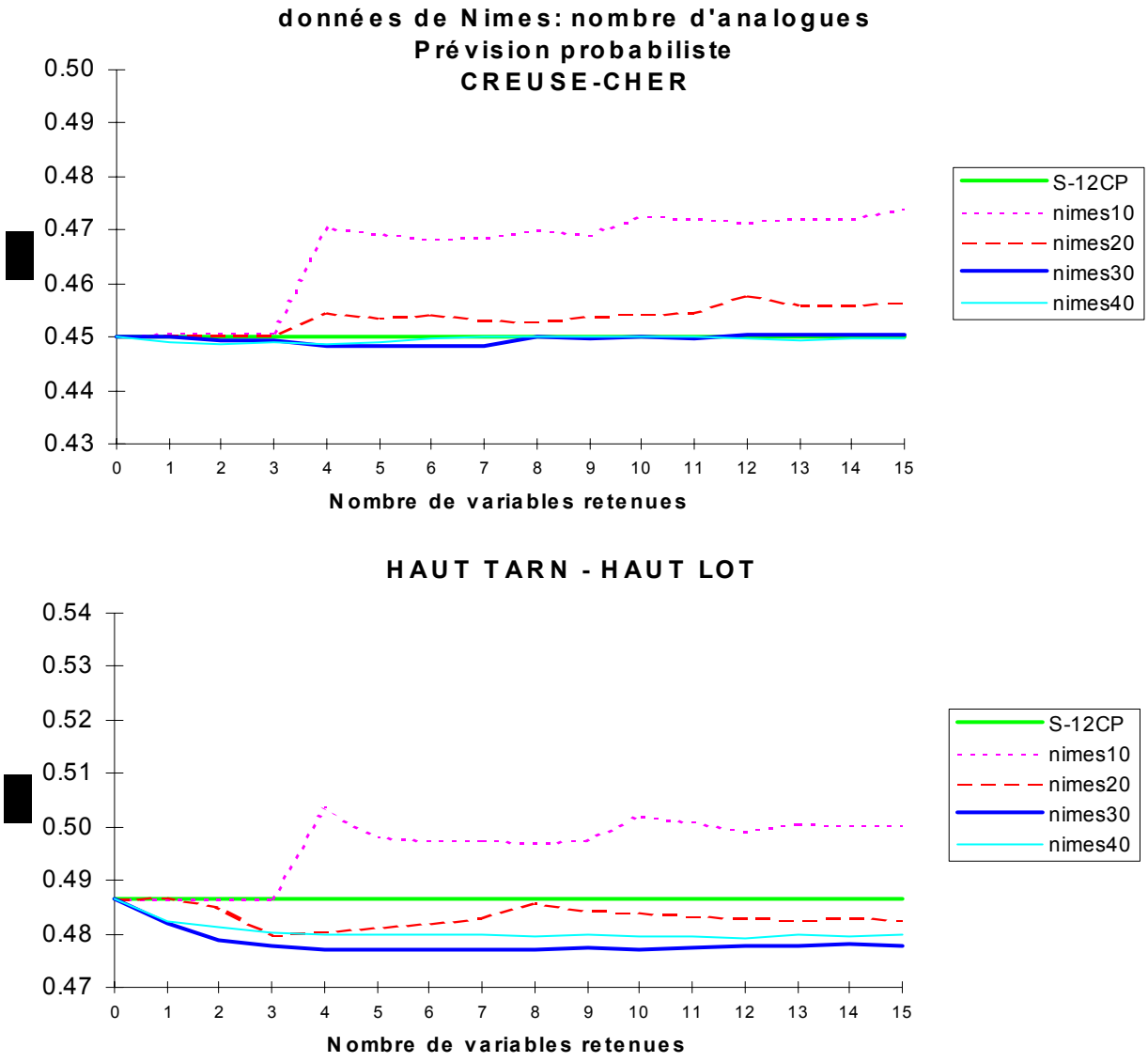


figure V-9: Choix du nombre d'analogues

Le choix d'un nombre $\mathbf{N_2 = 30}$ analogues apparaît comme évident pour l'ensemble des 9 bassins testés. C'est celui qui sera utilisé par la suite.

On peut aussi noter qu'un palier semble atteint autour de 6 variables retenues, ce que nous vérifierons ci-après. En outre, tous les bassins ne réagissent pas de la même manière à l'apport d'information locale. Ainsi, par exemple, le bassin Creuse-Cher, situé loin de Nîmes, semble ne pas subir son influence puisqu'aucun gain n'est apporté par l'analogue locale. Par contre, les bassins plus proches de Nîmes y sont plus sensibles. Mais nous verrons cela de manière plus approfondie dans les paragraphes suivants.

V.2.3.b Choix de l'échéance - Détermination du palier

Les résultats du paragraphe précédent proviennent des données de Nîmes à 00 h. Cependant, comme les données à 12 h sont disponibles, nous avons voulu les utiliser car elles sont mieux situées par rapport à la pluie, tout comme les données à 24 h en prévision parfaite (cf. figure V-10 et annexe V-8). Mais ces dernières (à 24 h) ont été très vite abandonnées car il faudrait les utiliser en prévision, ce qui dégraderait trop les performances. Par contre les données à 00 h semblent les plus intéressantes.

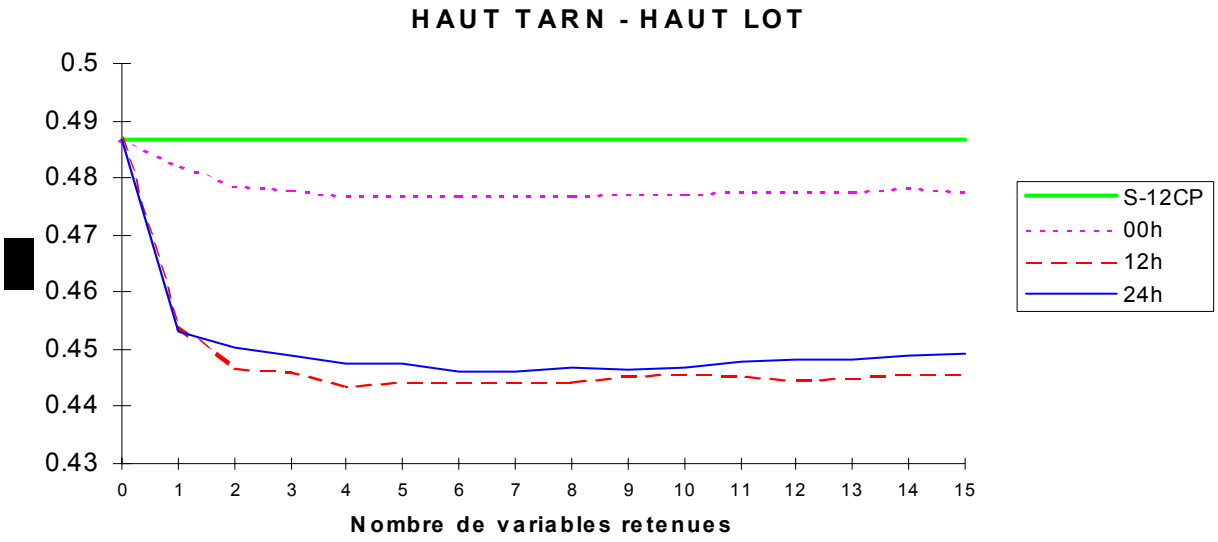


figure V-10: Choix de l'échéance

La sélection ascendante des variables locales à 00 et 12 h, effectuée sur les 33 bassins français, a permis de déterminer un palier avec 6 variables locales (cf. figure V-11). Au-delà, l'introduction d'autres variables n'apporte pas de gain significatif. Et l'on retrouvera en annexe V-9 les scores obtenus par bassin pour 6 variables locales.

Enfin, les résultats sont toujours meilleurs lorsque les **données de Nîmes à 12 h** sont utilisées.

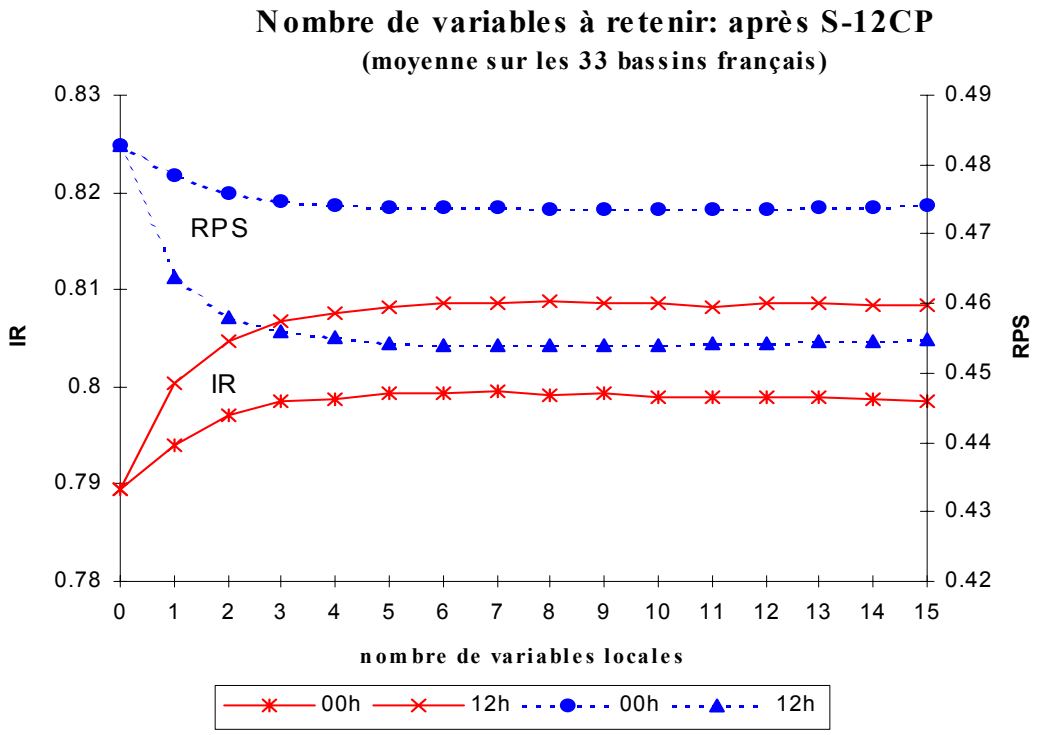


figure V-11 : détermination du palier

V.2.3.c Choix de la méthode d'analogie synoptique

Si nous avons travaillé d'abord sur la méthode S-12CP, pourtant moins performante que la méthode TW-GR, c'est uniquement une question de chronologie. En effet, les premiers essais présentés ici (choix du nombre d'analogues, du palier, de l'échéance) ont été effectués avant d'entamer ceux sur le score de Teweles-Wobus. C'est pourquoi, lorsque la méthode TW-GR s'est avérée plus performante que la méthode S-12CP, nous avons appliqué, sans les remettre en cause, les meilleurs résultats de l'analogie locale obtenus en partant de S-12CP à savoir :

- sélection de **30 analogues**,

- avec une distance euclidienne de la forme (V-4) utilisant **6 variables locales à 12 h** choisies par sélection ascendante.

On trouvera donc en annexe V-10 le même tableau que pour la méthode S-12CP, donnant pour chaque bassin le score obtenu en utilisant 6 variables locales, soit à 00 h, soit à 12 h dans la distance

euclidienne servant de critère d'analogie locale. La synthèse, à savoir la moyenne des 33 bassins, est donnée dans le tableau V-3 ci-dessous:

	prévision pluie/non pluie IR	prévision en classes RPS*100
méthode S-12CP	**78.9**	**48.3**
+ Nîmes 00h	79.9	47.4
+ Nîmes 12h	80.9	45.4
méthode TW-GR	**80.9**	**44.7**
+ Nîmes 00h	82.0	44.0
+ Nîmes 12h	82.6	42.6

tableau V-3: récapitulatif

Des gains de l'ordre de 1 % sont à noter avec l'analogie locale à 00 h et 2 % avec celle à 12 h. De plus, celle-ci n'est pas sensible au choix de la méthode d'analogie synoptique, c'est donc la meilleure, la méthode **TW-GR + Nîmes 12 h** que nous retiendrons.

V.2.3.d Les variables retenues

Sur l'ensemble des 6 variables retenues par bassin, soit 6 x 33 = 198, nous avons examiné celles qui sortaient le plus souvent, par type (T, Td, H ...) et par niveau de géopotentiel (cf. figure V-12). On peut remarquer que:

- quel que soit le type de prévision et quelle que soit la méthode d'analogie synoptique choisie, les répartitions sont équivalentes ce qui donne aux résultats une certaine robustesse,

- les variables locales retenues sont pour la plupart des variables de basses couches (950 à 700 hPa),

- les variables retenues sont en premier lieu celles donnant une idée de l'état hygrométrique de l'atmosphère: H et T-Td représentant environ 35% des variables retenues. Elles sont d'ailleurs bien corrélées (cf. annexe V-11). Viennent ensuite les altitudes des niveaux de pression dans 20-30% des cas.

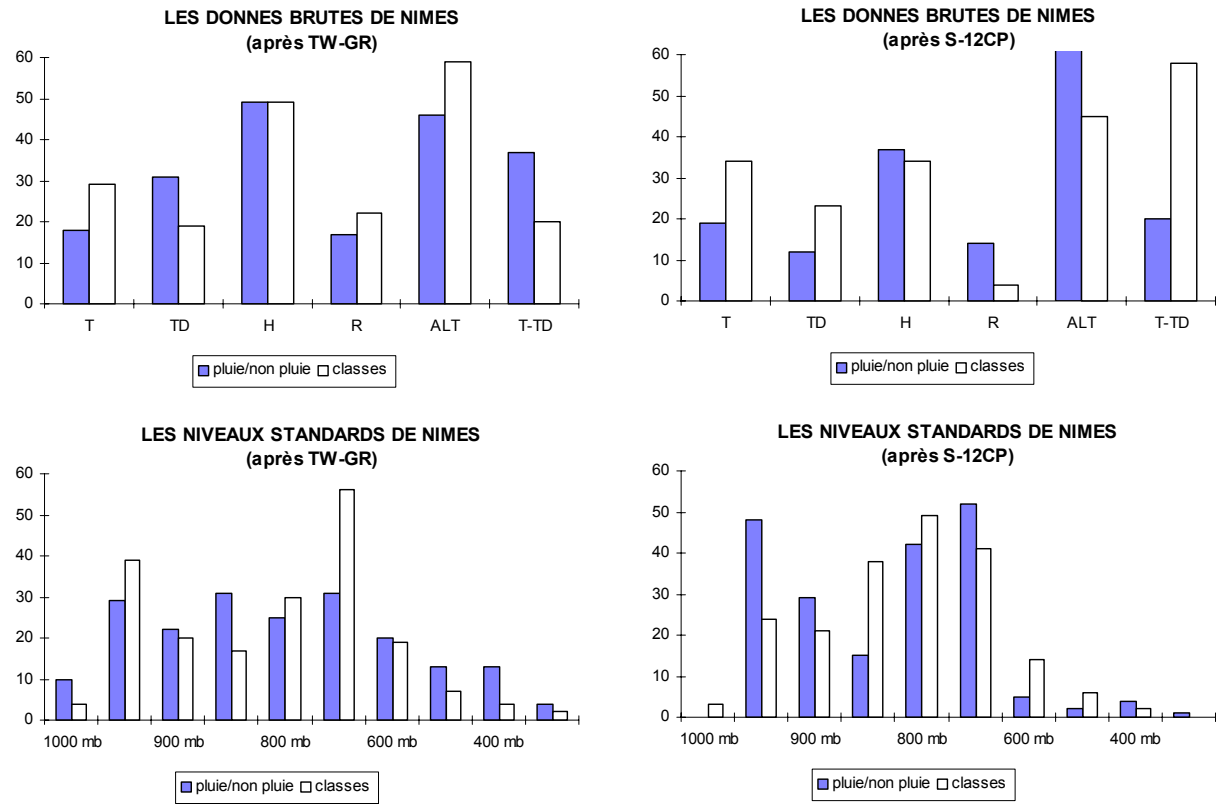


figure V-12 : répartition des variables locales retenues

V.2.3.e Conclusion

Même en utilisant les données brutes de Nîmes à 12 h, les gains sont relativement faibles, de l'ordre de 2 %. Cependant il ne faut pas oublier qu'ils sont moyennés sur les 33 bassins ce qui peut fausser les conclusions. En effet, si l'on dresse la carte de l'intensité des gains bassin par bassin pour la méthode TW-GR + Nîmes 12 h (cf. figure V-13 pour la prévision probabiliste et annexe V-12 pour la prévision pluie / non pluie), seuls les bassins situés dans un rayon, d'une centaine de kms autour de Nîmes sont sensibles à l'apport d'information locale venant de Nîmes. C'est ce que nous avions pressenti au §V.2.2.a.

On peut donc parler de l'existence d'un rayon d'influence de l'information locale, estimé à 150-200 km, voire même un domaine elliptique axé Sud-Ouest/Nord-Est dans le sens des vents dominants.

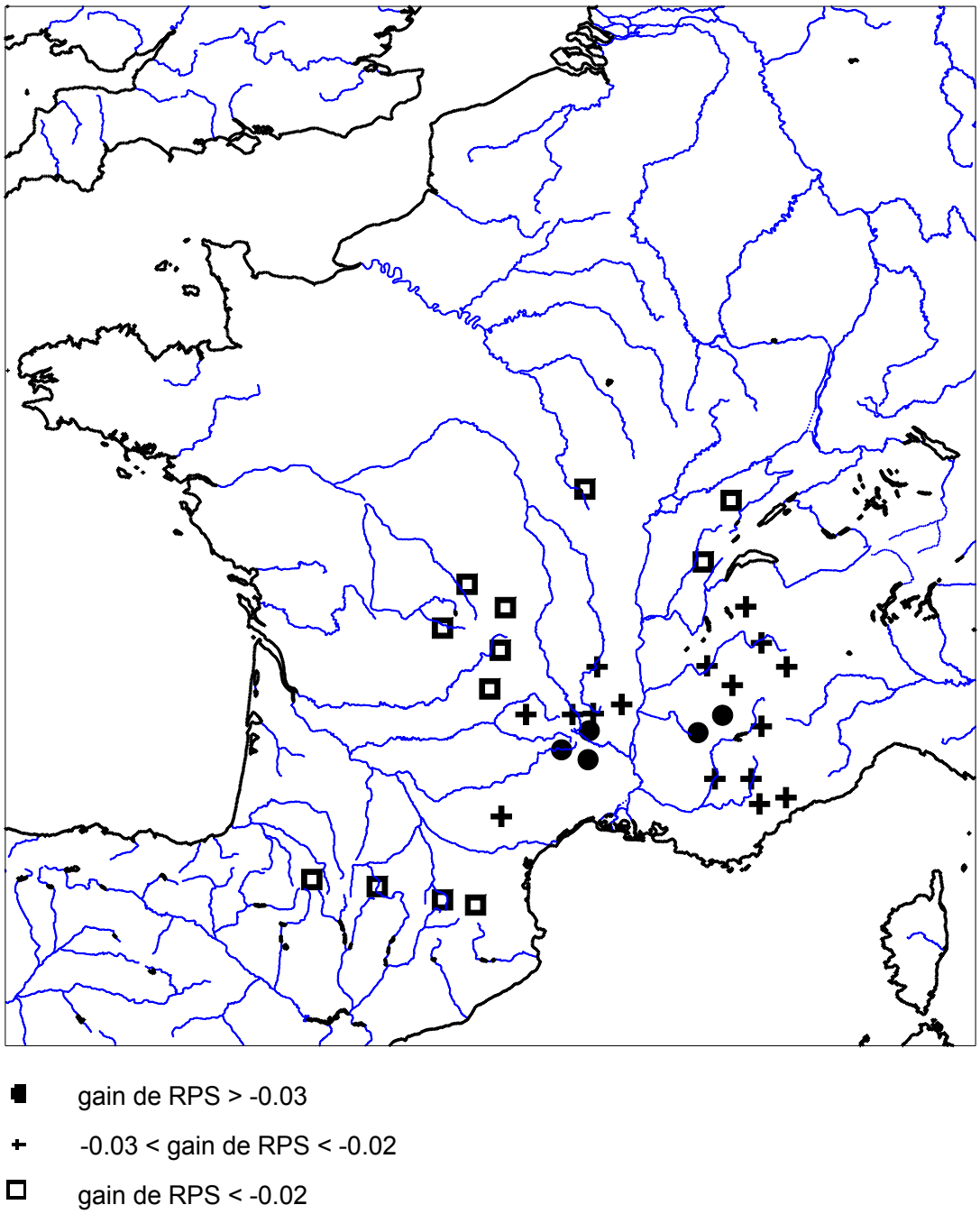


figure V-13 : rayon d'influence des variables locales de Nîmes

V.2.4 Utilisation des différents index **

Cette fois, les variables locales ne sont plus les données brutes mais les 17 index (indices d'instabilité et paramètres) présentés au §V.2.1.b. On leur a appliqué le même traitement qu'aux

données brutes au paragraphe précédent, ceci pour les 9 bassins témoins. Mais les performances finales, obtenues cette fois avec seulement 3 indices sélectionnés (cf. un exemple en annexe V-13), sont nettement inférieures à celles obtenues avec les 6 données brutes, que ce soit en utilisant les indices à 00 h ou à 12 h. D'ailleurs, cela se confirme si l'on regarde les corrélations entre les indices et les pluies d'une part et les données brutes et les pluies d'autre part, qui sont plutôt inférieures dans le premier cas (cf. annexe V-14).

V.2.5 Quelques vérifications

Pour terminer, nous avons voulu vérifier si notre façon d'utiliser l'information locale était la meilleure, même si cela peut sembler peu intéressant à première vue.

Jusque là, une hiérarchie des deux types d'analogie a été utilisée: d'abord l'analogie synoptique puis ensuite celle au niveau local parce que cela semblait le plus logique. Cependant on aurait pu faire l'inverse car comme nous l'avancions au début du paragraphe §V.2, on peut se demander s'il faut d'abord regarder les conditions météorologiques synoptiques puis celles locales (ce que nous avons supposé) ou l'inverse. En effet, faut-il d'abord regarder si la situation locale est fortement instable, puis demander aux conditions synoptiques s'il y a risque d'alimentation entretenue en air humide de basse couche ou procéder à l'inverse ?

Enfin, pourquoi les 2 analogies ne seraient-elles pas à mettre sur le même plan ?

V.2.5.a L'analogie locale seule

Pour évaluer les performances de la seule analogie locale, le principe de la méthode S-12CP (sélection ascendante des CP à introduire dans la distance euclidienne) a été appliqué aux 60 données brutes du radiosondage de Nîmes à 12 h, ceci pour les bassins situés dans le rayon d'influence de Nîmes. Pour les autres, vu que le radiosondage ne leur apporte rien, une analogie basée sur ces seules données ne peut être efficace.

Les bassins ont donc été choisis grâce aux cartes de répartition des gains avec l'introduction de l'information locale. Ce sont les bassins du Sud-Est et des Cévennes, repérés avec des croix et des ronds sur la figure V-13, numérotés 5, 6, 7 et 14 à 31 (pour les numéros, se reporter à la carte du chapitre I, figure I-2).

Pour ces 21 bassins et pour une prévision en pluie / non pluie, la courbe d'évolution du score en fonction du nombre de variables retenues présente (cf. un exemple en annexe V-15) la même forme que pour la méthode S-12CP: lente montée et palier vers 12 variables (voire moins).

Néanmoins les performances atteintes, quelle que soit la forme de la courbe, ne dépassent jamais celles de la meilleure méthode d'analogie synoptique, la méthode TW-GR. Des résultats similaires ont été obtenus en prévision probabiliste en classes, ce qui prouve la nécessité d'une analogie synoptique.

Il s'agit maintenant de déterminer la meilleure façon de combiner ces deux analogies.

V.2.5.b Les 2 analogies, synoptique et locale, combinées

Là encore le principe de la méthode S-12CP a été repris mais cette fois la sélection des variables à introduire dans le critère de sélection se fait entre :

- les 48 CP des champs de géopotentiel (12 CP à 00h, 12 à 24h pour les champs de géopotentiel 700 et 1000 hPa),

- les 60 données brutes du radiosondage de Nîmes à 12h.

En prévision probabiliste par classes, pour les 21 bassins considérés, les résultats sont à peu près équivalents, que l'on utilise l'analogie synoptique suivie de l'analogie locale (méthode S-12CP + Nîmes 12 h) ou l'analogie combinée (cf. annexe V-15). Cependant, dans le deuxième cas, un nombre plus important de variables est requis, de l'ordre de 20, pour atteindre le palier.

En outre, dans un but opérationnel, cette dernière méthode est sans doute moins pratique que celle faisant l'analogie en deux temps. Effectivement, les données utiles de Nîmes sont celles à 12 h alors

que la prévision se fait vers 06 h. Or leurs prévisions, contrairement à celles des champs de géopotentiels, ne sont pas très fiables. On s'orienterait donc vers une relance de la méthode de prévision en milieu de journée après avoir reçu les données locales de Nîmes mesurées à 12 h. Il semble donc plus pragmatique d'utiliser la méthode TW-GR pour faire une première prévision vers 06 h, puis de l'affiner en milieu de journée avec une sélection par analogie locale des analogues du matin.

V.2.5.c L'analogie locale avant l'analogie synoptique

Cette fois, l'analogie se fait d'abord au niveau local, en utilisant les résultats obtenus au paragraphe, puis une 2ème sélection est faite avec la méthode TW-GR :

- 50 analogues sont sélectionnés avec une distance euclidienne utilisant les 3 variables sélectionnées parmi les 60 données brutes du radiosondage de Nîmes (§ V.2.5a),

- sur ces 50, les 30 meilleurs par la méthode TW-GR sont extraits.

Mais pour les 21 bassins soumis à l'influence de l'information locale de Nîmes, les performances sont à peu près équivalentes à celles obtenues avec la seule analogie locale (cf. annexe V-16) : 75.5 % contre 75.8 %

V.3 Conclusion du chapitre V

Les données synoptiques que nous avons pu introduire dans le critère d'analogie, à savoir les champs de température, d'humidité et de géopotentiel à divers niveaux de pression, n'ont pas fourni de résultats concluants mais quelques pistes pour le futur.

En effet, quelques remarques encourageantes sont à faire :

- Tout d'abord, au vu des performances obtenues, il pourrait être intéressant, sinon de **remplacer le champ de géopotentiel 700 hPa**, **au moins de le compléter par celui à 500 hPa.**

- Ensuite, **l'introduction d'une information sur l'état hygrométrique en basses couches (900 hPa)**, en plus de celle contenue indirectement dans les champs de géopotentiel 700 et 1000 hPa, semble nécessaire pour améliorer la prévision même si cela n'a pu être vérifié que sur 6 années de données.

Toutefois, au sujet de ces données synoptiques, nous sommes conscients que de nombreuses autres données sont sans doute très intéressantes pour prévoir la pluie (tourbillon géostrophique ou géopotentiel à des niveaux plus élevés par exemple).

C'est pourquoi, avec la perspective de disposer des fichiers du NCEP/NCAR réanalysés sur 1955-1995, il sera plus aisé de vérifier les remarques faites ci-dessus et de tester un grand nombre de variables sur une période commune avec le fichier historique initial, assez longue (1955-1993).

L'utilisation de données locales comme celles du radiosondage de Nîmes s'est avérée, quant à elle, intéressante :

- sous forme de données brutes à 12 h (température, température du point de rosée, humidité relative, rapport de mélange, altitude à différents niveaux de pression) et non d'indices d'instabilité,

- en sélection de 2ème niveau, après celle sur les données synoptiques (méthode TW-GR ou S-12CP),

- et seulement **dans un rayon d'influence de l'ordre de 150-200 km** autour de Nîmes.

Des perspectives analogues seraient à exploiter autour des radiosondages de Palma de Majorque, Milan, Bordeaux...

CHAPITRE VI :

EXTENSION à des BASSINS FRONTALIERS

et

VALIDATION sur les AUTOMNES 1994,1995 et 1996

Introduction

Ce dernier chapitre se décompose en deux parties relativement distinctes. Dans un premier temps, les meilleurs résultats obtenus dans les chapitres III, IV et V ont été appliqués à une dizaine de bassins appartenant à des régions frontalières de la France (Catalogne pour l'Espagne, Ligurie et Piémont pour l'Italie) et présentant des caractéristiques identiques en terme de système pluvieux (cf. chapitre II, § II.2.3).

Ensuite, une validation de ces meilleurs résultats a été faite sur les automnes 1994, 1995 et 1996, pour tous les bassins où les données étaient disponibles (France et Italie).

VI.1 Extension à des bassins versants frontaliers

Suite à nos échanges scientifiques, nous avons sélectionné des bassins situés sur les frontières Sud (Catalogne-Espagne), et Sud-Est (Ligurie et Piémont, Italie) de la France (cf. chapitre II, §II-2). Ils sont représentés sur la figure VI-1, sur laquelle sont aussi numérotés les groupements français frontaliers.

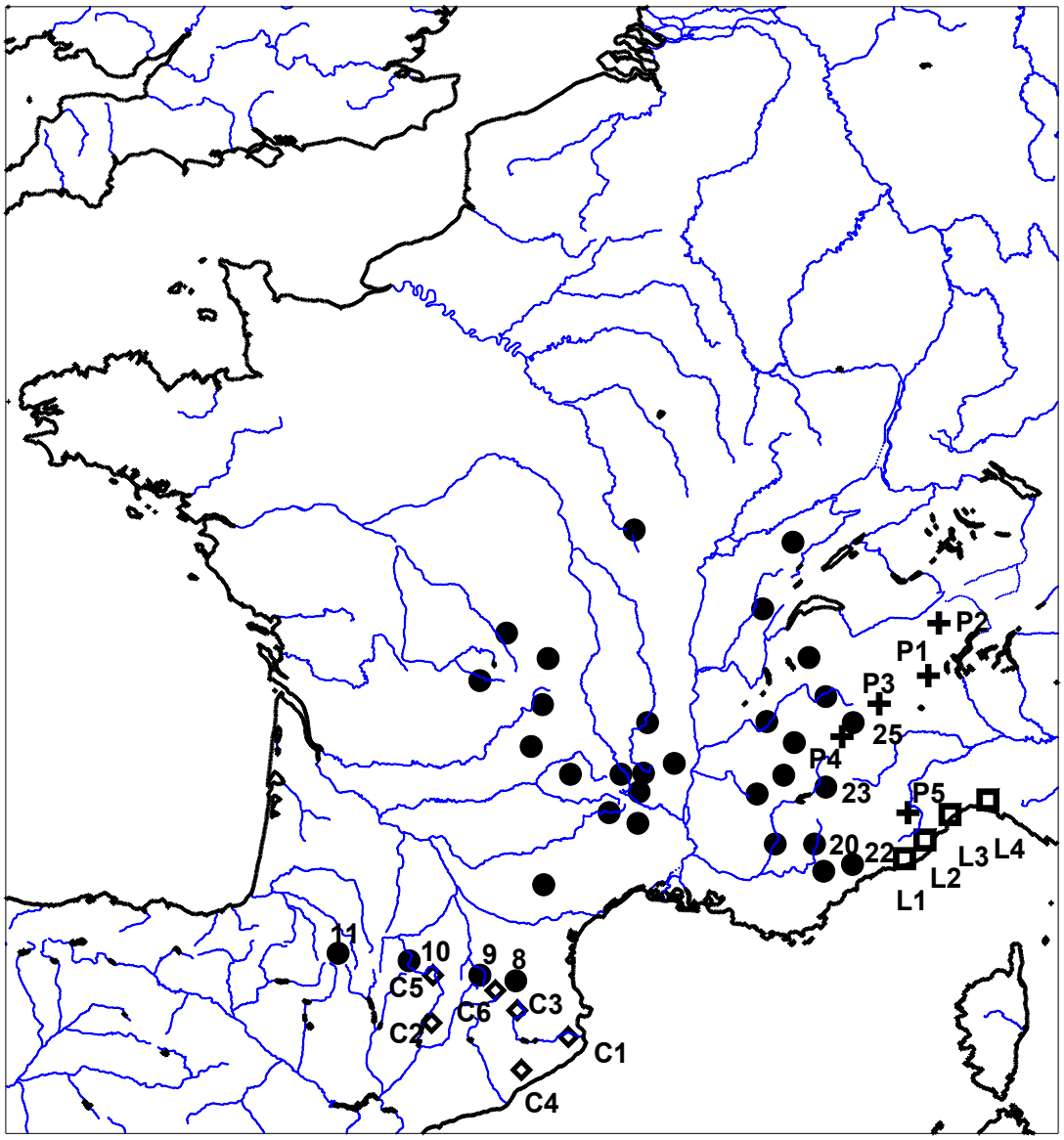
VI.1.1 Les groupements espagnols

Pour l'Espagne, les données journalières de pluie pour 6 bassins versants catalans définis au chapitre 2 (cf. II.2.2c et figure II-14) sont disponibles uniquement sur les automnes de 1970 à 1990. C'est pourquoi, afin de comparer les performances avec celles des bassins français, les méthodes choisies ont été, pour ces derniers, relancées sur cette même période, toujours en validation croisée.

Les bassins français utilisés pour la comparaison sont les 4 bassins des Pyrénées (cf. figure VI-1) :

- Pyrénées Est (8),
- Ariège - Vicdessos (9),

- Pique - Garonne - Salat (10),
- Gaves (11).



Catalogne:
C1. Emporda
C2. Pallars Jussa
C3. Bages
C4. Valles
C5. Vall d'Aran
C6. Ripolles

Ligurie:
L1. Roya-Nervia-Taggia
L2. Centa
L3. Savona
L4. Genova

Piémont:
P1. Sésia
P2. Toce
P3. Orco
P4. Dora Riparia
P5. Tanaro

figure VI-1 : les groupements frontaliers

Les méthodes retenues dans les chapitres précédents sont les suivantes :

i) méthode S-12CP : sélection des analogues avec une distance euclidienne utilisant 12 Composantes Principales CP à 00 ou 24 h (choisies par sélection ascendante),

ii) méthode S-12RS : même méthode que la précédente mais avec 12 données de RadioSondage RS à 00 ou 24 h,

iii) méthode TW-GR : sélection des analogues avec le critère de Teweles-Wobus avec des données à 00 ou 24 h en points de grille (grille GR à définir),

iv) méthode TW-GR + Nîmes : sélection des analogues en 2 temps, d'abord avec la méthode TW-GR puis avec les données de Nîmes à 00 ou 12 h,

v) ainsi que 3 méthodes de références : *la climatologie, la persistance et la méthode initiale dite de référence*.

VI.1.1.a La méthode S-12CP et les méthodes de référence

Des résultats similaires à ceux observés sur les bassins français ont été obtenus, et ce pour les deux types de prévision (pluie / non pluie et prévision en classes de pluie), à savoir (cf. annexe VI-1) :

- palier atteint pour environ 12 CP retenues,
- la climatologie et la méthode de référence ont toujours des performances inférieures,
- pour la persistance, elle est toujours moins bonne en prévision en classes,
- les performances obtenues sont du même ordre que celles des 4 bassins français des Pyrénées pour cette période.

Remarque : En annexe VI-1, seules les courbes d'efficience pour la prévision en pluie / non pluie ont été données car la prévision en classes présente des comportements identiques.

VI.1.1.b La méthode S-12RS

Lorsque l'on utilise les données de RadioSondage RS à la place des CP, là encore un palier est atteint autour de 12 RS sélectionnées. Quant aux performances réalisées en moyenne sur les 6 groupements catalans pour les méthodes S-12CP et S-12RS, si en pluie / non pluie elles sont

équivalentes, en prévision probabiliste un léger gain est à noter avec les RS (cf. tableau VI-1 ci-dessous).

Enfin, si l'on regarde les stations de RadioSondage les plus sélectionnées (cf. annexe VI-2), on remarque, par rapport aux stations sélectionnées pour les bassins français (cf. figure IV-5), un recentrage vers le Sud, donc sur la zone d'étude, très net en particulier pour la prévision en pluie / non pluie.

	prévision pluie/non pluie IR	prévision en classes RPS x 100
méthode S-12CP	75.3	45.7
méthode S-12RS	75.2	44.2
méthode TW-GR3EB1	**75.6**	**43.1**

tableau VI-1 : performances des différentes méthodes pour la Catalogne

VI.1.1.c La méthode TW-GR

Comme pour les bassins français sur la période totale 1953-1993, la localisation optimale de la grille de taille n° 3 (cf. figure IV-6) a été effectuée en maximisant la performance moyenne pour les 6 groupements, et ceci pour les 2 types de prévision. Trois grilles ont donné des résultats intéressants (cf. figure VI-2):

- 3E avec IR = 76.0 % pour la prévision pluie / non pluie,
  RPS = 43.9 pour la prévision probabiliste,
- 3EB1 avec IR = 75.8 % et RPS = 43.5,
- 3EB2 avec IR = 75.6 % et RPS = 43.1,

B1 indiquant que la grille est située un point de grille en-dessous de la grille 3E, B2, 2 points de grille.

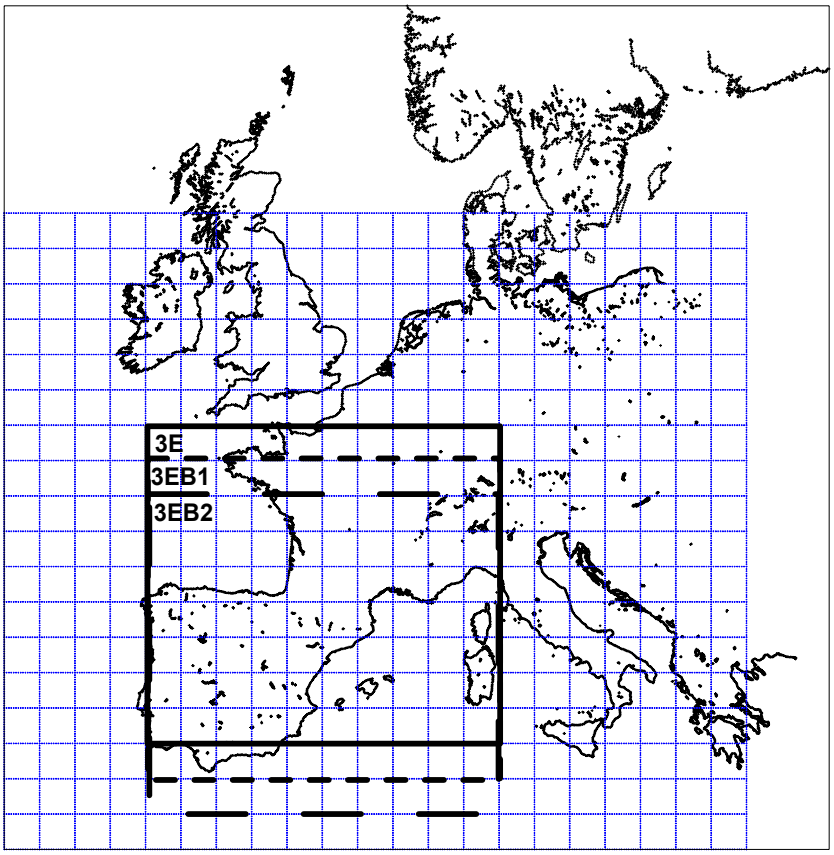


figure VI-2 : optimisation de la grille pour les groupements catalans

Le meilleur compromis nous a semblé être la **grille 3EB1 pour les groupements catalans**, grille située un peu plus au Sud et à l'Ouest de celle retenue pour les bassins français. Et, après optimisation de la grille, comme pour les bassins français, la **méthode TW-GR3EB1** est la plus performante (cf. tableau VI-1).

Remarque : par acquis de conscience, nous avons vérifié qu'en utilisant une grille de taille plus importante (n°4, cf. figure IV-6), cela n'apportait rien quelque soit sa position.

VI.1.1.d Utilisation des données brutes de Nîmes

L'utilisation du critère de sélection de deuxième niveau avec les données brutes de Nîmes entraîne une nouvelle diminution de la période d'étude qui passe de 70-90 à 70-83, soit seulement 14 ans. Aussi, si les résultats principaux restent les mêmes que ceux observés sur les groupements français pour la période 53-93, c'est-à-dire :

- palier à 6 données brutes retenues,
- données à 12 h plus intéressantes qu'à 00 h (cf. tableau VI-2),
- mêmes données retenues,

on ne peut cependant pas comparer les performances obtenues.
On trouvera cependant les performances obtenues lorsque l'on utilise les données brutes de Nîmes après la méthode TW-GR3EB1 dans le tableau VI-2.

période 70-83	prévision pluie/non pluie IR	prévision en classes RPS x 100
méthode TW-GR3EB1	76.1	43.1
TW-GR3EB1 + Nîmes 00h	78.5	42.0
TW-GR3EB1 + Nîmes 12h	**78.7**	**41.7**

tableau VI-2: performances des méthodes utilisant les données de Nîmes pour la Catalogne

Si les résultats sont intéressants avec les données de Nîmes à 12 h, on peut cependant penser que l'utilisation d'un radiosondage plus proche de la Catalogne donnerait de meilleurs résultats. On pense en particulier à Palma de Majorque, mieux situé par rapport à cette région, pour voir arriver en particulier les flux de Sud.

VI.1.2 Les bassins italiens

Contrairement aux groupements français et espagnols, les groupements italiens correspondent en fait à des bassins hydrologiques parfois importants : de 300 km^2 environ pour la Ligurie à 1000 km^2 pour le Piémont. Ce ne sont donc pas à proprement parler des groupements, homogènes au niveau de la pluviométrie, mais bel et bien des bassins versants. Aussi, même si ces bassins sont relativement petits, en particulier pour la Ligurie, ils sont moins optimaux que des groupements constitués de stations à pluviométrie similaire.

En outre, comme pour la Catalogne, se pose le problème des périodes disponibles puisque les données du Piémont n'ont pu être collectées que sur la période 1953-1986 et celles de la Ligurie sur 1953-1992. Notre traitement a donc surtout valeur de test et de démonstration pour aller éventuellement au-delà.

VI.1.2.a La méthode S-12CP et les méthodes de référence

Comme pour les groupements catalans, 4 bassins français ont été choisis pour servir de comparaison (cf. figure VI-1):

- Verdon (20),
- Var-Tinee-Roya (22),
- Haute Durance (23),
- Mont Cenis (25).

Si des commentaires identiques à ceux effectués au §VI.1.1.a sur les groupements catalans peuvent être faits pour le Piémont, il faut noter les performances relativement médiocres des bassins de Ligurie par rapport aux bassins français (cf. annexe VI-3), sans doute dues à ce découpage en bassins versants et non en groupements pluviométriquement homogènes.

Remarque : Pour la comparaison des 4 bassins français avec les bassins du Piémont, la méthode S-12CP a été relancé sur la période 53-86. Par contre, pour celle avec les bassins de Ligurie, connus sur la période 53-92, les résultats finaux des groupements français sur 53-93 ont été retenus.

VI.1.2.b Autres méthodes

Là encore, le gain obtenu en utilisant les RS plutôt que les CP est un peu plus important en prévision en classes (cf. tableau VI-3). De même, les RS sélectionnées le plus souvent se trouvent cette fois concentrées plus à l'Est et au Sud-Est (cf. annexe VI-4).

Enfin, pour la méthode TW-GR utilisant le score de Teweles-Wobus, le choix de la position de la grille diffère suivant les régions :

- pour le Piémont, on peut hésiter entre 3 grilles (cf. figure VI-3):

  3I avec IR = 78.4 % et RPS = 43.7 (* 100),

  3IH1 avec IR = 78.6 % et RPS = 44.0 (H1 : un point au-dessus),

  3JH1 avec IR = 78.5 % et RPS = 43.9,

mais le compromis entre les 2 prévisions nous incite à prendre la **grille 3JH1,** située au Nord-Est et englobant mieux la région du Piémont.

- pour la Ligurie, la grille la plus intéressante est la **grille 3J**, située un peu plus au Sud que la précédente, comme les bassins de Ligurie par rapport à ceux du Piémont.

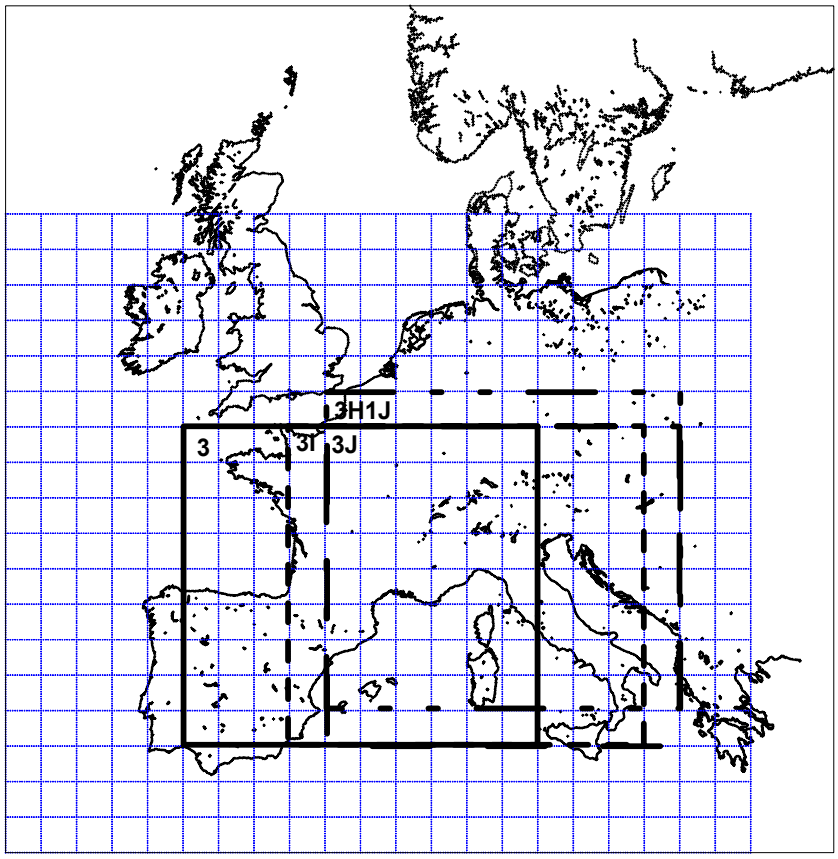


figure VI-3 : optimisation de la grille pour les groupements italiens

	prévision pluie/non pluie IR	prévision en classes RPS x 100	
méthode S-12CP	73.8	52.7	LIGURIE
méthode S-12RS	74.0	49.4	
méthode TW-GR3J	**74.0**	**49.7**	
méthode S-12CP	77.8	48.2	PIEMONT
méthode S-12RS	78.0	46.7	
méthode TW-GR3JH1	**78.5**	**43.9**	

tableau VI-3: performances des différentes méthodes pour l'Italie

VI.1.2.c Utilisation des données brutes de Nîmes

Cette fois, la validation croisée peut se faire sur la même période que pour les bassins français, 1954-1983. Et, comme pour les bassins espagnols (cf §VI.1.1.d), des comportements similaires à ceux des bassins français ont été observés (cf. tableau VI-4).

période 54-83	prévision pluie/non pluie IR	prévision en classes RPS x 100	
méthode TW-GR3J	73.9	50.8	LIGURIE
TW-GR3J + Nîmes 00h	74.5	50.6	
TW-GR3J + Nîmes 12h	**75.1**	**49.8**	
méthode TW-GR3JH1	78.8	44.1	PIEMONT
TW-GR3JH1 + Nîmes 00h	79.6	43.8	
TW-GR3JH1 + Nîmes 00h	**79.9**	**42.8**	

tableau VI-4: performances des méthodes utilisant les données de Nîmes

VI.1.3 Conclusion

Finalement des résultats similaires à ceux des groupements français ont été obtenus avec ces bassins frontaliers (cf. §IV.5). En effet :

i) l'utilisation des Composantes Principales n'apparaît donc plus indispensable puisque les performances de la méthode S-12RS sont plutôt légèrement supérieures à celles de la méthode S-12CP,

ii) l'utilisation des données en points de grille avec comme critère d'analogie le score de Teweles-Wobus apporte un gain non négligeable (méthode TW-GR) après optimisation de la grille. Celle-ci est d'ailleurs recentrée sur la zone d'étude, suivant les bassins concernés (cf. figure VI-4) puisqu'on remarque un décalage vers :

* le Nord-Est pour le Piémont,
* l'Est pour la Ligurie,
* le Sud-Ouest pour la Catalogne.

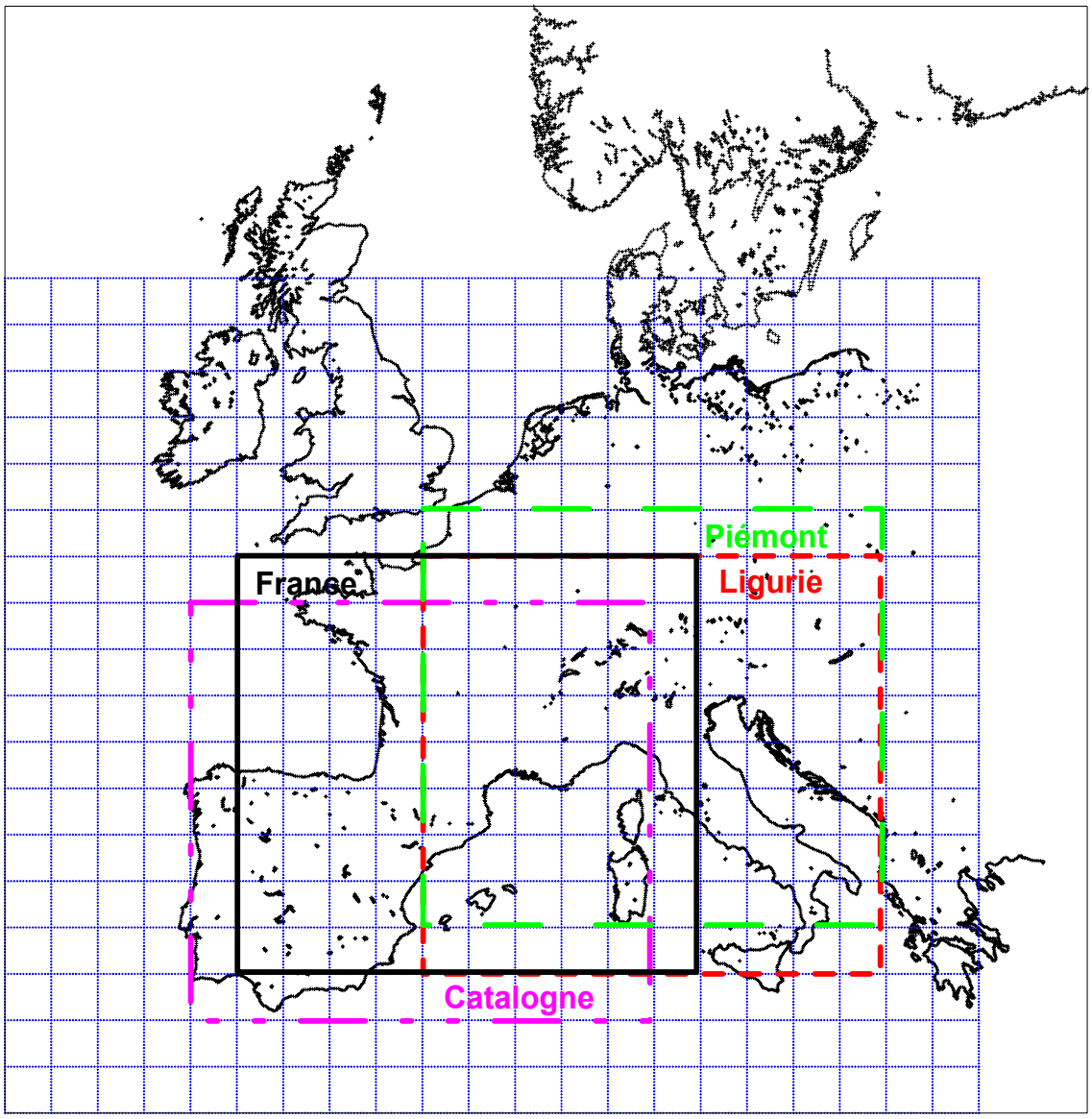


figure VI-4 : les grilles optimales pour les différentes régions

iii) l'utilisation des données thermodynamiques du radiosondage de Nîmes (données brutes) à 12 h en deuxième niveau de sélection est relativement intéressante pour les bassin italiens. Sur la figure VI-4, on trouvera la carte des gains du chapitre V (figure V-21) pour la prévision probabiliste, à laquelle ont été adjoints les gains des bassins italiens. La même carte pour la prévision en pluie / non pluie est donnée en annexe VI-5.

Remarque : Les gains des bassins catalans n'ont pas été donnés car ils correspondent à une période de 14 ans (70-83), trop différente de celle des autres (54-83).

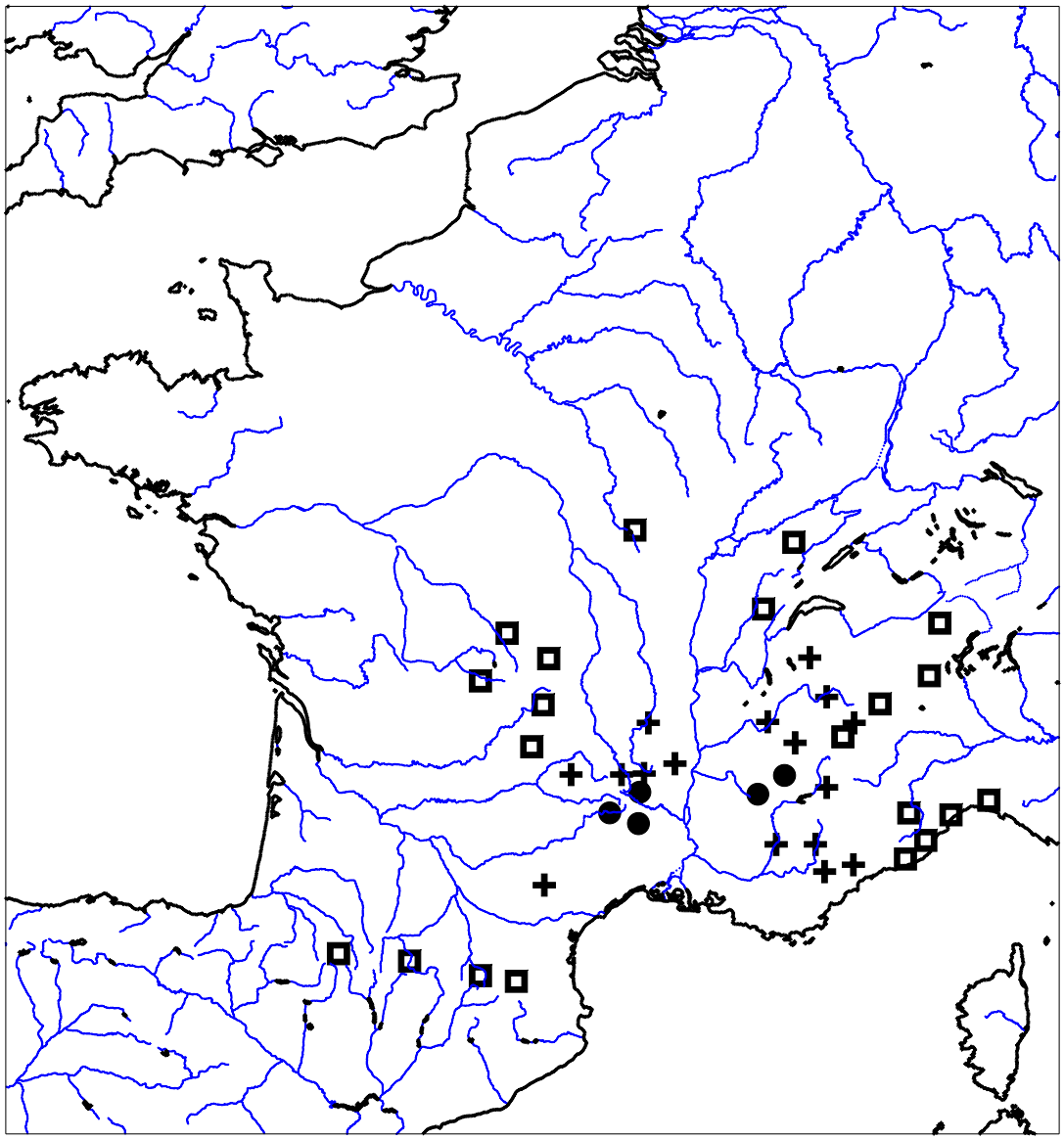


figure VI-5 : rayon d'influence des variables locales de Nîmes

A la lueur de ces résultats, il pourrait être envisager d'implanter la méthode TW-GR au moins dans les deux régions italiennes. Pour cela, une collecte plus fournie des données de pluie doit être effectuée, afin de réaliser de réels groupements à pluviométrie homogène sur la période la plus longue possible.

Pour les bassins espagnols, un gros travail de collecte de données est nécessaire avant toute chose.

VI.2 Validation sur les automnes 1994, 1995 et 1996

Quand il s'agit de performances, plusieurs problèmes se posent. Il faut d'abord donner une évaluation moyenne des performances sur une longue période (validation croisée sur 1953-1993) mais également une idée « réaliste » de la réaction en temps réel de la méthode, illustrée en particulier par quelques événements types (automnes 1994, 1995 et 1996).
Rappelons aussi que nous avons travaillé, jusque là, en différé, donc *en prévision parfaite* pour les champs prédicteurs utilisés à l'échéance de 24 h alors qu'en opérationnel, il s'agira de *champs prévus*.

Aussi, dans cette deuxième partie de chapitre, nous nous proposons de valider les résultats obtenus avec les meilleures méthodes retenues dans les chapitre III, IV et V sur les 3 derniers automnes.
Deux types de validation ont été réalisés :

i) Une validation quantitative avec évaluation des performances, grâce aux scores IR (indice de réussite en prévision en pluie / non pluie) et RPS, le Ranked Probability Score pour la prévision probabiliste.
Cette évaluation a été effectuée sur les automnes 1995 et 1996 pour :

- d'une part, comparer les performances obtenues pour la prévision à 24 h lorsque l'on utilise les champs prédicteurs à 24 h en prévision parfaite (ce qui a été fait en validation croisée) ou en prévision opérationnelle,

- d'autre part, apprécier la perte de performance de la prévision pour des échéances supérieures (48, 72 et 96 h) quand on utilise à 24, 48, 72 et 96 h des champs prévus.

ii) Une validation qualitative par l'intermédiaire de graphes présentant prévision et observation pour les principaux épisodes pluvieux des automnes 1994, 1995 et 1996.

Mais avant de pouvoir faire ces validations, il a fallu récupérer les données de géopotentiel - observées et prévues - et de pluie pour ces trois automnes.

VI.2.1 Les données disponibles

VI.2.1.a Les géopotentiels 700 et 1000 hPa

Pour ce qui concerne les données des 37 radiosondages à 00 h, elles ont pu être récupérées pour les 3 automnes, sans aucune journée manquante. En outre, pour les automnes 95 et 96, leurs prévisions à 24 h ont aussi été archivées. Néanmoins, pour ces dernières, quelques journées manquantes sont quand même à déplorer : une en 1996 et 22 en 1995, en particulier du 1er au 18 novembre, à cause d'une grève à Météo-France.
Cependant, cela va nous permettre de mesurer les pertes de performances dues à l'utilisation des champs prévus à 24 h - et non plus les données à 00 h du jour suivant - dans les méthodes utilisant des champs à 00 et 24 h.

Ensuite, les Composantes Principales issues de l'Analyse en Composantes Principales (ACP) classique et de l'ACP de Processus (ACPP) et les valeurs aux points de la grille n°3 ont été calculées pour les champs observés et prévus. A ce propos, il nous a paru intéressant de noter les corrélations obtenues entre les champs prévus et observés, quelle que soit leur forme (RadioSondage RS, points de GRille GR, Composantes Principales CP).
Elles sont consignées dans le tableau VI-5 pour chaque champ avec :

- *pour les RS*, le coefficient de corrélation moyen sur l'ensemble des 37 stations de RadioSondage,

- *pour les CP de l'ACPP,* le coefficent de corrélation moyen sur les 12 premières CP utilisées dans la méthode S-12CP,

- *pour les CP de l'ACP,* le coefficient de corrélation moyen sur les 6 premières CP, utilisées dans la méthode de référence (celui pour la première CP de l'épaisseur de la couche 700/1000 hPa est de 0.99),

- *pour les GR :* le coefficent de corrélation moyen sur les 110 points de la grille n°3 de la méthode TW-GR.

	champ 700 hPa	champ 1000 hPa
RadioSondages RS	0.97	0.96
CP par ACP (6)	0.98	0.96
CP par ACPP (6)	0.97	0.955
CP par ACPP (12)	0.93	0.905
grille GR n°3	0.97	0.95

tableau VI-5 : corrélation observation / prévision à 24 h

Les champs prévus à 24 h sont donc dans l'ensemble très bien corrélés avec ceux observés. La moins bonne corrélation correspond aux 12 CP calculées par ACPP. Cependant on peut vérifier que cela n'est pas dû à l'ACPP elle-même mais seulement au fait de prendre un plus grand nombre de variables. En effet, les coefficients de corrélation sont à peu près identiques si l'on retient 6 CP calculées par ACP ou ACPP.

Les fortes valeurs de ces coefficients de corrélation nous laissent espérer une diminution relativement faible des performances en passant des champs prédicteurs à 24 h analysés (prévision parfaite) à des champs prévus.

VI.2.1.b Les données de pluie

Pour les pluies, tous les groupements français ont pu être reconstitués pour les 3 automnes sauf :

- le groupement 6 (Haut Tarn - Haut Lot) seulement sur 1994 et 1995,
- le groupement 17 (Romanche - Arc inférieur) sur 1994 et 1996,
- les groupements 29 (Gard - Cèze) et 33 (Cure) pour lesquels aucune donnée n'était disponible (stations abandonnées).

Enfin, quelques données pluviométriques sur la région Piémont pour l'automne 94 permettront une première validation des différentes méthodes, en particulier pour l'épisode particulièrement violent de début novembre qui a affecté la région.

VI.2.2 Validation quantitative de la prévision

Trois méthodes ont été retenues pour être validées :

- *la méthode de référence* qui utilise les seules CP (calculées par ACP classique) à 00 h,
- *la méthode S-12CP*, avec les CP à 00 et 24 h,
- *la méthode TW-GR*, avec les champs en points de grille à 00 h et 24 h.

Elles présentent l'avantage de bien marquer l'évolution des résultats obtenus. En effet, à partir de la méthode de référence, la sélection des CP à introduire dans le critère de sélection des analogues a permis d'aboutir à une méthode S-12CP, plus performante. Par la suite, avec des données en points de grille et le score de Teweles-Wobus pour extraire les analogues, la méthode TW-GR s'est avérée encore plus intéressante.

Remarque : On aurait aussi pu considérer la méthode S-12RS, mais celle-ci ayant des performances quasiment équivalentes à S-12CP, il ne nous a pas semblé utile d'en faire aussi la validation. Et si la méthode S-12CP a été choisie, plutôt que S-12RS, c'est uniquement une question de chronologie dans le travail : la validation de S-12CP a été faire avant de bâtir la méthode S-12RS.

VI.2.2.a Prévision de la pluie à 24 h

Pour la prévision à 24 h faite le jour courant C à 00 h, les données nécessaires sont donc celles à 00 h le jour C et celles à 24 h, soit à 00 h le jour C+1.

Or, pour le calage des méthodes sur la période 1953-1993 (sélection des 12 CP pour la méthode S-12CP, et optimisation de la grille pour TW-GR), les données de la journée courante C à 24 h utilisées étaient celles observées le jour C+1 à 00 h, données connues ; c'était donc de la *prévision parfaite*.

Il faut donc s'attendre, en opérationnel, à une diminution des performances puisque, en faisant la prévision le jour C vers 06 h, les données du jour C à 00 h sont des données observées mais celles du lendemain à 00 h sont des données prévues.

Cette vérification a pu être faite sur les automnes 1995 et 1996 pour lesquels les prévisions des géopotentiels à 24 h ont été archivées. Ainsi, les scores de réussite en prévision pluie / non pluie et en prévision probabiliste ont été calculés sur cette période pour les cas suivants :

i) méthode de référence (REF) avec les données observées à 00 h, nous servant toujours de référence,

ii) méthode S-12CP avec les données observées à 00 h et *en prévision parfaite* à 24 h,

iii) méthode S-12CP (prév) avec les données observées à 00 h et *prévues* à 24 h,

iv) méthode TW-GR avec les données observées à 00 h et *en prévision parfaite* à 24 h,

v) méthode TW-GR (prév) avec les données observées à 00 h et *prévues* à 24 h.

Les résultats, moyennés sur les 31 groupements disponibles, sont consignés dans le tableau VI-6 suivant :

	prévision pluie/non pluie IR	prévision en classes RPS x 100
méthode de référence	69.4	66.4
méthode S-12CP	72.6	56.4
méthode S-12CP (prév)	72.3	56.6
méthodeTW-GR	77.6	48.7
méthode TW-GR (prév)	**77.4**	**49.4**

tableau VI-6 : comparaison des performances prévision / observation

On remarque une très légère perte de performance due au passage aux données prévues pour les méthodes S-12CP et TW-GR. C'est ce que nous pressentions dans le paragraphe précédent, à la lecture des coefficients de corrélation entre données prévues et observées.
De plus, la hiérarchie des méthodes observée sur la période 1953-1993 est respectée: la méthode TW-GR reste nettement supérieure à la méthode S-12CP qui, elle-même, apporte un gain significatif par rapport à la méthode de référence.

VI.2.2.b Estimation de la prévision des champs à 48, 72 et 96 h

Dans un deuxième temps, nous avons voulu voir comment se dégradait la prévision au fur et à mesure que l'échéance augmentait (24, 48, 72 et 96 h). Pour cela, les performances des 3 méthodes pour des prévisions à 24, 48, 72 et 96 h (respectivement notées C, C+1, C+2, C+3) ont été

comparées. Mais auparavant, il a été nécessaire de faire une estimation des prévisions à 48, 72 et 96 h des champs de géopotentiels 700 et 1000 hPa qui seront utilisées dans les différentes méthodes (cf. tableau VI-7) et qui n'ont pas été archivées.

		données				
méthode	échéance	00 h	24 h	48 h	72 h	96 h
REF	C	x				
	C+1		x			
	C+2			x		
	C+3				x	
S1-12CP	C	x	x			
et	C+1		x	x		
TW-GR	C+2			x	x	
	C+3				x	x

tableau VI-7 : prévisions nécessaires pour les différentes échéances

Les données qu'il nous faut sont donc :

- les données observées à 00 h (C),
- les données prévues à 24 (C+1), 48 (C+2), 72 (C+3) et 96 h (C+4).

Or, si les deux premières sont connues, il a fallu faire une estimation des prévisions à 48, 72 et 96 h puisqu'elles n'ont pas été archivées.

Pour ne pas s'égarer dans les indices, référons-nous à la figure VI-6 ci-dessous :

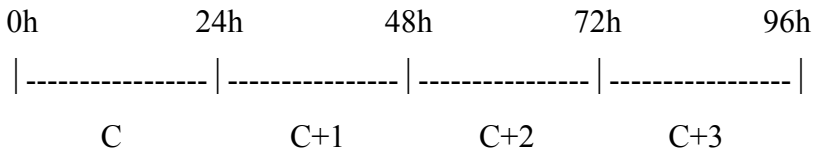


figure VI-6 : les échéances

Plaçons-nous au jour C. Et notons, pour la variable V, représentant un champ prédicteur :

- V_o (C) l'observation faite le jour C,
- V_p (C+1) à Cla prévision à C+1 (24 h) faite le jour C,
- V_p (C+2) à Cla prévision à C+2 (48 h) faite au jour C, etc...

Une estimation de la prévision à 48 h, V_p (C+2) à C, pourrait être faite par la prévision à 24 h faite au jour C+1 (V_p (C+2) à C+1), celle de la prévision à 72 h par la prévision à 24 h faite au jour C+2 ... :

$$V_p \text{ (C+2) à C} = V_p \text{ (C+2) à C+1}$$
$$V_p \text{ (C+3) à C} = V_p \text{ (C+3) à C+2} \qquad \text{(VI-1)}$$

Mais cela nous a semblé être trop optimiste aussi une erreur a été rajoutée. C'est l'écart entre la prévision de C+1 (prévision à 24 h faite à C) et l'observation faite à C+1 qui permet d'intégrer *l'erreur déjà faite le jour C sur la prévision à C+1* (cf. figure VI-7) :

$$V_p \text{ (C+2) à C} = V_p \text{ (C+2) à C+1} + V_p \text{ (C+1) à C} - V_o \text{ (C+1)} \qquad \text{(VI-2)}$$

Avec le même raisonnement, les prévisions à 72 et 96 h sont estimées de la façon suivante :

$$V_p \text{ (C+3) à C} = V_p \text{ (C+3) à C+2} + V_p \text{ (C+2) à C+1} - V_o \text{ (C+2)} + V_p \text{ (C+1) à C} - V_o \text{ (C+1)} \qquad \text{(VI-3)}$$

$$V_p \text{ (C+4) à C} = V_p \text{ (C+4) à C+3} + V_p \text{ (C+3) à C+2} - V_o \text{ (C+3)} + V_p \text{ (C+2) à C+1} - V_o \text{ (C+2)} + V_p \text{ (C+1) à C} - V_o \text{ (C+1)} \qquad \text{(VI-4)}$$

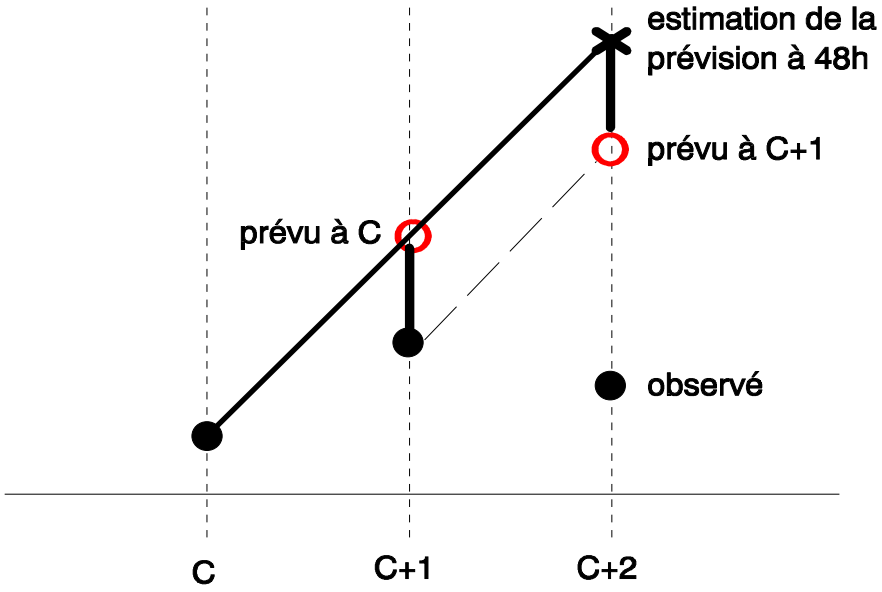


figure VI-7 : estimation de la prévision à 48 h

VI.2.2.c Prévision de la pluie à des échéances supérieures

Lorsque nous avons été en possession des prévisions des champs à 24, 48, 72 et 96 h, les performances des 3 méthodes ont pu être calculées. La figure VI-8 présente, pour les 2 types de prévision, l'évolution des performances au fur et à mesure de l'augmentation de l'échéance.

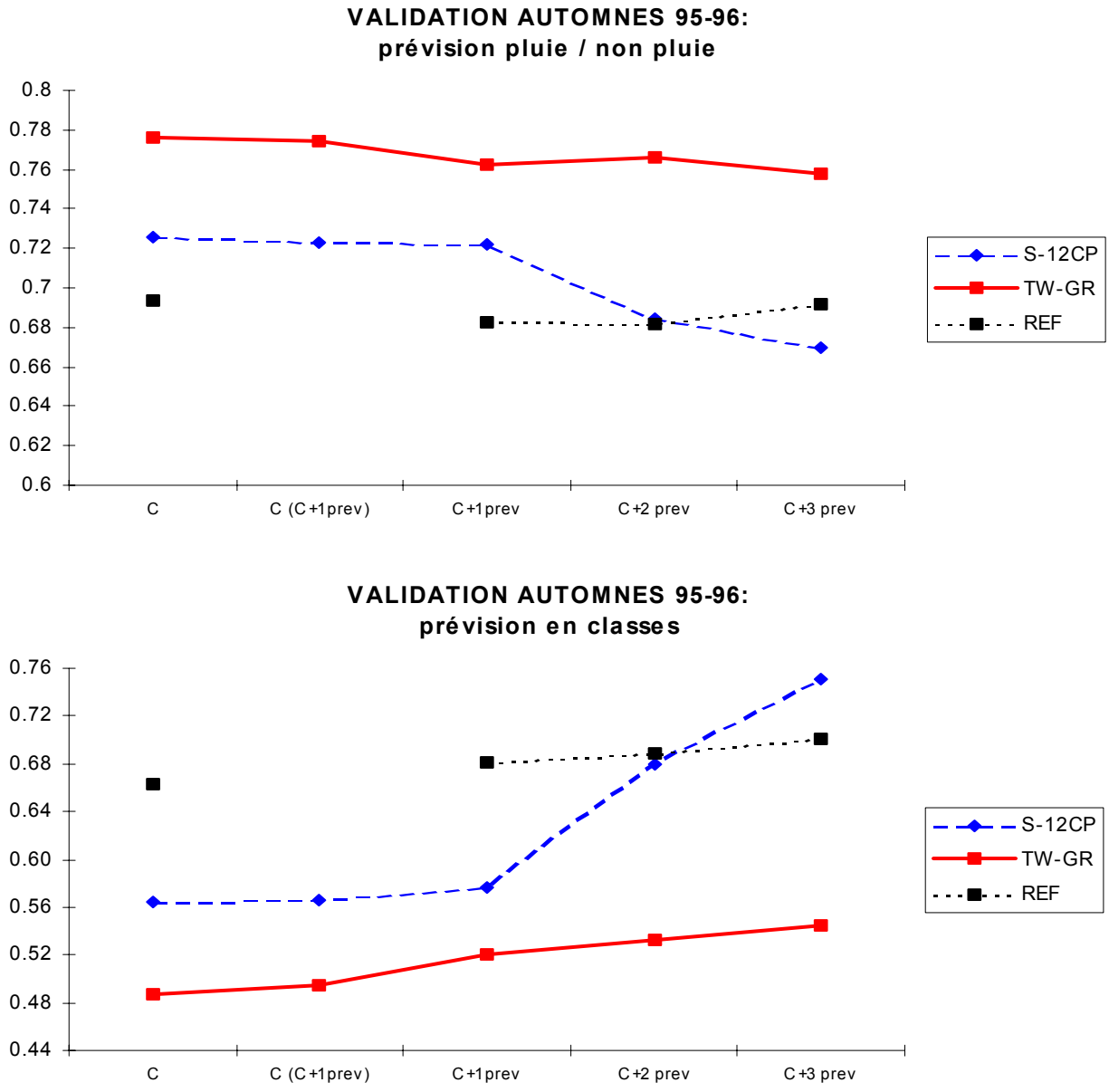


figure VI-8 : évolution des performances avec l'échéance

Cela nous amène à quelques remarques :

- dans l'ensemble, **quand l'échéance augmente, les performances diminuent**,
- cependant **jusqu'à 48 h, les pertes restent minimes** pour les 3 méthodes,

- **au-delà (72 et 96 h),** si les méthodes TW-GR et de référence perdent peu en efficacité, **la méthode S-12CP se dégrade** considérablement pour être même dépassée par la méthode de référence à 96 h,

- enfin, quelque soit l'échéance, la **méthode TW-GR est toujours la meilleure**.

Donc, là encore, la méthode TW-GR est bien la plus intéressante, avec une dégradation limitée de ses performances quand l'échéance augmente. Seule la méthode de référence, tout en restant nettement en-deçà en terme de performance, fait aussi bien voire mieux. Cela peut s'expliquer par le fait qu'elle n'utilise les champs prédicteurs qu'à une échéance (00 h pour la prévision à 24 h, 24 h pour celle à 48 h, etc...), ce qui diminue l'erreur due à l'estimation de la prévision de ces champs par rapport aux deux autres méthodes qui les utilisent à deux échéances (00 et 24 h pour la prévision à 24 h, etc...).

Quant à l'effondrement de la méthode S-12CP pour des prévisions supérieures à 48 h, on peut l'attribuer au fait que les 12 CP de l'ACPP prévues à 24 h sont de qualité un peu moins bonne que les autres champs prédicteurs. En effet, si l'on se rappelle des coefficients de corrélation entre les champs prévus et observés (cf. tableau VI-5), ils étaient relativement plus faibles pour les 12 CP par ACPP que pour les autres.

De plus, la dégradation est encore amplifiée car cela se répercute dans l'estimation des CP prévues à 72 et 96 h qui utilise (cf. éq. VI-3 et VI-4) le cumul des prévisions à 24 h : V_p(C+4) à C+3, V_p(C+3) à C+2, ...

VI.2.3 Validation qualitative de la prévision

Il s'agit maintenant d'apprécier les qualités de la prévision par rapport à l'observation sur quelques épisodes pluvieux, à l'aide de graphes. En effet, pour un prévisionniste, plus que des indices statistiques, c'est la façon dont réagit une méthode de prévision à un événement qui est essentielle.

Bien évidemment, il n'était pas envisageable, contrairement à la validation quantitative de la prévision faite au paragraphe précédent, de travailler sur tous les bassins disponibles. Aussi, nous sommes-nous focalisés sur des bassins présentant des pluies à caractère extrême durant ces 3 automnes, et en particulier, un bassin caractéristique des Cévennes, **la Loire supérieure**. En effet,

un certain nombre de violents épisodes (retours d'Est et coups de Sud) ont affecté ce groupement durant les 3 derniers automnes comme le montrent les histogrammes de la figure VI-9 :

* 23 septembre 1994 100 mm,
* 19 octobre 1994 150 mm,
* 4 novembre 1994 130 mm,
* 18 septembre 1995 110 mm,
* 4 octobre 1995 110 mm,
* 11 et 12 novembre 1996 140 et 190 mm.

Enfin, quelques graphes seront présentés pour un bassin du Piémont, surtout pour l'épisode particulièrement violent du début novembre 1994.

VI.2.3.a Représentation de la prévision

Des graphes superposant prévision et observation seront présentés toujours pour les 3 mêmes méthodes que nous rappelons ici :

i) méthode de référence,

ii) méthode S-12CP,

iii) méthode TW-GR.

Un problème s'est ensuite posé quant à la représentation graphique de la prévision. En effet, jusqu'à maintenant, nous avons travaillé en prévision pluie / non pluie ou en prévision probabiliste. La première n'étant pas assez informative et la seconde trop complexe à représenter, nous avons finalement choisi de revenir à la représentation utilisée dans la méthode de référence, à savoir :

- les quantiles 20, 60 et 90%,
- et la moyenne des analogues,

calculés à partir de la distribution empirique des précipitations observées lors des situations retenues comme analogues. Le seul changement intervenant dans les différentes méthodes se situe donc dans le calcul de la prévision, après la sélection des analogues.

Pour cela, comme la prévision quantilique est plus proche d'une prévision probabiliste que d'une prévision d'occurrence de pluie, les méthodes retenues S-12CP et TW-GR sont celles qui ont été calées avec la prévision probabiliste. Ainsi, les 12 CP retenues dans la méthode S-12CP sont celles sélectionnées en minimisant le Ranked Probability Score de la prévision probabiliste.

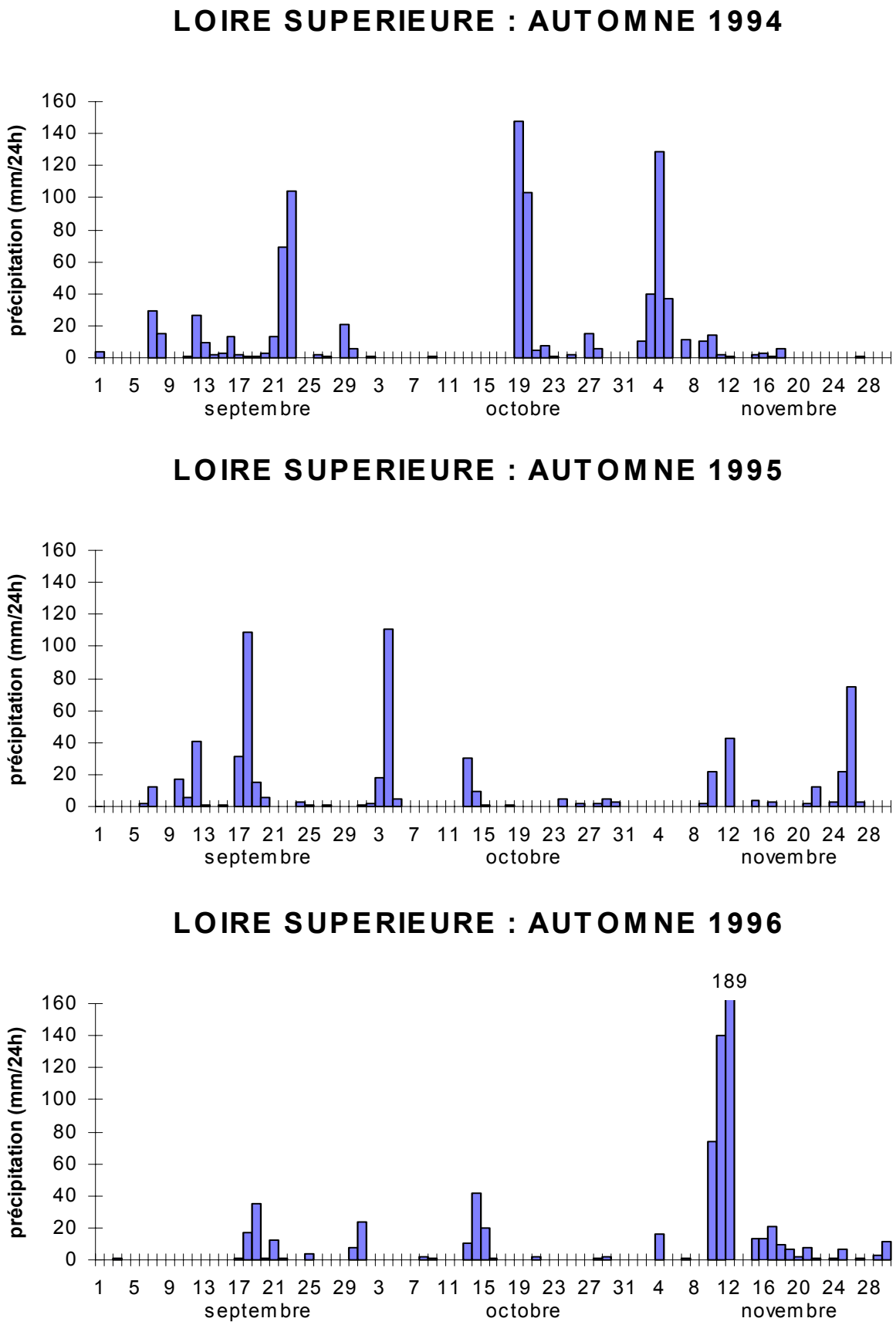


figure VI-9 : hyétogrammes des automnes 1994, 1995 et 1996 pour la Loire supérieure

VI.2.3.b Prévision à 24 h

Les performances des 3 méthodes ont été comparées en utilisant, pour les données à 24 h dans les méthodes S-12CP et TW-GR, des prévisions parfaites. L'automne 1994 nous a semblé intéressant pour cette étude, et en particulier les mois d'octobre 1994, avec un pic à 150 mm le 19 octobre (cf. figure VI-10), et de novembre 1994 avec 130 mm le 4 (cf. figure VI-11).

Remarque : Avant toute chose, les utilisateurs quotidiens de la méthode de référence nous avait prévenu d'un manque de réactivité certain du quantile 20%. Celui-ci, presque toujours à zéro n'apporte aucune information, et ce, quelque soit l'épisode. Nous avons pu le vérifier sur les quelques événements choisis aussi n'en parlerons-nous pas.
De plus, les utilisateurs se basent, pour déterminer un état d'alerte, d'abord sur le quantile 90%, plus réactif, et ensuite sur la moyenne des analogues. Aussi nous sommes-nous focalisés sur ceux-ci.

Octobre 1994 :

A propos de la prévision par la méthode de référence pour le mois d'octobre 1994 (figure VI-10a), on peut noter :

- la trop grande réactivité du quantile 90% pendant la période non pluvieuse du 1er au 18 octobre,
- le défaut d'alerte pour le 19 puisque seul le pic du lendemain a été perçu, un peu en-deçà de ce qui a été observé (quantile 90% à 80 mm pour 100 mm observé),
- la surestimation des pluies de la fin du mois avec :
    * un décalage de 24 h pour la prévision des petites pluies des 22 et 27,
    * une fausse alerte les 30 et 31.

Pour ce qui est de la prévision par la méthode S-12CP (figure VI-10b), un mieux global a été observé avec :

- une meilleure appréhension de la période non pluvieuse,
- une mise en alerte pour les pics des 19 et 20, même s'ils restent sous-estimés (quantile 90% à 90 mm pour 150 observés le 19 et 60 prévus pour 110 observés le lendemain),
- une bonne prévision des petites pluies de la fin du mois,
- la disparition de l'alerte du 30 et 31.

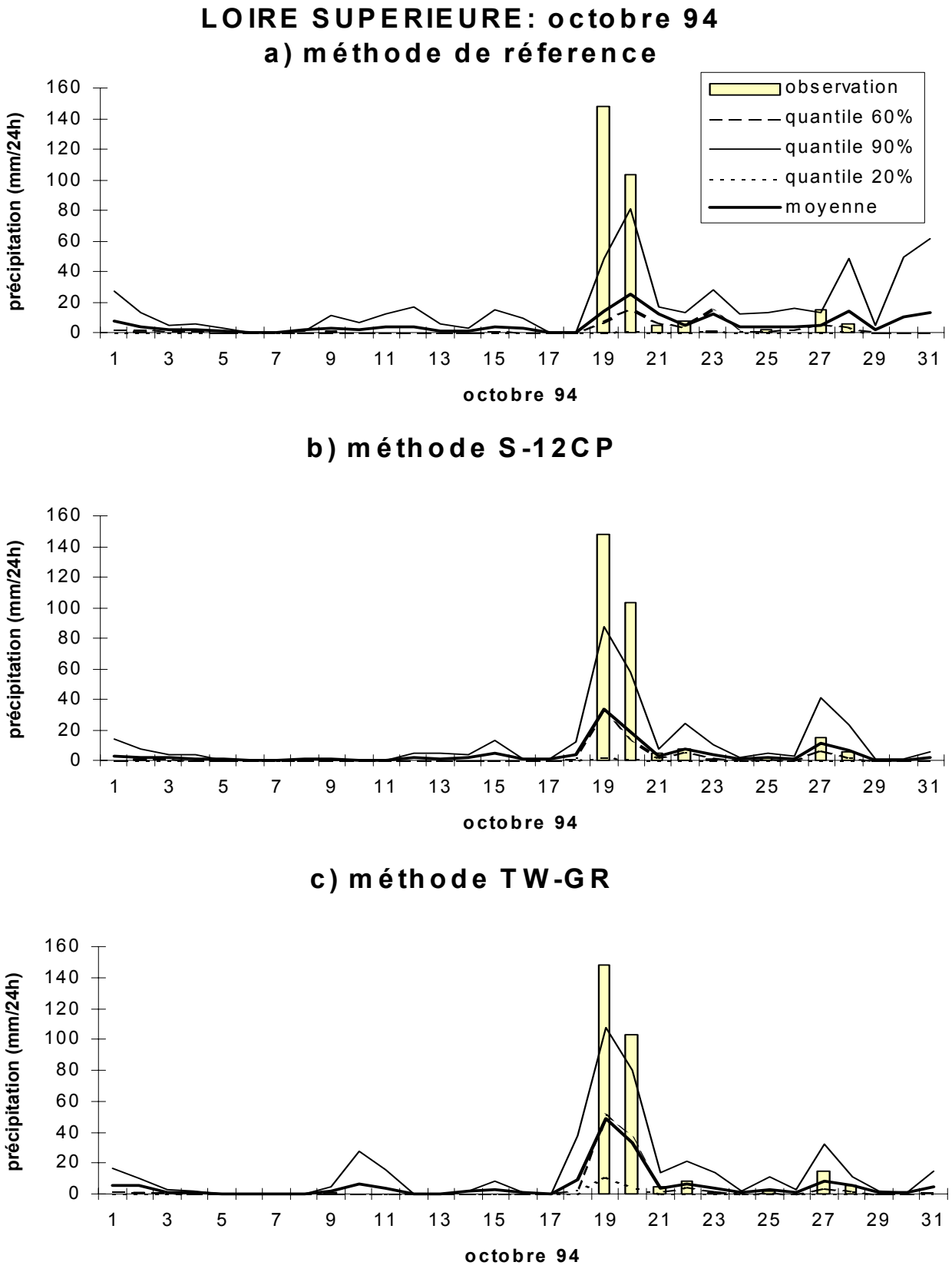


figure VI-10 : Loire supérieure, octobre 94

a) méthode de référence, b) méthode S-12CP, c) méthode TW-GR

Enfin, avec la méthode TW-GR, la prévision est encore meilleure puisque le pic du 19-20 octobre a été prévu avec des intensités plus importantes: 110 mm pour le 19 et 80 pour le 20. Et à part un petit écart du quantile 90% le 10, les résultats sont similaires à ceux de la méthode S-12CP.

Novembre 1994 :

Si la méthode de référence a bien prévu le pic du 4 (quantile 90% à 110 mm pour 130 observés), elle a surestimé les pluies du 3 et 5. La méthode S-12CP, quant à elle, a mieux prévue l'évolution de l'événement pluvieux. En contrepartie, le quantile 90% pour le 4 novembre est un peu plus faible avec 90 mm. Enfin, et comme pour le mois d'octobre, c'est la méthode TW-GR qui reproduit le plus fidèlement l'observation, avec un quantile 90% pour le pic du 4 novembre égal à la valeur observée, et une moyenne des analogues proche des pluies moins importantes observées autour du pic.

Ces quelques résultats graphiques nous confortent dans l'idée de remplacer la méthode de référence par la méthode TW-GR. Celle-ci s'est révélée la plus performante, que ce soit :

- à la lecture des résultats quantitatifs sur la période 1953-1993 pendant le calage de la méthode,
- lors de l'évaluation quantitative sur les automnes 1995 et 1996,

et nous en avons la confirmation ici.

VI.2.3.c Prévision à 24 h avec champs prédicteurs observés ou prévus

Dans le paragraphe §VI.2.2.a, nous avons montré numériquement que la prévision à 24 h par les méthodes S-12CP et TW-GR n'était pratiquement pas dégradée par le passage en opérationnel, avec des champs prédicteurs à 24 h prévus. Cela se vérifie qualitativement, et pour la méthode TW-GR, et pour la méthode S-12CP comme on peut le constater, pour le mois de novembre 1996, sur la figure VI-12 (méthode TW-GR) et l'annexe VI-6 (méthode S-12CP).

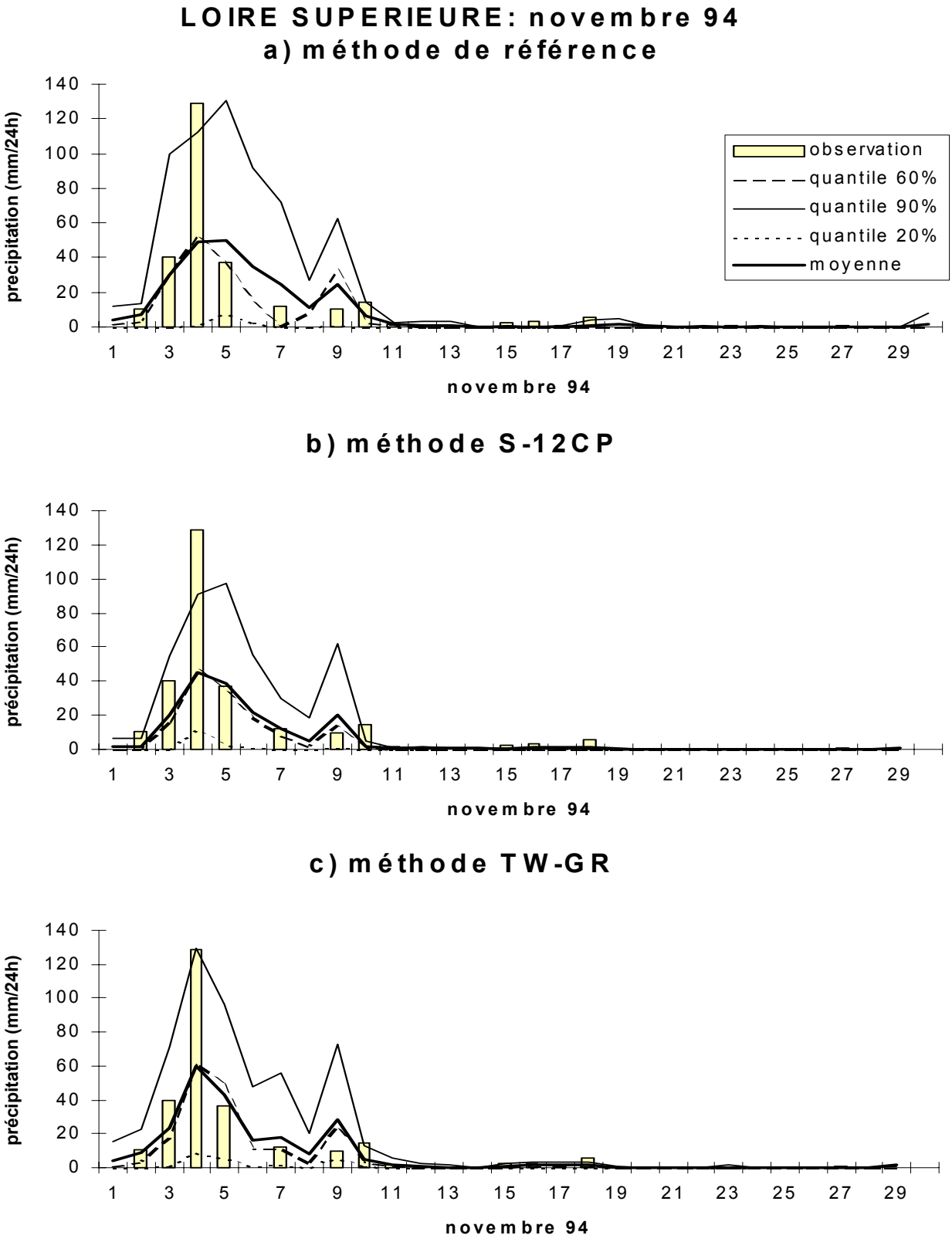


figure VI-11 : Loire supérieure, novembre 94

a) méthode de référence, b) méthode S-12CP, c) méthode TW-GR

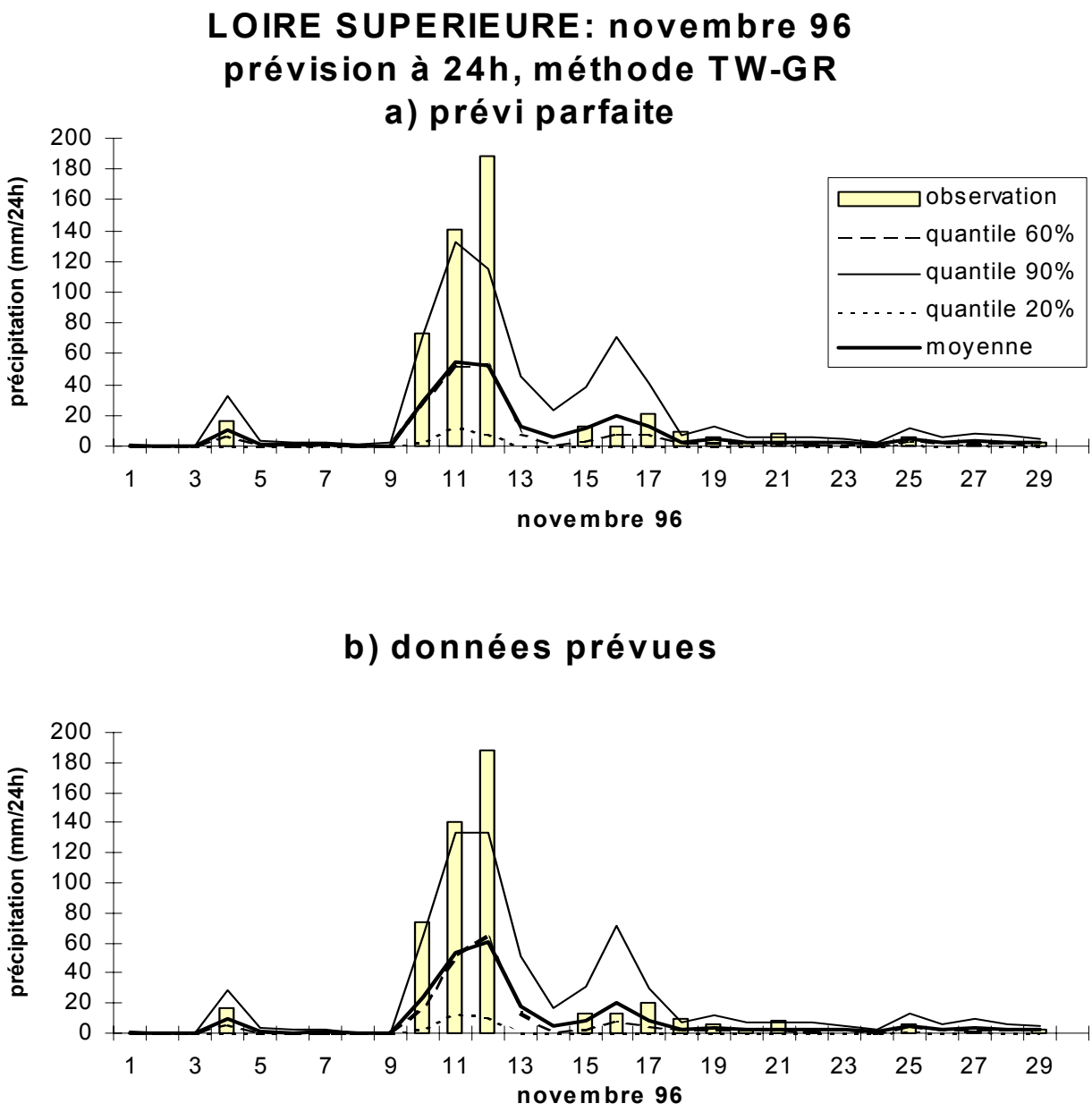


figure VI-12 : Loire supérieure, novembre 96, méthode TW-GR
a) prévi parfaite, b) champs prévus

VI.2.3.d Prévision à des échéances supérieures

Il s'agit maintenant de valider graphiquement les résultats obtenus au paragraphe §VI.2.2b, à savoir une dégradation des performances de la prévision avec l'augmentation de l'échéance, en particulier à 72 et 96 h pour la méthode S-12CP.

Pour cela, sur un même graphique ont été portées, pour un jour donné, les prévisions faites 24 (avec les champs prévus), 48, 72 et 96 h auparavant. Et comme il n'était pas possible de mettre les 3 quantiles et la moyenne des pluies des analogues sur un même graphe, seulement deux types de graphes ont été tracés : avec le quantile 90% et la moyenne, qui sont aussi les plus utilisés par les prévisionnistes.

Quelques exemples illustrent très bien les résultats du paragraphe §VI.2.2.b. C'est le cas de la prévision par la moyenne des analogues, pour le mois d'octobre 96 (cf. figure VI-13), où l'on remarque une dégradation beaucoup plus marquée à 72 et 96 h pour la méthode S-12CP. Elle se retrouve avec le quantile 90% (annexe VI-7).

Remarque : pour ce mois d'octobre 1996, il est intéressant de remarquer que, contrairement aux quelques exemples présentés ci-dessus, les 3 méthodes donnent des résultats équivalents en terme de prévision. Cependant les pluies de ce mois y sont beaucoup moins importantes.

Pour un épisode plus sévère, comme celui du mois de novembre 96, on retrouve les mêmes résultats pour la moyenne des analogues (annexe VI-8). Pour le quantile 90% (cf. annexe VI-9), c'est encore la même chose entre S-12CP et TW-GR. Pour la méthode de référence, il est plus difficile de juger car le 13 novembre un nombre insuffisant d'analogues n'a pas permis de déterminer les quantiles et seule la moyenne a été donnée comme prévision.

VI.2.3.e Extension au Piémont

En novembre 1994, des pluies exceptionnelles ont générées des crues catastrophiques sur les versants italiens du Sud des Alpes en Ligurie et Piémont. Possédant quelques données pluviométriques sur les bassins du Piémont pour ce mois-ci, c'était l'occasion de vérifier les capacités de mise en alerte des méthodes de prévision dans cette région lors de tels événements.

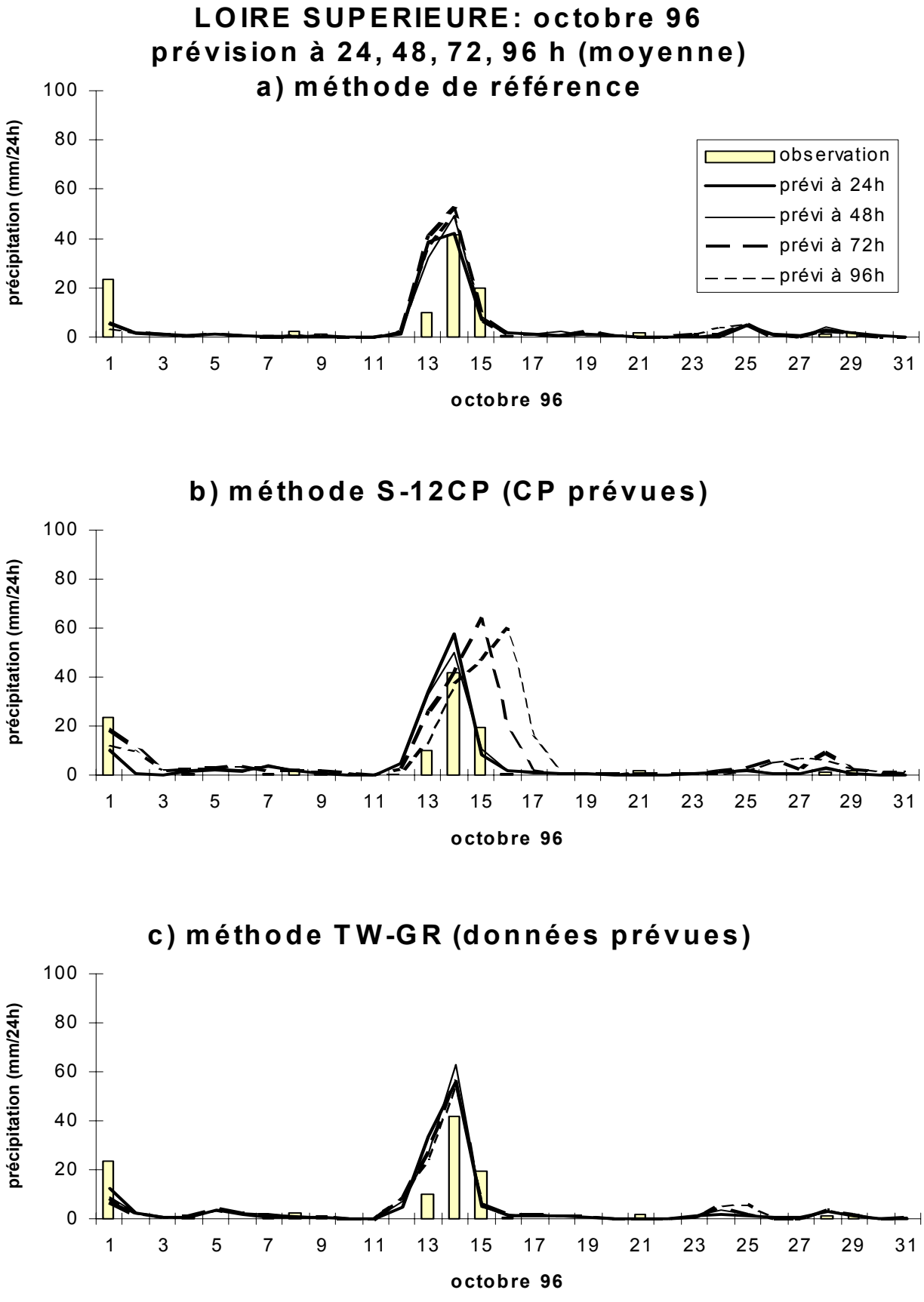


figure VI-13 : Loire supérieure, octobre 96, prévision à 24, 48, 72 et 96 h (moyenne)

Deux bassins ont pu être reconstitués en novembre 1994, même si toutes les stations n'étaient pas disponibles. Il s'agit de :

- Dora Riparia avec 3 stations sur 6,
- et du Tanaro avec 4 stations sur 7 (cf. figure VI-1).

La figure VI-14 présente, pour le mois de novembre 94 et le bassin Tanaro, la prévision à 24 h faite par les 3 méthodes et comparée à la pluie moyenne observée sur les 4 stations du bassin. En annexe VI-10 sont joints les mêmes graphiques pour le bassin Dora Riparia avec comme pluie observée, la moyenne des 3 stations disponibles.

On constate, là encore, une nette amélioration des prévisions avec la méthode TW-GR, que ce soit pour la période pluvieuse, où le quantile 90% a permis une mise en alerte avec une pluie prévue égale à la pluie observée, et pour la période sèche. La méthode S-12CP a, elle aussi, bien reproduit la forme des pluies, mais en sous-estimant leur quantité.

VI.2.4 Conclusion

Les deux types de validation, quantitative et qualitative, ont permis de confirmer ce qui est apparu lors de la validation croisée sur la période 1953-1993 : la supériorité de la méthode TW-GR. De plus, on a pu noter que:

- elle s'est avérée **performante**, que ce soit pour prévoir :
  * les épisodes pluvieux, extrêmes ou non, avec **une nette diminution des défauts d'alerte** par rapport à la méthode de référence,
  * et les périodes sèches, avec **moins de fausses alertes** que la méthode de référence,
- le passage en opérationnel (champs prédicteurs en prévision parfaite à 24 h $\rightarrow$ champs prévus à 24 h) n'a pas altéré ses qualités,
- **aux échéances supérieures, la dégradation de la prévision reste minime**.

Par contre, même si elle est dans l'ensemble meilleure que la méthode de référence, la méthode S-12CP présente quelques inconvénients comme :

- une nette baisse de ses performances pour des échéances supérieures à 48 h,
- une tendance à sous-estimer les pluies lors d'épisodes violents.

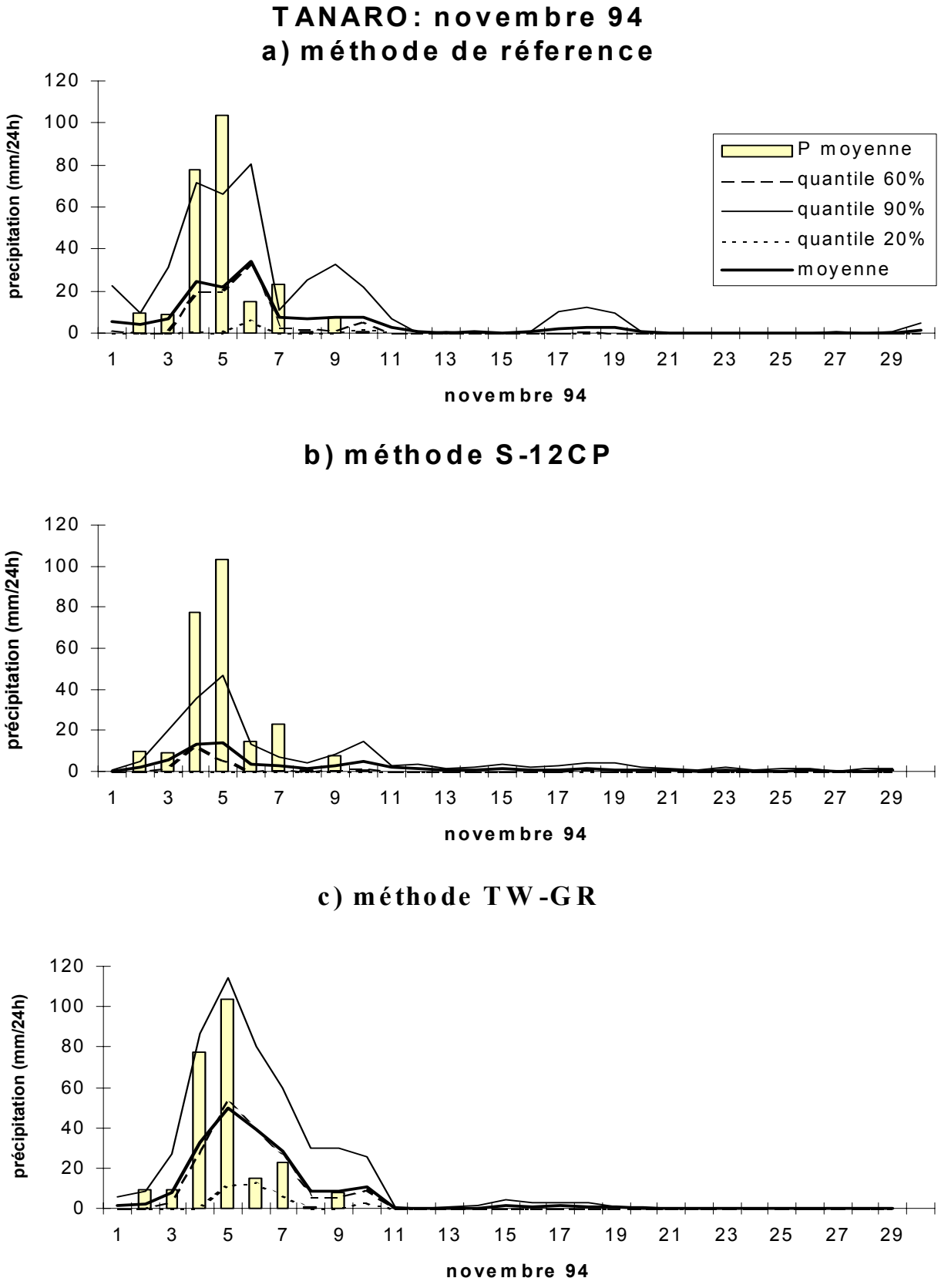


figure VI-14 : Tanaro, novembre 1994, prévision à 24 h par les 3 méthodes

VI.3 Conclusion du chapitre VI

Finalement, la supériorité de la méthode TW-GR, présentée numériquement en calibration sur la période 1953-1993 dans les chapitres précédents, s'est confirmée en validation sur les trois derniers automnes, de manière qualitative et quantitative, que ce soit au niveau :

- des qualités de sa prévision à 24 h,
- de la régularité de sa prévision pour des échéances supérieures,
- de son passage en opérationnel.

Elle semble donc au point pour être utilisée de manière opérationnelle, à la place de la méthode de référence. De plus, et quelle que soit la méthode (référence ou TW-GR), l'utilisation des champs de géopotentiels prévus à 24 h, 48 h etc., ne dégrade pas sensiblement la capacité de prévision. Notamment, elle semble permettre une prévision des pluies qui se dégrade beaucoup moins que celle élaborée de manière physico-déterministe par les modèles météorologiques. Il faut donc utiliser les qualités hydrodynamiques de ceux-ci pour une bonne prévision des champs de géopotentiels. Par contre, la méthode des analogues, en court-circuitant la propagation des erreurs dans la modélisation thermodynamique reste beaucoup plus robuste au fil des échéances.

De plus, après ses résultats prometteurs sur le Piémont pour l'événement de novembre 1994, les régions frontalières de Ligurie et du Piémont sont vivement intéressées par l'extension de la méthode à leurs bassins particulièrement sensibles, même s'il faudrait, pour cela, reprendre la formation de groupements de stations à pluviométrie homogène.

CONCLUSION GENERALE

et

PERSPECTIVES

1. Conclusions

Le but de notre étude était une reconsidération complète de la méthode de prévision de pluie par recherche de situations analogues, élaborée dans les années 70 par Duband au Service Ressources en Eau de Electricité de France (Division Technique Générale).

Cette « remise à plat » s'est faite de la façon la plus systématique possible, en fonction des données et des moyens de calcul disponibles à l'heure actuelle (1995-1997). Nous nous sommes donc penchés sur :

① les prédicteurs:

- en reconsidérant les prédicteurs initiaux (champs de géopotentiels 700 et 1000 hPa),
- et la forme sous laquelle les utiliser :
  * Composantes Principales (CP),
  * données brutes des 37 RadioSondages (RS),
  * points de grille obtenus par interpolation des RS,
- et en y rajoutant de nouvelles données à utiliser comme prédicteurs, qu'elles soient locales ou synoptiques,

② les prédictands,

- en introduisant de nouveaux groupements de pluie sur lesquels faire la prévision afin d'évaluer les possibilités d'extension de la méthode à des bassins frontaliers :
  * en Italie (Ligurie, Piémont),
  * en Espagne (Catalogne),

③ le critère d'analogie.

Parti d'une distance euclidienne utilisant un certain nombre de CP, nous avons tenté de :

- sélectionner et/ou pondérer les variables à introduire dans cette distance euclidienne,
- remplacer ce critère par un autre comme
  * le score de Teweles-Wobus,
  * un coefficient de corrélation,
  * une combinaison de critères.

A la suite de tous ces essais, une nouvelle méthode s'est imposée, la **méthode TW-GR**, qui utilise :

- comme prédicteurs les *champs de géopotentiels initiaux* (700 et 1000 hPa) *à 00 et 24 h,* fournis aux *points d'une grille* de taille intermédiaire qui a été optimisée, tout comme sa localisation suivant la zone d'étude,

- avec comme critère de sélection des analogues, *le score de Teweles-Wobus*.

Cette proposition a été validée sur différents bassins et sur divers épisodes, plus ou moins violents, survenus dans les 3 derniers automnes (1994, 1995 et 1996). Ses qualités prévisionnelles, pressenties par les scores de réussite obtenus sur la période de calage (1953-1993), n'ont pas été démenties à la validation puisque l'on constate :

- **une nette baisse des fausses alertes** par rapport à la méthode de référence,

- **une diminution des défauts d'alerte** : les épisodes intenses ont tous été prévus, même s'ils sont parfois encore un peu sous-estimés,

- **une robustesse appréciable:** la prévision se dégrade peu quand l'échéance augmente, contrairement à la méthode S-12CP par exemple.

Actuellement, cette méthode largement rénovée est en cours d'implantation à EDF pour une mise en service opérationnel la plus rapide possible sur les bassins français.
Elle a aussi été testée sur quelques bassins frontaliers catalans (Espagne) et du côté italien en Ligurie et Piémont. Si du travail reste à faire au niveau de la collecte des données pluviométriques, les résultats sont cependant tout aussi encourageants que ceux obtenus sur les bassins français.

Enfin, sur cette méthode TW-GR, peut se greffer une sélection d'analogues de 2ème niveau, faite à l'aide de variables locales d'un radiosondage proche de la zone d'étude. Cependant, seules les données à 12 h étant réellement intéressantes, il faudrait envisager, pour un passage en opérationnel, de faire cette 2ème sélection en milieu de journée, après avoir reçu les données du radiosondage de 12 h, et non le matin avec des données prévues dont la qualité n'est pas suffisante.
De plus, les données d'un seul radiosondage (Nîmes) ont été testées, et nous avons vu qu'elles avaient un rayon d'influence de l'ordre de 150 km. Il faudrait donc envisager de récupérer d'autres radiosondages pour les bassins auxquels Nîmes n'apporte pas d'information.

2. Perspectives

Pour terminer, quelques perspectives d'utilisation et d'amélioration de cette méthode TW-GR ont été envisagées et sont brièvement évoquées ci-dessous:

2.1 Perspectives d'utilisation

Jusqu'à maintenant, la méthode de prévision était utilisée à EDF pour gérer les risques de crues sur une trentaine de bassins versants équipés d'aménagements hydroélectriques. Au stade actuel, la nouvelle méthode présente un intérêt certain pour assurer :

i) la mise en vigilance jusqu'à 2 ou 3 jours en avance.

En effet, la qualité de prévision des champs synoptiques fournis par les modèles météorologiques reste très acceptable, même à des échéances éloignées. Ce n'est pas le cas de leurs prévisions quantitatives de pluie, pour diverses raisons évoquées en introduction:

- inadéquation de la taille des mailles aux zones pluviométriques homogènes,
- paramétrisation et non-linéarité des processus microphysiques,
- propagation des erreurs.

Par contre, l'approche par analogues traite mieux la répartition spatiale des pluies et n'amplifie pas les erreurs de prévision des champs de géopotentiels. C'est donc un outil de mise en vigilance très intéressant à 2 à 4 jours (par exemple pour gérer le suivi des week-ends) qui pourrait être exploité, au-delà des services d'EDF, sur d'autres bassins versants à risque, notamment dans le Sud-Est.

ii) la mise en alerte, au niveau journalier.

Celle-ci est particulièrement efficace pour les épisodes d'automne souvent marqués par une recrudescence vespérale. En effet, on fournit une prévision sur 24 h dès 7 h du matin, donc souvent avec une anticipation de 6 à 12 h sur la période intense de précipitation. Et surtout, on peut :

- affiner et actualiser à 12 h avec des données locales,
- et combiner avec les prévisions quantitatives de précipitations des modèles, raisonnablement fiables à 3, 6 et 12 h.

2.2 Autres perspectives d'utilisation

i) Une perspective peu utilisée mais souvent demandée par les prévisionnistes est la *fourniture de dates de situations analogues* (évidemment hiérarchisées) auxquelles ils puissent se « raccrocher » pour visualiser des situations réelles susceptibles de se reproduire.
Certaines pourraient même alimenter des systèmes d'information comme Symposium à Météo-France ou être transmises au public.

ii) Il est évident que les systèmes de prévision hydrométéorologique, notamment en zone méditerranéenne, ne peuvent se contenter de prévision en cumuls journaliers. Ils ont besoin de *scénarios de hyétogramme, heure par heure*.
Les modèles physico-déterministes peuvent en proposer, mais un ou quelques uns pour une prévision d'ensemble, et sans indice de qualité sur le scénario.
Une possibilité aujourd'hui est d'utiliser un générateur statistique de scénarios horaires. De tels générateurs existent. Ils sont en général conditionnés par le passé, déjà réalisé, de l'épisode en cours, mais pas sur l'avenir (sauf prévision immédiate à 1 ou 2 h par extrapolation radar).
La prévision probabiliste permettrait de contraindre ou de censurer les scénarios, de manière à ne prendre en compte que ceux qui « rentrent » dans la fourchette de prévision probabiliste à 24 h, en introduisant, en plus des contraintes imposées par les pluies déjà observées, une contrainte sur le cumul journalier des pluies à venir.
On espère que cette information supplémentaire introduite dans le générateur permettra une amélioration sensible *des prévisions de débit à échéance de 6 ou 12 h,* ce qui sera validé sur des bassins de l'Ardèche, avec Topmodel comme modèle de prévision.

iii) Enfin, la méthode intéresse d'autres utilisateurs comme par exemple les *régions du Piémont et de Ligurie* qui souhaitent l'implanter chez eux afin de mieux prévoir les épisodes pluvieux extrêmes, comme celui de novembre 1994.

2.3 Perspectives d'amélioration

i) Une reconsidération des groupements français est à faire tout d'abord, et ceci pour deux raisons :

* certaines stations pluviométriques ont disparu, ce qui rend difficile, voire impossible, la reconstitution de certains groupements,

* certains groupements sont à revoir pour des questions d'homogénéité pluviométrique, au regard des travaux de Champeaux et Tamburini (1996).

ii) L*'utilisation de la sélection de 2ème niveau* après la méthode TW-GR, nécessite l'introduction d'autres radiosondages comme

* Palma de Majorque pour les bassins catalans et pyrénéens,

* Bordeaux pour les bassins situés plus à l'Ouest,

* Milan pour les bassins italiens.

iii) Finalement l'objection principale mise en avant pour l'utilisation de cette méthode était la nécessité de constituer, ou d'acquérir, un *fichier historique* de situations d'au moins 40 ans.
Cela n'existait pas au niveau européen, et de manière très hétérogène chez les atmosphériciens américains. Cependant, les nécessités de la recherche climatique ont conduit à la constitution de tels fichiers. Par exemple, le NCEP et le NCAR (National Centers for Environmental Prediction et National Center for Atmospheric Research) ont réanalysé, avec le même modèle, 40 années de données synoptiques à tous les niveaux. Et le Centre Européen envisage de faire de même.
C'est une occasion inespérée de tester et d'utiliser l'approche par analogue, puisque son principal handicap, le fichier historique des situations synoptiques, est en passe d'être résolu.
Il ne reste alors qu'à constituer le fichier local de la variable prédictand (pluie moyenne sur un groupement ou bassin) dont souvent l'utilisateur dispose déjà.

iv) Une cartographie des prévisions peut aussi être envisagée afin de montrer la cohérence spatiale qui existe pour certains épisodes.

iv) Bien souvent, des modèles de type physico-déterministe se retrouvent en compétition avec des méthodes plus ou moins statistiques alors que le plus efficace serait sûrement de *les*

combiner. Un exemple en est le bénéfice que retire la prévision par analogues de la qualité de la prévision des champs de géopotentiels à 3 ou 4 jours.

Tout ceci en fait, à nos yeux, et nous nous sommes employés à le démontrer, une approche pragmatique, fédérant les points forts des modèles à la fois statistiques et déterministes, et susceptible d'assurer un progrès réel et une transition améliorée dans l'attente de modèles plus performants.

De plus, elle est aisément généralisable à de nouveaux bassins et tire même avantage d'une certaine « continuité » géographique. Il serait donc souhaitable qu'elle se diffuse plus largement et que les gestionnaires des grands systèmes d'annonce de crue l'inscrivent parmi leurs outils opérationnels.

REFERENCES BIBLIOGRAPHIQUES
et
BIBLIOGRAPHIE

REFERENCES BIBLIOGRAPHIQUES

Barnett T.P., Preisendorfer R.W., 1977: Short-term prediction of global climate using multi-dimensional analog methods. *Proceedings of the 5th AMS Conference on Probability and Statisctics in Atmospheric Sciences, 15-18 Nov. 1977, Las Vegas - USA, pp. 11-15.*

Barnett T.P., Preisendorfer R.W., 1978: Multifield analog prediction of short-term fluctuations using a climate state vector. *Journal of Atmospheric Sciences, vol.35,n°10, pp.1771-1787.*

Barnston A.G., Livezey R.E., 1986: An operational multifield analog prediction system. *Proceedings of the first WMO Workshop on the diagnosis and prediction of monthly ans seasonal Atmospheric variations over the globe, Long-Range forecasting Res. Rep. Ser. 6, Vol II, Tech. Doc., WMO/TD87, pp. 671, WMO, Geneva, Switzerland.*

Bell V.A., Moore R.J., 1997: A water-balance storm model for short-term rainfall and flood forecasting at the catchment scale using radar and satellite data. *Worshop expert meeting on "Integrated systems for real-time flood forecasting and warning", Padova, Italy.*

Bengio Y., 1991: Artificial neural networks and their application to sequence recognition. *Ph D. Thesis, Mc Gill University, Montréal.*

Bergen R.E., Harnack R.P., 1982: Long-range temperature prediction using a simple analog approach. *Monthly Weather Review, Vol 110, pp. 1082-1099.*

Berlin V., Cendrier D., 1986: Typologie de stations de mesures pluviométriques dans le bassin de la moyenne Durance - Analyse en Composantes Principales de Processus et fonctions empiriques orthogonales appliquées à un réseau de stations de radiosondage. *Rapport de stage de fin d'études de l'Ecole de la Météorologie Nationale, 264 p.*

Bidner, 1970: The air force global weather central Severe WEAther Threat (SWEAT) index - A preliminary report. *Air Weather Service Aerospace Sciences review, AWS, RP 105-2, 70-3, 2-5.*

Bois Ph., Obled Ch., Thalamy J., 1981: Etude des liaisons entre champs de pression et températures sur l'Europe avec la durée d'insolation: Application à la possibilité de la prévision d'insolation. *C.R. du colloque "Météorologie de l'énergie solaire", PIRDES-CNRS, Toulouse-France, pp. 145-177.*

Bolognesi R., 1993: Premiers développements d'un modèle hybride pour le diagnostic spatial des risques d'avalanches.

Bouhaddou O., 1984: Analyse en composantes principales et interpolation de processus. Méthode et simulation. *Thèse de 3ème cycle USMG, Grenoble, 165 pp.*

Bowen D., 1976: Long-range weather forecasting. *Water Power and Dam Construction.*

Braud I., 1990: Etude méthodologique de l'analyse en composantes principales de processus bidimensionnels - Effets des approximations numériques et de l'échantillonnage et utilisation pour la simulation de champs aléatoires - Application au traitement des températures mensuelles de surface de la mer sur l'Atlantique intertropicale. *Thèse de doctorat de l'Institut National Ploytechnique de Grenoble, 214 p.*

Brier G.W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review, Vol 78, pp. 1-3.*

Browning K.A., Collier C.G., 1989: Nowcasting of precipitations systems. *Rev. Geophysics, Vol 27 (3), pp. 345-370.*

Champeaux J.-L., Tamburini A., 1996: Zonage climatique de la France à partir des séries de précipitations (1971-90) du réseau climatique d'état. *La Météorologie - 8ème série N° 14, pp. 44-54.*

Davalo E., Naïm P., 1990: Des réseaux de neurones. *Ed. Eyrolles, 232p.*

David C.L., Smith J.S, 1971: An evaluation of seven stability indices as predictors of sever thunderstorms and tornadoes. *Preprints, Seventh Conf. Sever Local Storms, Kansas City, Amer. Meteor. Soc., pp. 105-109.*

Déqué M., Royer J.F., Veysseire J.M., 1988: Estimation de la qualité des prévisions probabilistes par classes ordonnées. *Note de travail n°207 de l'Etablissement d'Etudes et de Recherche Météorologiques.*

Deville, 1974: Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE, n°15, pp. 1-101.*

Dimopoulos I., Lek S., Lauga J., 1996: Modélisation de la relation pluie-débit par les réseaux connexionnistes et le filtre de Kalman. *Hydrological Sciences Journal, Vol. 41, pp. 179-193.*

Dolcine L., 1997: Prévision quantitative à très courte échéance de la pluie. Modèle global adapté à l'information radar. *Thèse de l'Université J. Fourier, Grenoble, 183 pp.*

Duband D., 1970: Reconnaissance dynamique de la forme des situations météorologiques. Application à la prévision quantitative des précipitations. *Thèse de 3ème cycle de la faculté des sciences de Paris.*

Duband D., 1971: Prévision quantitative des précipitations en automne. *Rapport interne - EDF-DTG.*

Duband D., 1974: Reconnaissance dynamique de la forme des situations météorologiques. Application à la prévision quantitative des précipitations. *Congrès de la Société Hydrotechnique de France, XIIIèmes journées de l'Hydraulique, Paris-France.*

Duband D., 1980: Dynamic selection of analogue flow patterns to produce quantitative precipitation forecasts, *WMO symposium on Probabilistic and Statistical Methods in Weather Forecating, Nice-France, Septembre, pp. 487-492.*

Duband D., 1981: Prévision spatiale des hauteurs de précipitations journalières. *La Houille Blanche, n°7/8, pp. 497-511.*

Ducroc V., 1997 : Apport scientifique de la modélisation à échelle fine pour l'étude des phénomènes exceptionnels convectifs du Sud-Est. *Atmosphère et Climat, Lettre d'information de Météo-France n°73, pp. 23-24.*

Durot K., 1996: critique et validation des mesures physico-chimiques aux abords des centrales nucléaires sur la Loire: étude du pH. *Rapport de DEA, Université Joseph Fourier, Grenoble-France, 58 p.*

Epstein E.S., 1969: A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology, Vol 8, pp. 985-987.*

Fessant F., 1995: Prédiction de séries temporelles par réseaux de neurones artificiels: application aux séries temporelles ionosphériques. *Thèse de l'Université de Rennes, 161 pages.*

Galway J.G., 1956: The lifted index as a predictor of latent instability. *Bull. Amer. Meteor. Soc., Vol. 37, 10, 528-529.*

Georges J.J., 1960: Weather forecasting for aeronautics. *Academic Press, pp. 407-415.*

Gibergans J., 1995: Application de la méthode des analogues à la Catalogne. *Rapport interne de stage à EDF-DTG*

Glahn H.R., Lowry D.A., 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor., Vol 11, pp. 1203-1211.*

Gordon N., 1987: Statistical very short-range forecating via analogues. *Proceedings of the Symposium on Mesoscale Analysis and Forecasting, Vancouver-Canada, 17-19 August 1987, ESA SP-282.*

Grégoris Y., 1996: Modélisation de la propagation de débits à l'aide de Réseaux de Neurones. *Note interne de EDF, 29 pages.*

Gruza G.V., Rankova E.Ya., 1980: Long-range weather forecasting using a group of analogs and evaluation of meteorological predictability. *WMO symposium on Probabilistic and Statistical Methods in Weather Forecating, Nice-France, Septembre, pp. 269-276.*

Guilbaud S., 1994: Développements de la méthode de prévision quantitative des précipitations SENALOG: exploitation des champs 700 et 1000 mb et utilisation de l'analyse discriminante. *Rapport de DEA, Université Joseph Fourier, Grenoble-France, 58 p.*

Harnack R., Cammarata M., Dixon K., Lanzante J., Harnack J., 1985: Summary of U.S. seasonal temperature forecast experiments, *Proceedings of the 9th AMS Conference on Probability and Statistics in Atmospheric Sciences, Virginia Beach - USA, 9-11 August 1985, pp. 175-179.*

Kruizinga S., Murphy A., 1983: Use of an analogue procedure to formulate objective probabilistic temperature forecasts in the Netherlands. *Monthly Weather Review, Vol 111, pp. 2244-2254.*

Livezey R.E. and Barnston A.G., 1988: An operational multifield analog/antianalog prediction system for United States seasonal temperatures, Part I: System design and winter experiments. *Journal of Geophysical Research, Vol 93, pp. 10953-10974.*

Livezey R.E. and Barnston A.G., 1989: An operational multifield analog/antianalog prediction system for United States seasonal temperatures, Part II: Spring, Summer, Fall and intermediate 3-month period experiments. *Journal of Climate, Vol 2, pp. 513-541.*

Llasat M.C. and Puigcerver M., 1992: Pluies extrêmes en Catalogne - Influence orographique et caractéristiques synoptiques. *Hydrologie continentale, Vol. 7, n°2, pp. 99-115.*

Lorenz E., 1969: Atmospheric predictability as revealed by natural occuring analogues. *Journal of the Atmospheric Sciences, Vol 26, pp 636-646.*

Lorenz E., 1980: Nonlinear statistical weather prediction. *WMO symposium on Probabilistic and Statistical Methods in Weather Forecating, Nice-France, Septembre, pp. 487-492.*

Mandon S ,1985: Comparaison d'épisodes pluvieux intenses sur le Sud-Est de la France et de situations analogues au sens de la circulation générale. Recherche de variables discriminantes. *Rapport de stage de fin d'études de l'Ecole de la Météorologie Nationale, 176 p.*

Martin E., 1995: Modélisation de la climatologie nivale des Alpes françaises, application des techniques de régionalisation à l'étude de l'impact d'un changement climatique sur l'enneigement. *Thèse de doctorat de l'unversité Paul Sabatier, Toulouse 244 p.*

Miller R.C., 1967: Notes on analysis and severe storms forecasting procedures of the military weather warning center. *Tech. Rep., 200, AWS, USAF.*

Miller R.C., Bidner A., Maddox R.A., 1971: The use of computer products in severe weather forecasting (the SWEAT index). *Preprints, Seventh Conf. Severe Local Storms, Kansas City, Amer. Meteor. Soc., pp. 1-6.*

Moncrieff M.W., Miller M.J., 1976: The dynamics and simulation of tropical cumulonimbus and squall lines. *Q. J. R. Meteor. Soc., Vol.432, pp. 373-394.*

Murphy A.H., 1970: A note on the Ranked Probability Score. *Journal of Applied Meteorology, Vol 10, pp. 155-156.*

Murray R., 1974: Indicators of monthly mean temperature and rainfall for England and Wales based on antecedent monthly pressure anomalies over the Nothern hemisphere. *Meteor. Mag. , Vol 103, pp. 70-73.*

Nap J.L., Van den Dool H.M., Oerlemans J., 1981: A verification of monthly weather forecasts in the seventies. *Mon. Wea. Rev., Vol 109, 306-312.*

Navarre J.P.: Modèle de prévision des précipitations sur les Alpes, *Note interne du CEN - Météo-France, 7 p.*

Nieminen R., 1983: Operational verification of ECMWF forecast fields and results for 1980-1981. *Technical Report n°36, 11 p.*

Norrie D.H., De Vries G., 1973: The finite element method. *Academic Press.*

Obled Ch., 1979: Contribution à l'analyse des données en Hydrométéorologie: la prévision des phénomènes accidentels et l'analyse des champs spatiaux (application à la prévision des analogues à Davos et à l'analyse des épisodes pluvieux cévenols). *Thèse de Doctorat d'Etat es Sciences Physiques, INPG-USMG, Grenoble-France, 357 p.*

Obled Ch., Good W., 1980: Recent develoments in Avalanche Forecasting by statistical techniques: a methodological review and some applications to the Parsenn area (Davos-Switzerland). *Journal of Glaciology, vol XXV, n°92, pp. 315-346.*

Obled Ch., Creutin J.D., 1986: Some developments in the use of Empirical Orthogonal Functions for mapping meteorological fields. *Journal of Climate and Applied Meteorology, Vol. 25, n°9, pp. 1189-1204.*

Rumelhart D.E., Hinton G.E., Williams R.J., 1986: Learning intermational representations by error propoagation. *Parallel Distributed Processing: Explorations in the microstructures of cognition, ed. D/E/ Rumelhart & J. McClelland, Vol. 1, pp. 318-362, MIT Press, Cambridge, Massachusetts, USA.*

Saporta G., 1990 : Probabilité, Analyse de données et Statistique. *Edition Technip, 493 pp.*

Sénési S., Thépemier R.M., 1996: Indices d'instabilité et occurrence d'orage. *Soumis à la Météorologie en septembre 1996, 38 pages.*

Shabbar A. and Knox J., 1986: Monthly prediction by the analogue method, *Proceedings of the first WMO workshop on the diagnosis and prediction of monthly and seasonal atmospheric variations over the globe, Longe range forecasting Res. Rep. Ser. 6, Vol II, Tech. Doc. WMO/TD87, pp. 672-681.*

Showalter A.K., 1953: A stability index for thunderstorm forecasting. *Bull. Amer. Meteor. Soc., Vol 34, pp. 205-252.*

Sing S.V., R.H. Kripalani, P.M.M. Ismail and Mrs P. Saha, 1980: Forecasting monsoon precipitation by synoptic-cum-statistical methods, *WMO symposium on Probabilistic and Statistical Methods in Weather Forecating, Nice-France, Septembre, pp. 407-413.*

Teweles S. and Wobus H., 1954: Verification in prognostic charts, *Bulletin of the American Meteorology Society, Vol 35, n°10, pp. 455-463.*

Thalamy, J.,1981: Etude de quelques situations météorologiques ayant provaqué des crues sur les Cévennes. Prévision de durée d'insolation par des méthodes de plus proches voisins. *Rapport de stage de fin d'études de l'Ecole de la Météorologie Nationale, 93 p.*

Toth Z., 1989: Long-range weather forecasting using an analog approach. *Journal of Climate, Vol 2, 594-607.*

Vallée, J.L.,1986: Précipitations sur le Sud-Ouest du Massif Central et l'Est des Pyrénées. Optimisation du modèle EDF/DTG de prévision par recherche d'analogues. *Rapport de stage de fin d'études de l'Ecole de la Météorologie Nationale, 110 p.*

Van Den Dool H., 1987: A bias in skill in forecasts based on analogues and antianalogues, *Journal of Climate and Applied Meteorology, Vol 26, pp. 1278-1281.*

Van Den Dool H., 1989: A new look at weather forecasting through analogues, *Monthly Weather Review, Vol 117, pp. 2230-2247.*

Vermot-Desroches B., 1987: Modèle de reconnaissance des situations météorologiques pour la prévision quantitative des précipitations. *Rapport de stage de fin d'études de l'Ecole de la Météorologie Nationale,121 p.*

Verret R., 1985: Expermients on the selection of analogues used in preparing objective forecasts for days 3, 4 and 5. *CMC Information, Vol IV, n°1, pp. 2-17.*

Villé P., 1990: Compte-rendu de l'évaluation du modèle de recherche de situations analogues développé par J-P. Navarre. *Rapport interne du département Développement du SMIR/Centre-Est, 19 p.*

Wilson J., N. Yacowar, 1980: Statistical weather element forecasting in the Canadian Weather Service. *WMO symposium on Probabilistic and Statistical Methods in Weather Forecating, Nice-France, Septembre 1980, pp. 401-406.*

Woodcock F., 1980a: On the use of analogues to improve regression forecasts. *Monthly Weather Review, Vol 108, pp. 292-297.*

Woodcock F., 1980b: The use of analogues in statistical forecasts. *WMO symposium on Probabilistic and Statistical Methods in Weather Forecating, Nice-France, Septembre 1980, pp. 415-421*

Woodcock F., Keenan T.D, 1979: Objective tropical cyclone movement forecasts using synoptic and track analogue information. *Intermational Conference on Tropical Cyclones, Perth-Australia,25-29 nov.*

BIBLIOGRAPHIE

Bardossy A., Plate E., 1992 : Space-time model for daily rainfall using atmospheric circulation patterns. *Water Ressources Research, Vol 28, n°5, pp. 1247-1259.*

Bois Ph., 1991: Cours polycopié d'Introduction au traitement de données en hydrologie. *Ecole Nationale Supérieure d'Hydraulique et de Mécanique de Grenoble - INPG.*

Bois Ph., Obled Ch., 1976: Prévision des avalanches par des méthodes statistiques: Aspects méthodologiques et opérationnels. *La Houille Blanche, n°6/7, pp. 509-531.*

Braud I., Obled Ch., 1989: A Comparison between Analytical Eigenfunctions of a Covariance Kernel and their Numerical Approximations. *Proc. of 11th Conf. On Probability and Statistics in Atmospheric Sciences, Amer. Met. Soc., Monterey USA, October 1989, pp. 26-29*

Braud I., Obled Ch., Pham Dinh T., 1993: Empirical Orthogonal Functions (EOF) Analysis of Spatial Random Fields: Theory, Accuracy of the numerical approximations and Sampling Effects. *Stochastic Hydrology and Hydraulics, vol. 7, pp. 146-160*

Brier G.W. and Allen R., 1952: Verifcation of weather forecast. *Compendium of Meteorology, Boston, Amer. Meteor. Soc., pp. 841-848.*

Creutin J.D., Obled Ch., 1980: Elimination de variables et optimisation de réseaux de mesures. *2èmes journées Internationales "Analyse des Données et Informatique"- Versailles 1979- publ. in Data Analysis and Informatics, E. Diday Ed., North Holland Pub. Comp.,pp. 759-775.*

Duband D., 1992: Cours polycopié d'hydrologie approfondie. *Ecole Nationale Supérieure d'Hydraulique et de Mécanique de Grenoble - INPG.*

Epstein E.S. and Murphy A.H., 1965: A note on the attributes of probabilistic predictions and the probability score. *Journal of Applied Meteorology, Vol 4, pp. 297-299.*

Fenelon J.P., 1981 : Qu'est-ce-que l'Analyse des Données ? *Edition Lefonen.*

Foehn P., Good W., Bois Ph., Obled Ch., 1977: Evaluation and comparison of statistical and conventional methods to forecast avalanche hazard. *Journal of Glaciology, vol. 19, n°31, pp. 375-387.*

Gruza G.V., Reitenbakh R.G., 1973: Analogy principle in atmospheric processes predictability studies and in solution of weather forecasting problems, *Meteorology and Hydrology, n°11, pp 22-31.*

Gutzler D.S. and Shukla J., 1984: Analogs in the wintertime 500 mb height field, *Journal of the Atmospheric Sciences, Vol 41, n°2, pp. 177-189*.

Lebart L., Morineau A., Fenelon J.P., 1982 : Traitement des données statistiques - Méthodes et programmation. *Edition Dunod.*

Livezey R.E., Barnston A.G., Gruza G.V., Rankova E.Ya., 1994: Comparative skill of two analog seasonal temperature prediction systems: objective selection of predictors, *Journal of Climate, Vol 7, pp. 608-615*.

Llasat M.C. and Rodriguez R., 1991: Extreme rainfall events in Catalonia - The case of 12 Novembre 1988. *Natural Hazards, n°5, pp. 133-151.*

Mamassis N., Koutsoyiannis D., 1996 : Influence of atmospheric circulation types on space-time distribution of intense rainfall. *Journal of geophysical research, Vol. 101, n° D21, pp. 26267-26276.*

Murphy A.H., 1969: On the Ranked Probability Score. *Journal of Applied Meteorology, Vol 8, pp. 989-991.*

Murphy A.H. and Epstein E.S., 1967: Verification of probabilistic predictions: a brief review. *Journal of Applied Meteorology, Vol 6, pp. 748-755.*

Murphy A.H. and Epstein E.S., 1967: A note on probability forecasts and « hedging ». *Journal of Applied Meteorology, Vol 6, pp. 1002-1004.*

Murphy A.H. and Katz R.W. (edited by), 1985: Probability, statistics, and decision making in the atmospheric sciences. *Westview Press, 545 pp.*

Neumann Ch.J., 1977: Simulated analog models for the prediction of tropical cyclone motion, *Preprints of the 5th Conference on Probability and Statistics in the Atmospheric Sciences, AMS, 15-18 Nov. 1977, Las Vegas - USA, pp.47-52.*

Radinovic D., 1975: An analogue method for weather forecasting using the 500/1000 mb Relative Topography, *Monthly Weather Review, Vol 103, pp. 639-649*.

Ramis C., Alonso S. and Llasat M.C., 1995: A comparative study between two cases of extreme rainfall events in Catalonia. *Surveys in Geophysics, n°16, pp. 141-161.01*

Sanders F., 1963: On subjective probability forecasting. *Journal of Applied Meteorology, Vol 2, pp. 191-201.*

Sanders F., 1966: The verification of probability forecasts. *Journal of Applied Meteorology, Vol 6, pp. 756-761.*

Tenenhaus M., 1994 : Méthodes statistiques en gestion. *Edition Dunod, 373 pp.*

Triplet J.P. et Roche G., 1971: Météorologie générale. *Ecole Nationale de la Météorologie, 317 pp.*

Verret R. and Yacowar N., 1989: Improvement of numerical weather element forecasts by combining forecasts from different procedures, *Preprints of the 11th Conference on Probability and Statistics, AMS, Monterey - USA, 1-5 October 1989, pp. 58-63.*

Wiesman M.L. and Klemp J.B., 1986: Characteristics of isolated convective storms. *Mesoscale Meteorology and Forecasting, chapter 15, pp. 331-358.*

Yacowar N., 1975: Probability forecasts using finely tuned analogs. *Fourth Conference on Probability and Statistics in Atmospheric Sciences, Tallahassee, Florida, 18-21 nov., Preprints, pp.49-50.*

ANNEXES DES CHAPITRES II, III, IV, V et VI

ANNEXES DU CHAPITRE II

ANNEXE II-1:

Rappels sur l'Analyse en Composantes Principales

L'Analyse en Composantes Principales ou ACP est une technique de traitement de données qui permet principalement de condenser l'information de manière significative en en perdant le moins possible: partant d'une matrice $[X_{np}]$ de données avec p variables et n observations, on aboutit, après une ACP, à une matrice $[Z_{nq}]$ réduite en colonnes ($q<p$).

1. Principe

La méthode de recherche des Composantes Principales (CP) consiste à faire un changement d'axe selon le critère suivant:

- le premier axe est choisi de manière à minimiser la distance des points d'observations à une droite. La variance expliquée par la nouvelle composante Z_1 sur cet axe ainsi défini est la plus grande possible,

- le deuxième axe, perpendiculaire au premier, est tel que la variance expliquée par Z_2 soit la plus grande après celle de Z_1,

etc...

2. Notations

Soit p variables $(X_1,...,X_p)$ pour lesquelles on a n observations. On a donc une matrice $[X_{np}]$ de données:

$$\left[X_{np}\right]=\begin{bmatrix} x_{11} & x_{12} & x_{13} & . & . & x_{1p} \\ x_{21} & x_{22} & x_{23} & . & . & x_{2p} \\ . & . & . & . & . & . \\ . & & & & & . \\ . & & & & & . \\ x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ X_1 \quad X_2 \quad\quad\quad\quad X_p \end{matrix}$$

On peut calculer:

- la moyenne de chaque variable, pour k=1 à p: $M_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ (1)

- l'écart type de chaque variable, pour k=1 à p:

$$\sigma_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ik} - M_k\right)^2} \tag{2}$$

- les coefficients de corrélation r_{jk} entre X_j et X_k:

$$r_{jk} = \frac{\sum_{i=1}^{n}\left(x_{ij} - M_j\right)\left(x_{ik} - M_k\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - M_j\right)^2 \sum_{i=1}^{n}\left(x_{ik} - M_k\right)^2}} \tag{3}$$

- la variance de la variable Xj: $c_{jj} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - M_j\right)^2 = \sigma_j^2$ (4)

- la covariance entre les variables X_j et X_k:

$$c_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - M_j\right)\left(x_{ik} - M_k\right) \tag{5}$$

Nous noterons R la matrice de corrélation et C la matrice de variance-covariance, matrices toutes 2 symétriques:

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & . & . & r_{1p} \\ . & 1 & r_{23} & . & . & r_{2p} \\ . & & 1 & & & . \\ . & & & . & & . \\ . & & & & . & . \\ . & . & . & . & . & 1 \end{bmatrix} \qquad C = \begin{bmatrix} c_{11} & c_{12} & c_{13} & . & . & c_{1p} \\ . & c_{22} & c_{23} & . & . & c_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ . & . & . & . & . & c_{pp} \end{bmatrix}$$

3. Théorie

Les composantes principales ou CP sont calculées à partir des p relations suivantes:

$$Z_k = \sum_{j=1}^{p} \frac{\left(X_j - M_j\right)}{\sigma_j} a_{jk} \qquad \text{pour k=1 à p} \qquad (6)$$

Les vecteurs Z_1 à Z_p sont les Composantes Principales ou CP de la matrice $[X_{np}]$. Mais dans la réalité seules q CP sont retenues, q<p. Ainsi, à la matrice $[X_{np}]$ correspond une autre matrice de données réduite en colonne $[Z_{nq}]$:

$$\left[Z_{nq}\right] = \begin{bmatrix} z_{11} & z_{12} & . & z_{1q} \\ z_{21} & z_{22} & . & z_{2q} \\ . & . & . & . \\ . & & & \\ . & & & \\ z_{n1} & z_{n2} & . & z_{nq} \\ Z_1 & Z_2 & & Z_q \end{bmatrix}$$

dont les éléments sont calculés grâce aux q relations (6):

$$z_{jk} = \sum_{i=1}^{n} \frac{\left(x_{ji} - M_i\right)}{\sigma_i} a_{ik} \quad \text{pour k=1 à q} \qquad (7)$$

Les coefficients a_{ik}, cosinus directeurs de la CP d'ordre k, sont les coordonnées des vecteurs propres F_1 à F_p de la matrice de corrélation R précédente, vecteurs propres associés aux valeurs propres λ_k. Ces vecteurs propres sont calculés dans l'ordre des valeurs propres décroissantes.

Sous la forme matricielle, si F est la matrice p x p des cosinus directeurs:

$$[F] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & . & . & a_{1p} \\ a_{21} & a_{22} & a_{23} & . & . & a_{2p} \\ . & . & . & . & . & . \\ . & & & & & . \\ . & & & & & . \\ a_{p1} & a_{p2} & . & . & . & a_{pp} \end{bmatrix}$$
$$F_1 \quad F_2 \qquad\qquad F_p$$

alors les valeurs propres sont obtenues en résolvant l'équation:

$$\mathbf{R.F=F.\Lambda} \tag{8}$$

où Λ est la matrice diagonale des valeurs propres:

$$[\Lambda] = \begin{bmatrix} \lambda_1 & 0 & 0 & . & . & 0 \\ 0 & \lambda_2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & & & . & & . \\ . & & & & & 0 \\ 0 & 0 & . & . & 0 & \lambda_p \end{bmatrix}$$

Et l'équation (6) devient:

$$\mathbf{Z=X.F} \tag{9}$$

Remarque: ici, le calcul des composantes principales s'est effectué sur des variables *centrées réduites* mais il est possible de le faire sur des variables *centrées* uniquement. La matrice de variance-covariance C est alors utilisée à la place de la matrice de corrélation R.

L'équation (6) est alors remplacée par: $$Z_k = \sum_{j=1}^{n}\left(X_j - M_j\right)u_{jk}$$

et l' équation (8) par: $$\mathbf{C.F=F.\Lambda}$$

Cette méthode donne aux variables un poids proportionnel à leur variance alors qu'en utilisant la matrice de corrélation elles ont toutes le même poids. Dans ce cas là, si 2 variables sont liées par une relation linéaire, elles deviennent identiques après avoir été centrées et réduites.

4. Propriétés des Composantes Principales

Les composantes principales sont orthogonales entre elles, donc décorrélées.

Les λ_k étant les variances des nouvelles variables Z_k, on peut normer ces composantes pour obtenir des CP centrées réduites:

$$Y_k = \frac{Z_k}{\sqrt{\lambda_k}} \quad \text{avec } \lambda_k = \frac{1}{n}\sum_{i=1}^{n}\left(z_{ik}\right)^2 = \sigma^2_{Z_k} \qquad \text{pour k=1 à p} \tag{10}$$

Les λ_k, qui sont aussi les valeurs propres de la matrice de corrélation R ont la propriété d'avoir une somme égale à la trace de R donc à p:

$$\sum_{i=1}^{p}\lambda_i = \text{trace}(R) = p \tag{11}$$

Cela permet de connaître la contribution de chacune des CP à la variance totale du système à p dimensions. On peut alors déterminer le nombre q, q<p, de CP "utiles" ou significatives que l'on doit conserver afin que la perte d'information soit faible par rapport à la variance totale (quelques %).

Dans le cas d'une ACP sur des variables centrées l'équation (11) devient:

$$\sum_{i=1}^{p} \lambda_i = \text{trace}(C) = \sum_{i=1}^{p} c_{ii} = \sum_{i=1}^{p} \sigma_i^2 \qquad (12)$$

ANNEXE II.2 :

Approximation numérique de l'ACPP

1. Rappels

Cette annexe fait référence aux théories énoncées au § II.1.3. Nous rappelons ici les équations importantes:

$$X(\xi,\underline{x}) = \sum_{k=1}^{\infty} Z_k(\xi).F_k(\underline{x}) \qquad \text{(II-2)}$$

$$\int_D C(\underline{x},\underline{x}').F_k(\underline{x}')d\underline{x}' = \lambda_k.F_k(\underline{x}) \qquad \text{(II-4)}$$

$$Z_k(\xi) = \int_D X(\xi,\underline{x}).F_k(\underline{x})d\underline{x} \qquad \text{(II-5)}$$

où D est le domaine considéré. Il peut être physiquement défini (un bassin océanique par exemple) ou arbitrairement choisi (une région).

2. Approximation numérique

Soit une base de fonctions $e_i(x)$ pour i=1 à P ayant la structure d'un espace vectoriel sur D. Le processus $X(\xi,\underline{x})$ est interpolé dans cette base.
Si l'on choisit les $\underline{x}_i$ comme étant les P points de mesure du domaine D, on peut avoir les $e_i(\underline{x})$ de façon canonique, c'est-à-dire telles que $e_i(\underline{x}_j) = \delta_{ij}$.

Le processus interpolé X* est donc obtenu de la manière suivante:

$$X^*(\xi,\underline{x}) = \sum_{i=1}^{P} X(\xi,\underline{x}_i).e_i(\underline{x}) \tag{12}$$

où les $X(\xi,\underline{x}_i)$ sont les valeurs de X mesurées aux P points de mesure.

La covariance C^* du processus interpolé devient:

$$C^*(\underline{x},\underline{x}') = \sum_{i=1}^{P}\sum_{j=1}^{P} C(\underline{x}_i,\underline{x}_j).e_i(\underline{x}).e_j(\underline{x}') \tag{13}$$

Les fonctions propres, solutions de l'équation (II-4), sont elles aussi approximées sur la base des $e_i(\underline{x})$:

$$F_k^*(\underline{x}) = \sum_{i=1}^{P} f_{ik}.e_i(\underline{x}) \tag{14}$$

où les f_{ik} pour i=1 à P sont les valeurs à déterminer de $F_k(\underline{x})$ au point $\underline{x}_j$.

Et l'équation (II-4) devient, en remplaçant les vraies valeurs de C et F par leurs valeurs interpolées C^* et F^*:

$$\int_D \left[\sum_{i=1}^{P}\sum_{j=1}^{P} C(\underline{x}_i,\underline{x}_j).e_i(\underline{x}).e_j(\underline{x}')\right].\left[\sum_{m=1}^{P} f_{mk}.e_m(\underline{x}')\right]d\underline{x}' = \lambda_k.\sum_{i=1}^{P} f_{ik}.e_i(\underline{x}) \tag{15}$$

soit,

$$\sum_{i=1}^{P}\left[\sum_{j=1}^{P}\sum_{m=1}^{P} C(\underline{x}_i,\underline{x}_j).f_{mk}.\int_D e_j(\underline{x}').e_m(\underline{x}')d\underline{x}' - \lambda_k.f_{ik}\right].e_i(\underline{x}) = 0 \tag{16}$$

Or les $e_i(\underline{x})$ forment une base donc les termes entre crochets sont nuls et l'équation (5) devient:

$$\sum_{j=1}^{P}\sum_{m=1}^{P} C(\underline{x}_i,\underline{x}_j).f_{mk}.E_{jm} = \lambda_k.f_{ik} \quad \text{pour i=1 à P} \tag{17}$$

avec $E_{jm} = \int_D e_j(\underline{x}').e_m(\underline{x}')d\underline{x}'$

Sous forme matricielle cela donne

$$\mathbf{C.E.F = F.\Lambda} \qquad (18)$$

avec - **C** matrice carrée symétrique des variances-covariances entre 2 points du réseau:

$$C = \begin{bmatrix} c_{11} & c_{12} & c_{13} & . & . & c_{1p} \\ c_{12} & c_{22} & c_{23} & . & . & c_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ . & . & . & . & . & c_{pp} \end{bmatrix} \quad \text{où } c_{ij}=C(\underline{x}_i,\underline{x}_j)=C(\underline{x}_j,\underline{x}_i)=c_{ji},$$

- **E** matrice carrée symétrique des produits scalaires construits à partir des $e_i(\underline{x})$:

$$E = \begin{bmatrix} E_{11} & E_{12} & E_{13} & . & . & E_{1p} \\ . & E_{22} & E_{23} & . & . & E_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ . & . & . & . & . & E_{pp} \end{bmatrix}$$

- **F** matrice carrée des valeurs des fonctions propres aux points de mesure:

$$F = \begin{bmatrix} f_{11} & f_{12} & f_{13} & . & . & f_{1p} \\ . & f_{22} & f_{23} & . & . & f_{2p} \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ f_{p1} & . & . & . & . & f_{pp} \end{bmatrix}$$

$F_1 \quad F_2 \qquad\qquad F_p$

- et $\mathbf{\Lambda}$, matrice diagonale des valeurs propres λ_k:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & . & . & 0 \\ 0 & \lambda_2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & & & . & & . \\ . & & & & & 0 \\ 0 & 0 & . & . & 0 & \lambda_p \end{bmatrix}$$

Estimation des Composantes Principales $Z_k(\xi)$:

En utilisant la décomposition sur la base de fonction $e_i(x)$ l'équation (II-5) devient:

$$Z_k(\xi) = \int_D \sum_{i=1}^{P} X(\xi, \underline{x}_i).e_i(\underline{x}). \sum_{j=1}^{P} f_{jk}.e_j(\underline{x})d\underline{x}$$

puis,
$$Z_k(\xi) = \sum_{i=1}^{P}\sum_{j=1}^{P} X(\xi, \underline{x}_i).\int_D e_i(\underline{x}).e_j(\underline{x})d\underline{x}.f_{jk}$$

et
$$Z_k(\xi) = \sum_{i=1}^{P}\sum_{j=1}^{P} X(\xi, \underline{x}_i).E_{ij}.f_{jk} \qquad (19)$$

Soit

$$\mathbf{Z = X.E.F} \qquad (20)$$

où - **X** est la matrice des N réalisations du processus $X(\xi,\underline{x})$:

$$\left[X_{np}\right] = \begin{bmatrix} x_{11} & x_{12} & x_{13} & . & . & x_{1p} \\ x_{21} & x_{22} & x_{23} & . & . & x_{2p} \\ . & . & . & . & . & . \\ . & & & & & . \\ . & & & & & . \\ x_{n1} & x_{n2} & . & . & . & x_{np} \end{bmatrix}$$
$$X_1 \quad X_2 \qquad\qquad X_p$$

et - **Z** la matrice des composantes principales:

$$\left[Z_{np}\right] = \begin{bmatrix} z_{11} & z_{12} & z_{13} & . & . & z_{1p} \\ z_{21} & z_{22} & z_{23} & . & . & z_{2p} \\ . & . & . & . & . & . \\ . & & & & & . \\ . & & & & & . \\ z_{n1} & z_{n2} & . & . & . & z_{np} \end{bmatrix}$$

$$Z_1 \quad Z_2 \qquad\qquad Z_p$$

3. Calcul matriciel

Le calcul des CP se fait donc grâce à un calcul matriciel relativement simple. En effet, la matrice étant symétrique, elle peut s'écrire, grâce à une décomposition de Choleski, sous la forme:

$$E = Q\,Q^T$$

où Q est une matrice triangulaire inférieure.

Puis, si l'on multiplie à gauche chaque membre de l'équation (18) par Q^T, on obtient:

$$Q^T C(QQ^T)F = Q^T F\Lambda$$

soit, en posant $V = Q^T F$:

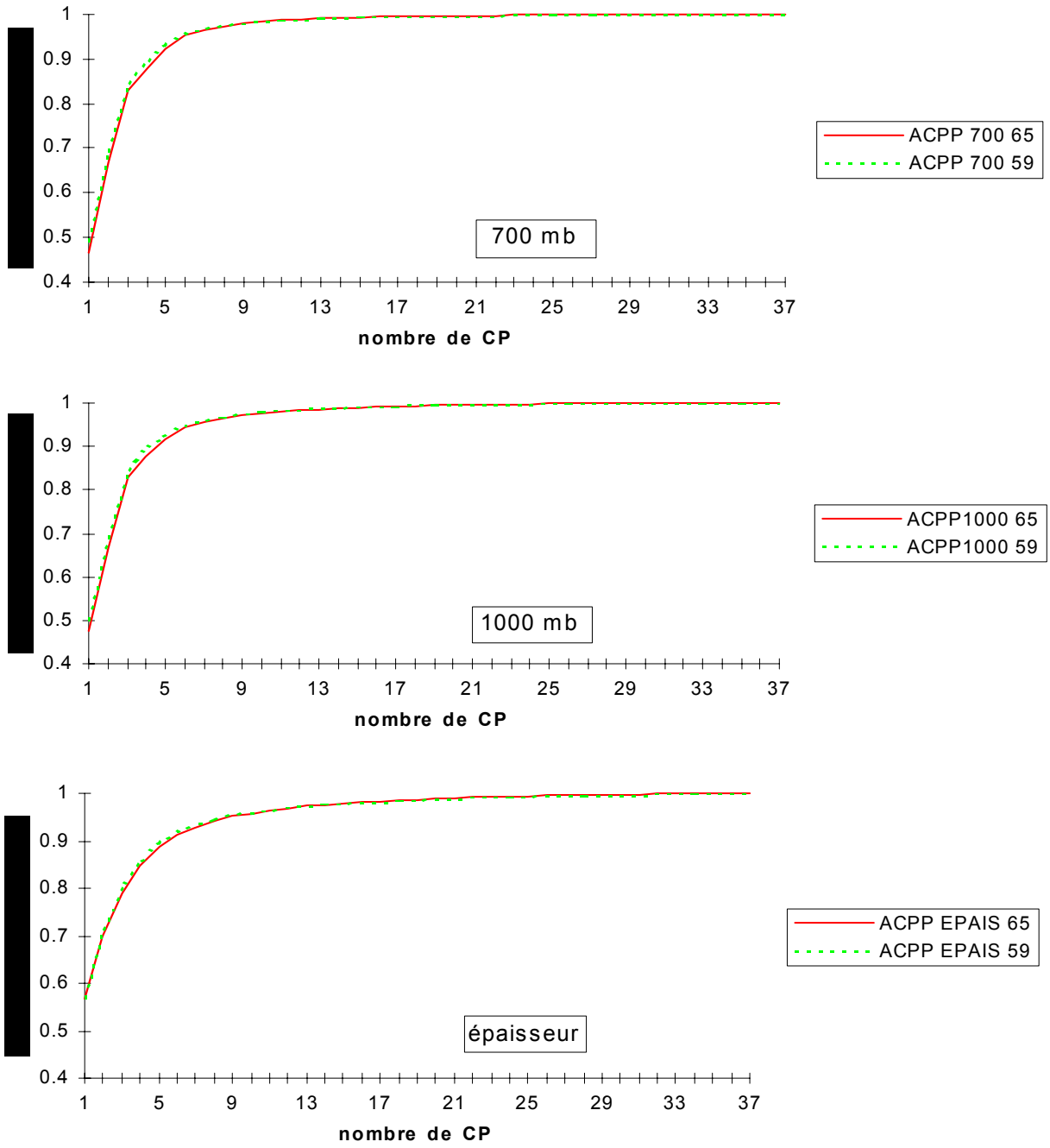
$$(Q^T CQ)V = V\Lambda$$

Et $Q^T CQ$ étant symétrique puisque C l'est aussi, on calcule ses vecteurs propres V de manière classique et on en déduit:

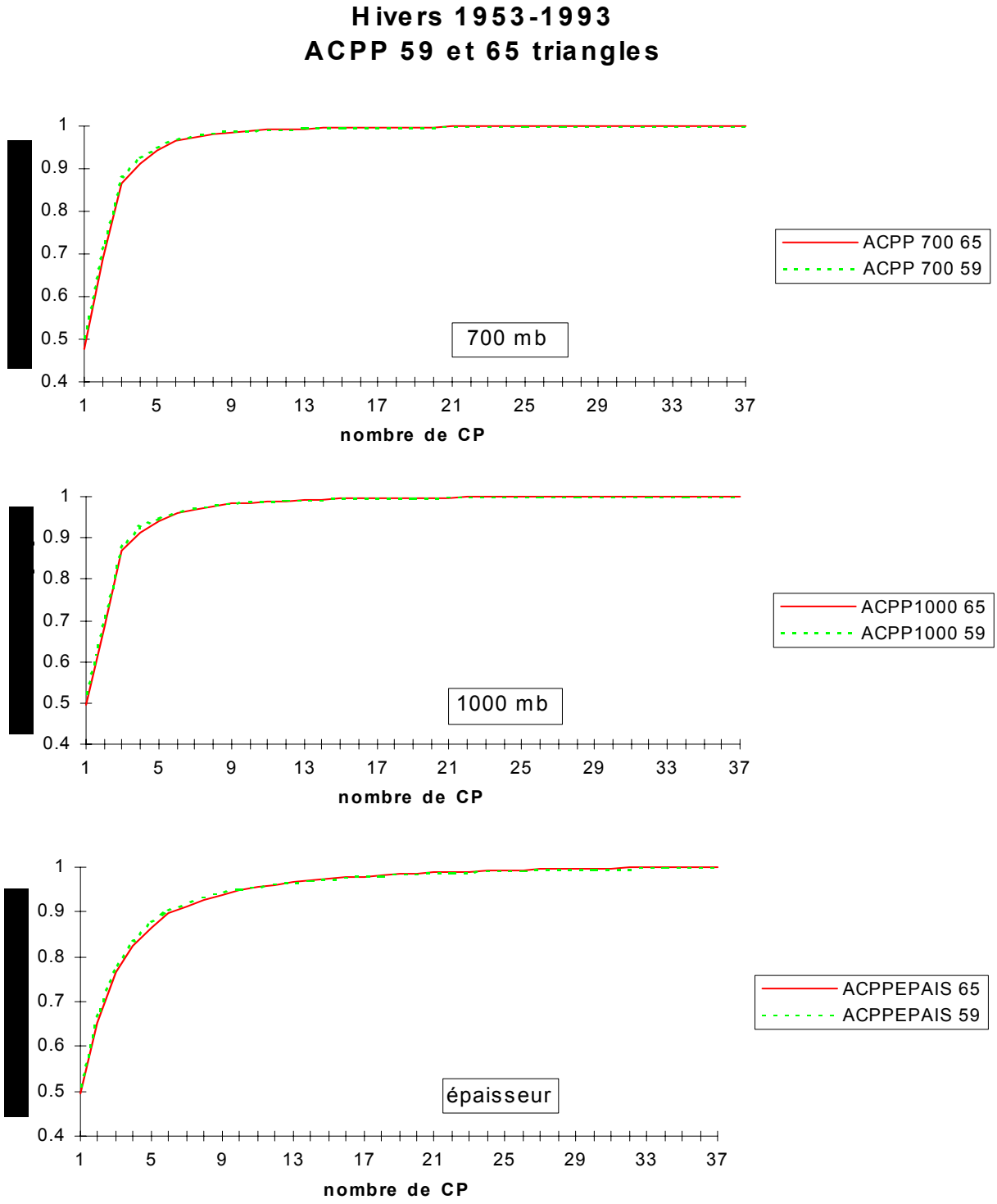
$$F = (Q^T)^{-1}\,V$$

ANNEXE II-3:

a) Choix du domaine - 59 ou 65 triangles - pour l'été



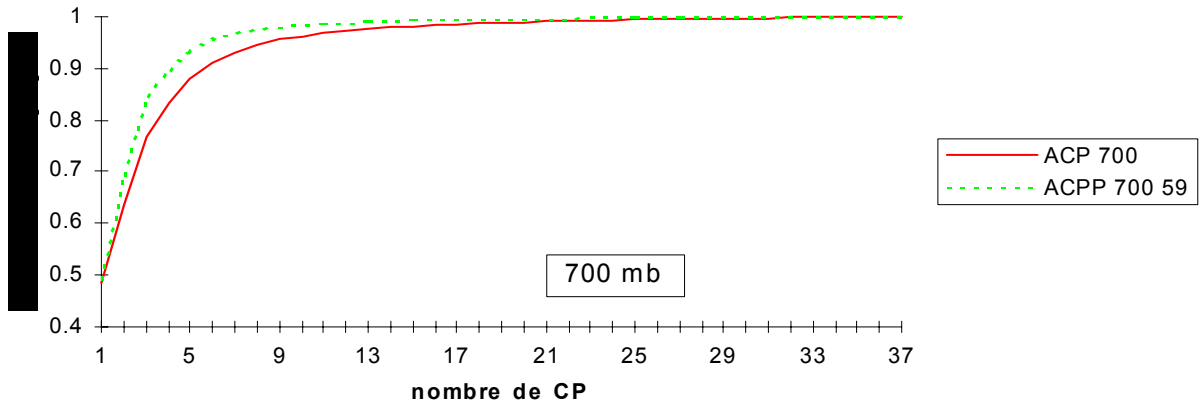
b) Choix du domaine - 59 ou 65 triangles - pour l'hiver

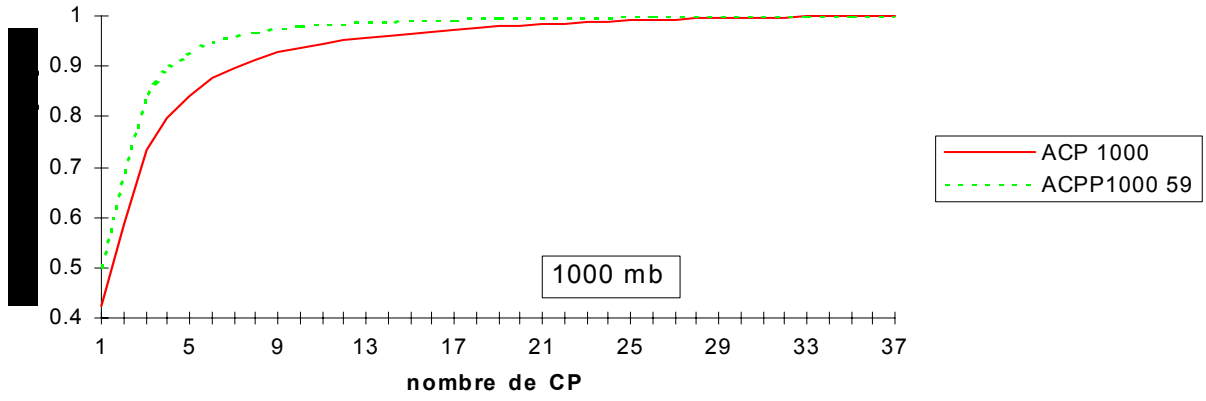


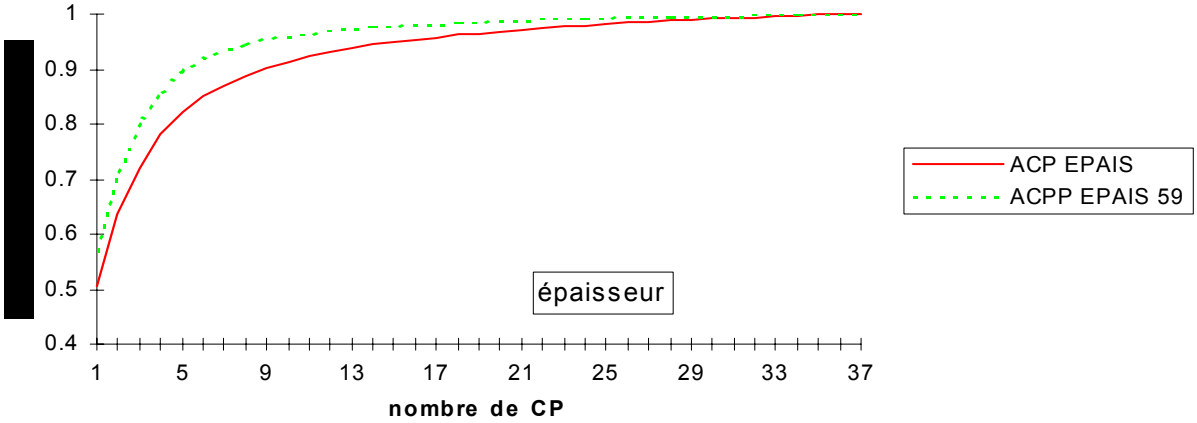
ANNEXE II-4:

a) Comparaison ACP / ACPP pour l'été

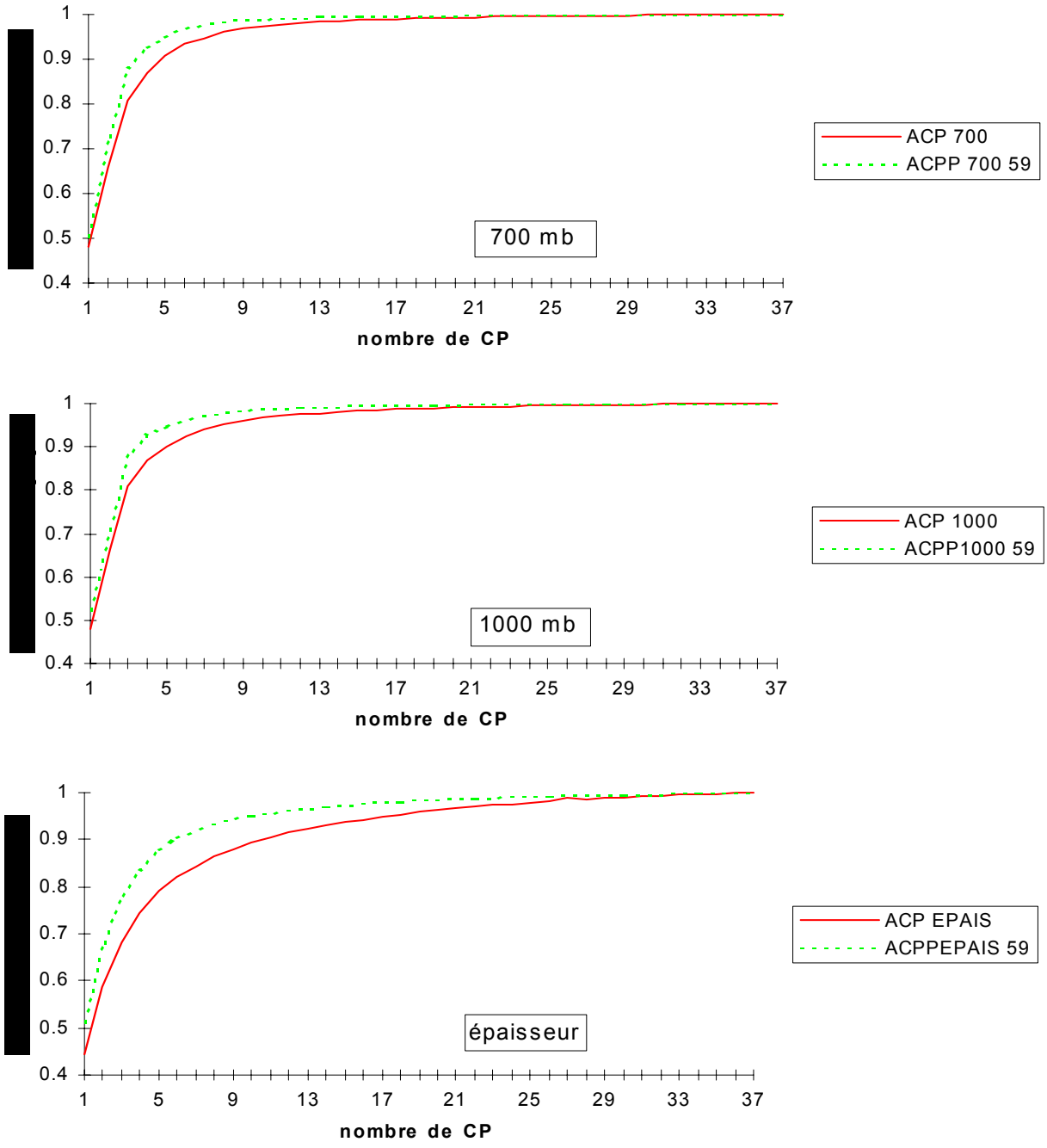
Etés 1953-1993
ACP et ACPP 59 triangles





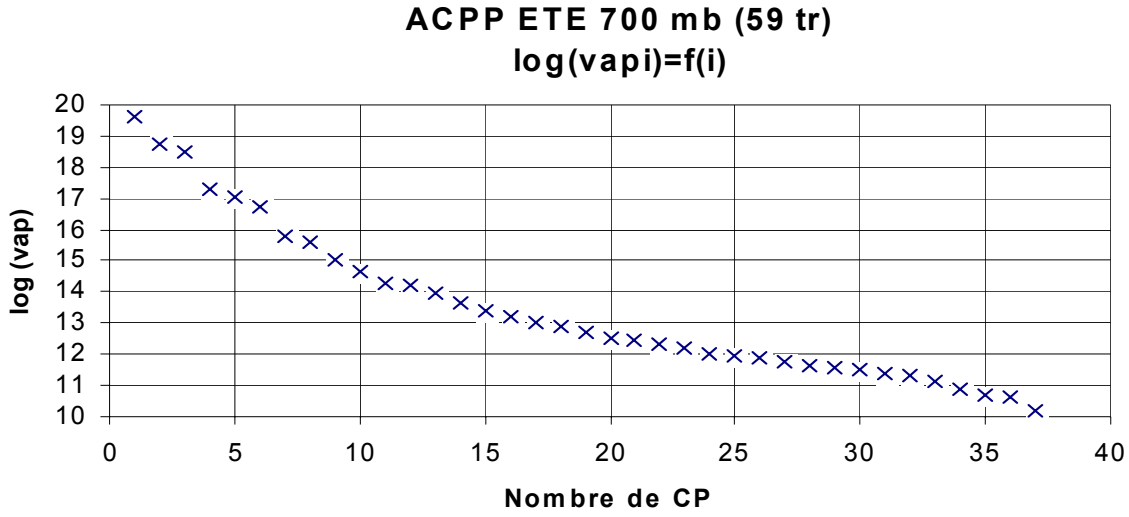


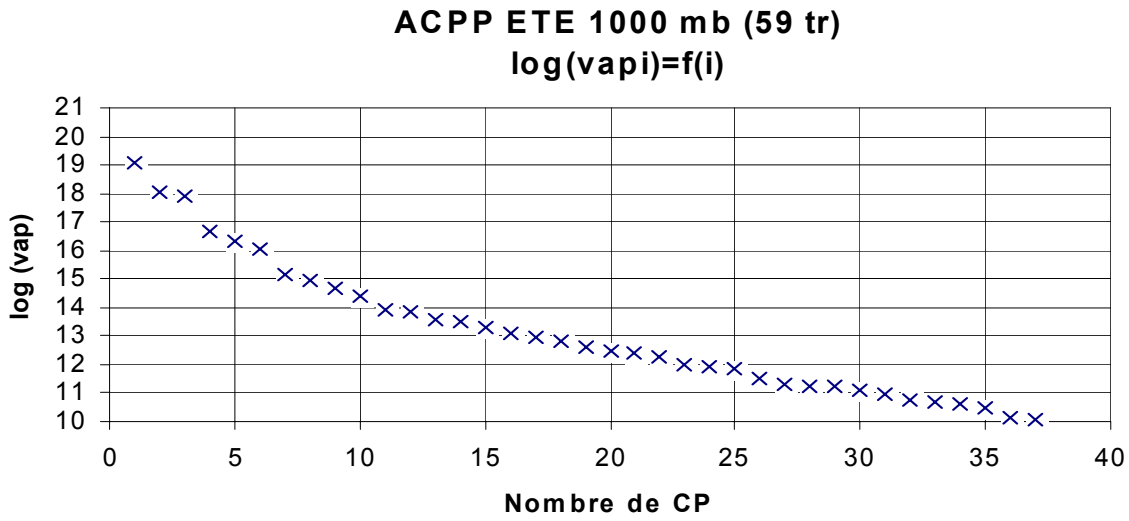
b) Comparaison ACP / ACPP pour l'hiver

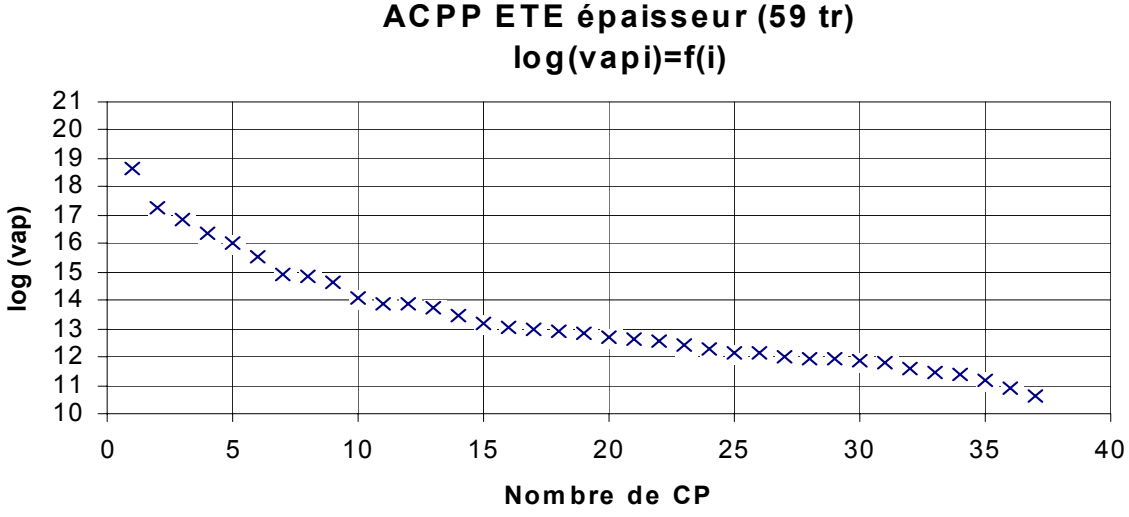


ANNEXE II-5:

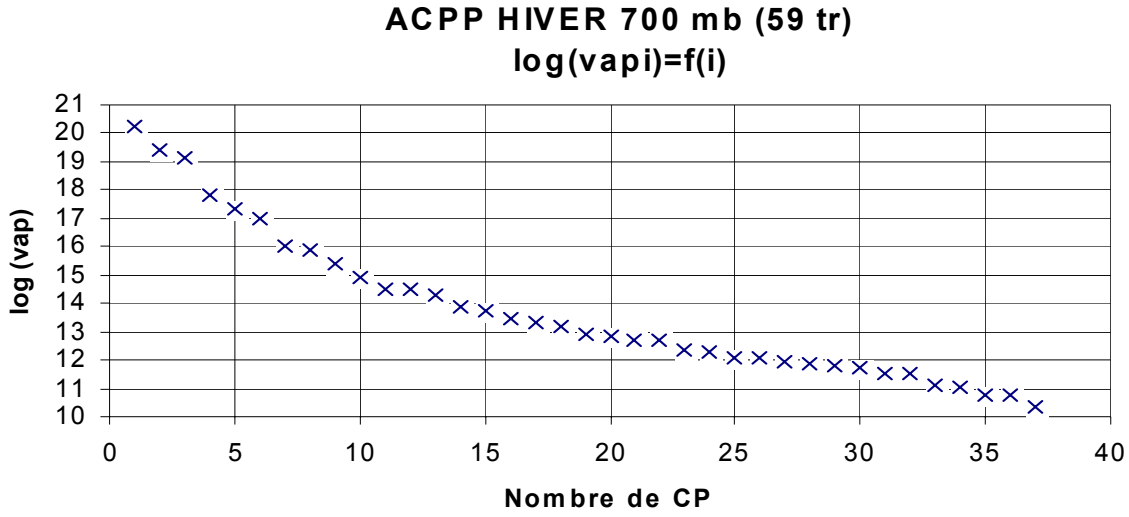
a) Méthode LEV pour l'été

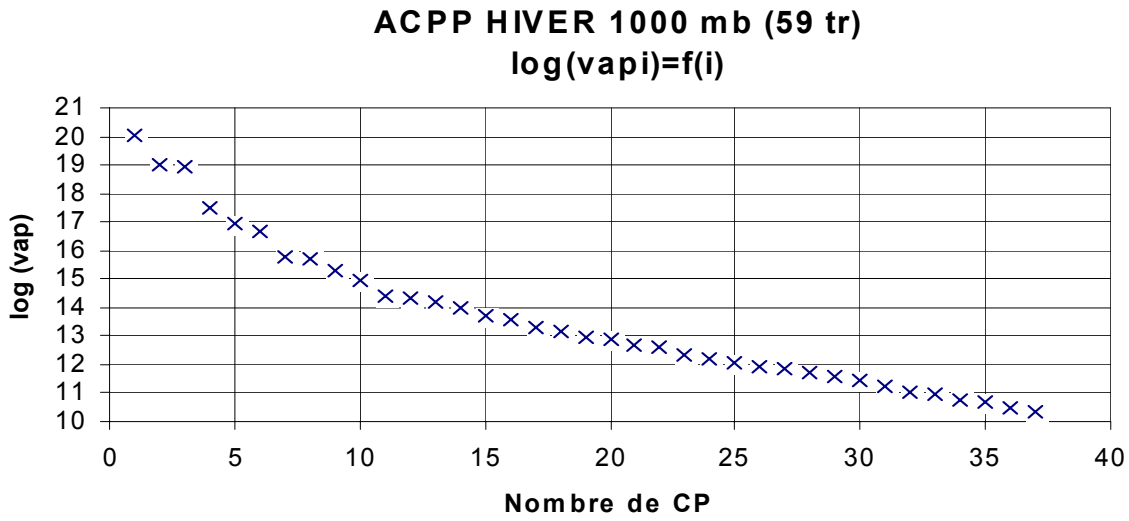


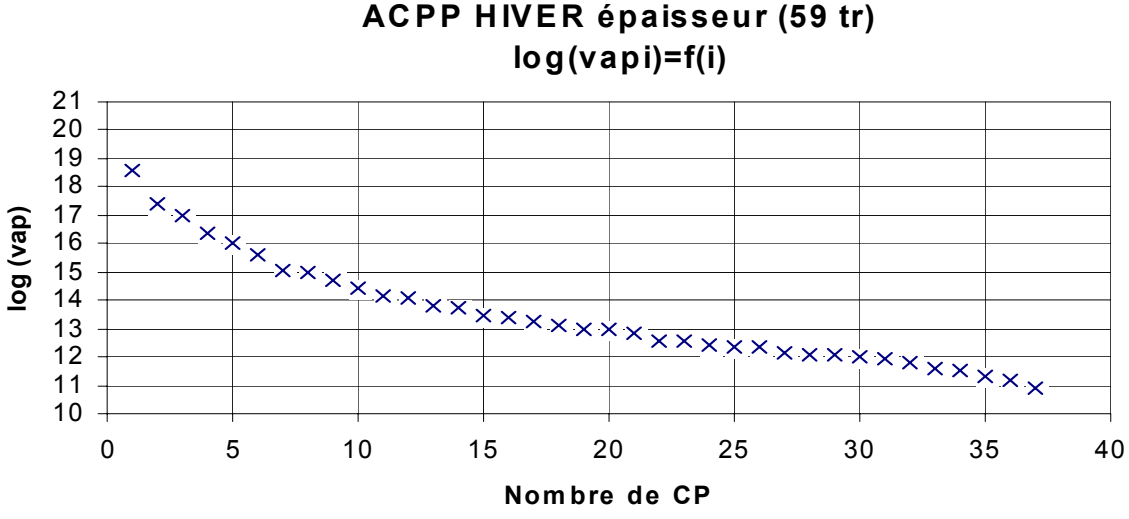




b) Méthode LEV pour l'hiver



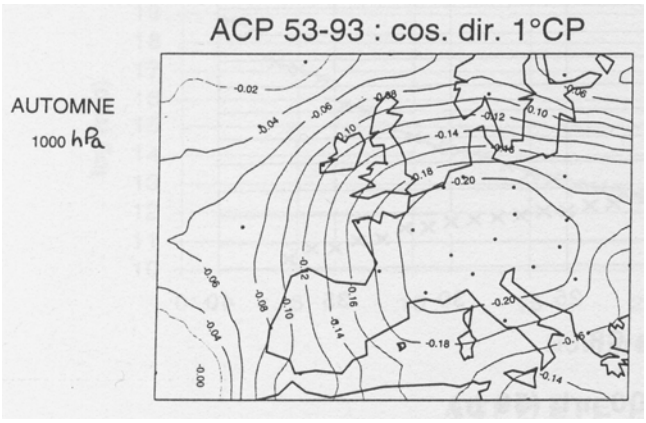


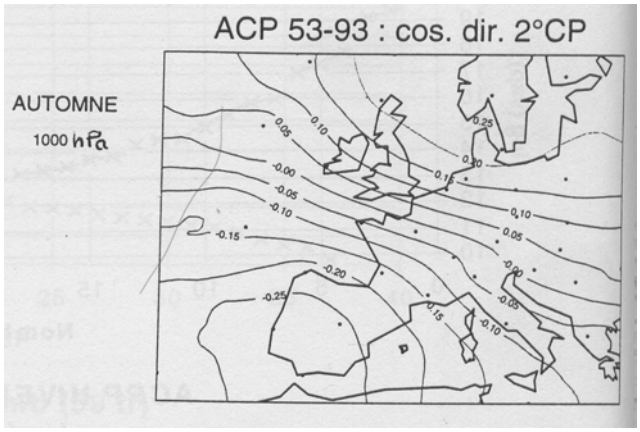


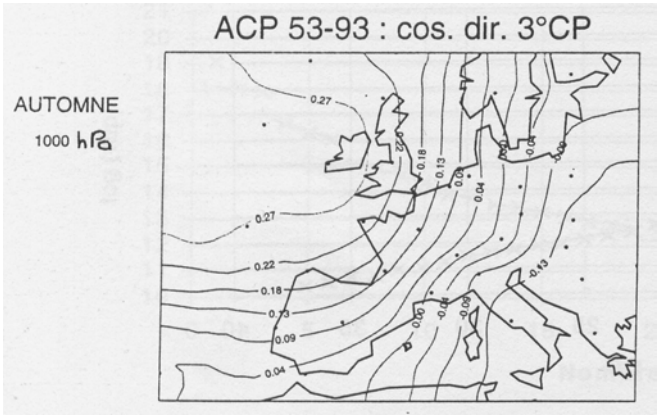
ANNEXE II-6:

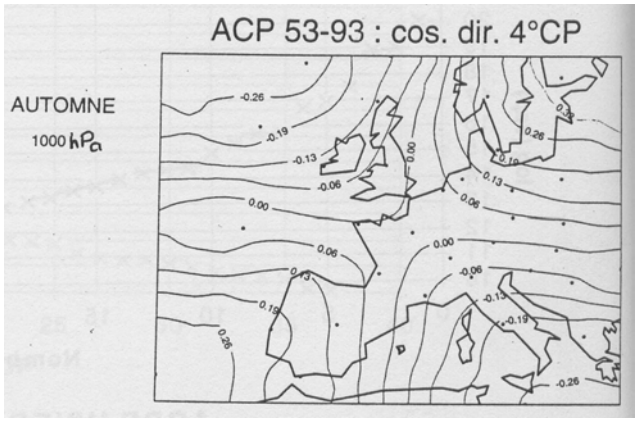
a) Cartes des 6 1ers cosinus directeurs de l'ACP :

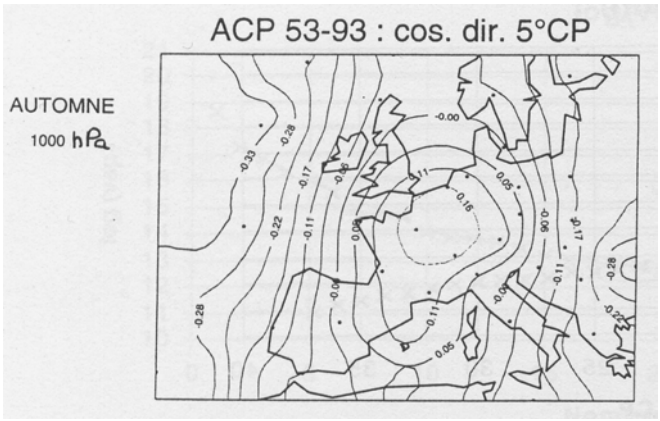
Géopotentiel 1000 hPa, automne

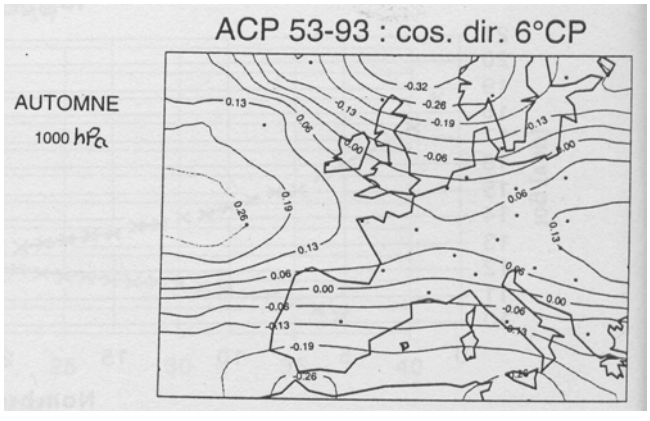






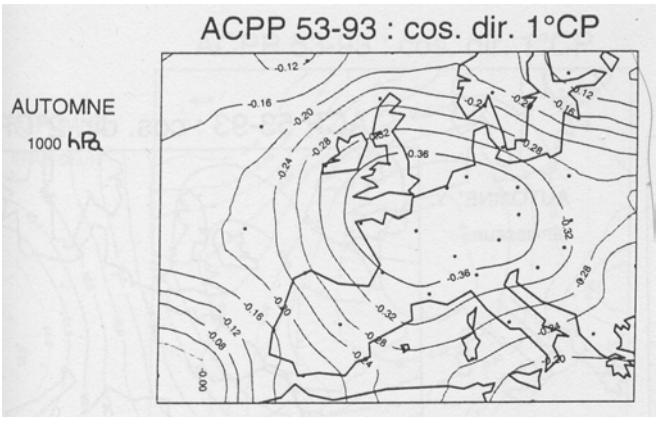


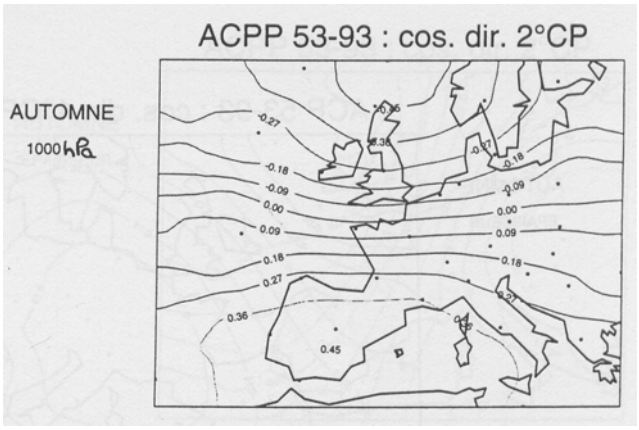


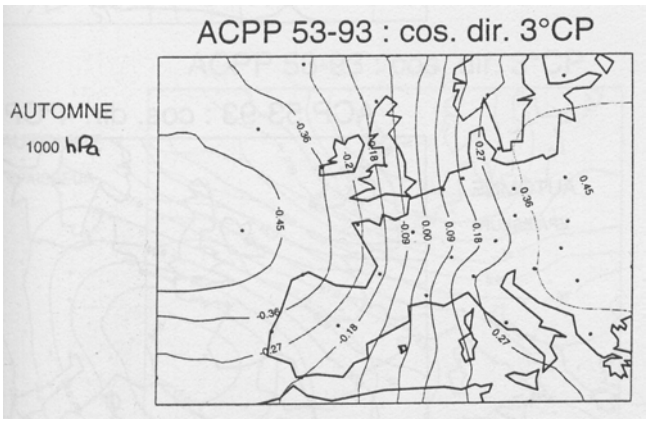


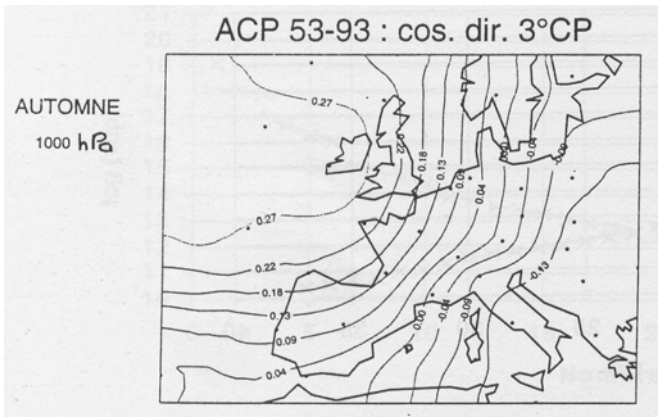
b) Cartes des 6 1ers cosinus directeurs de l'ACPP

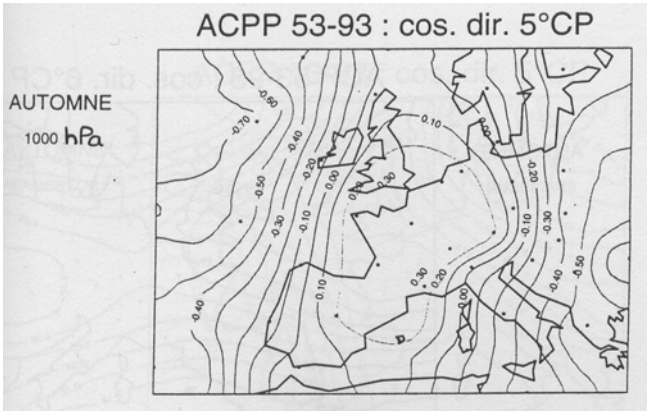
Géopotentiel 1000 hPa, automne

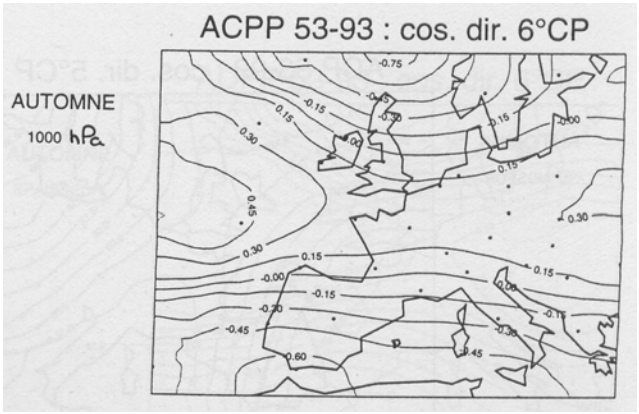






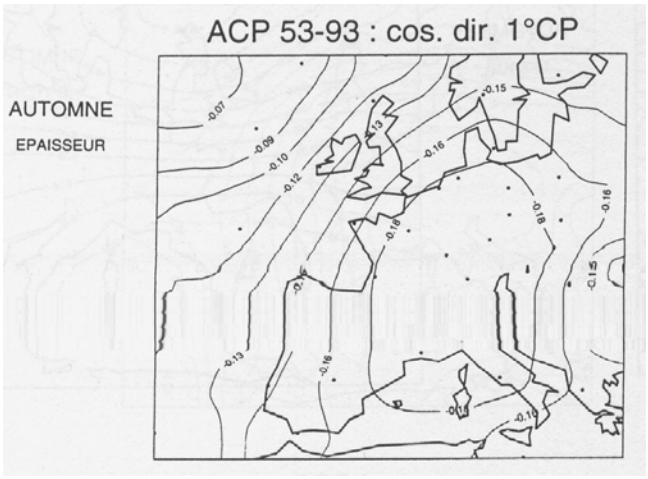


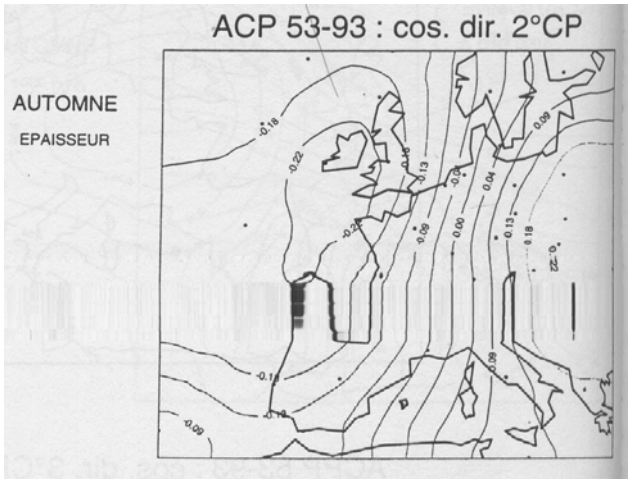


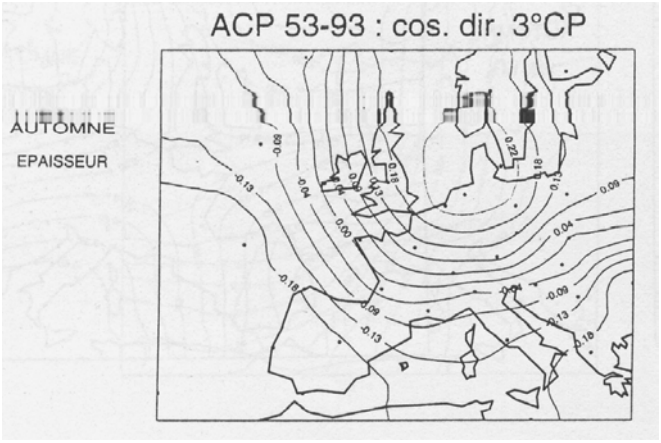


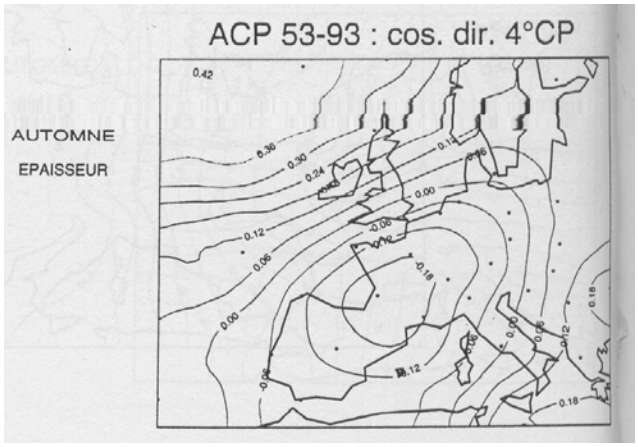
c) Cartes des 6 1ers cosinus directeurs de l'ACP

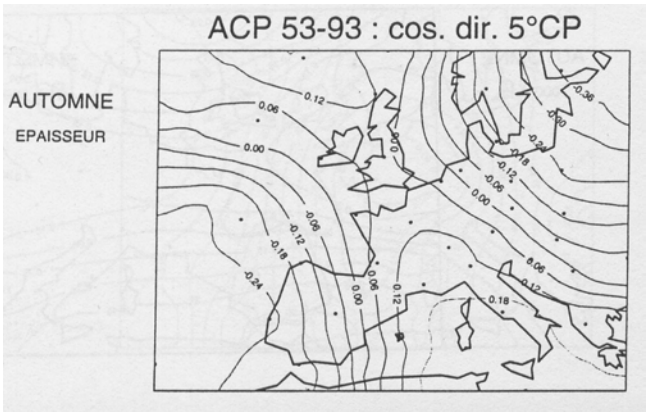
Epaisseur 700/1000 hPa, automne

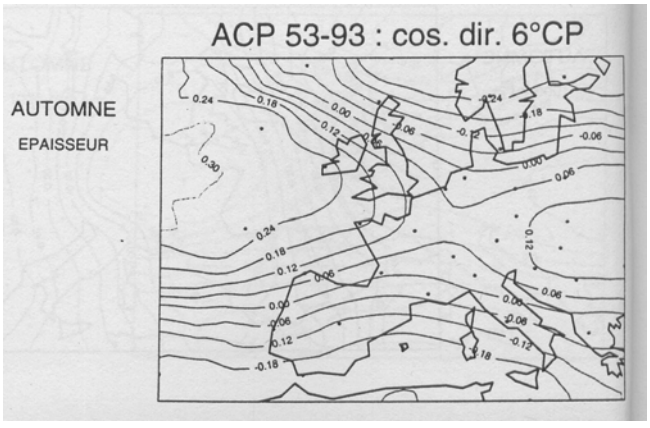






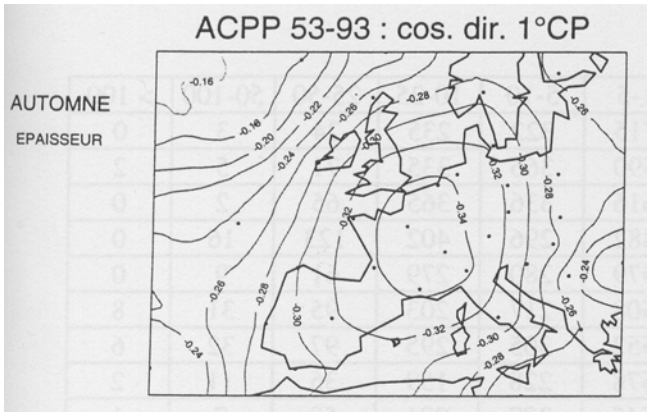


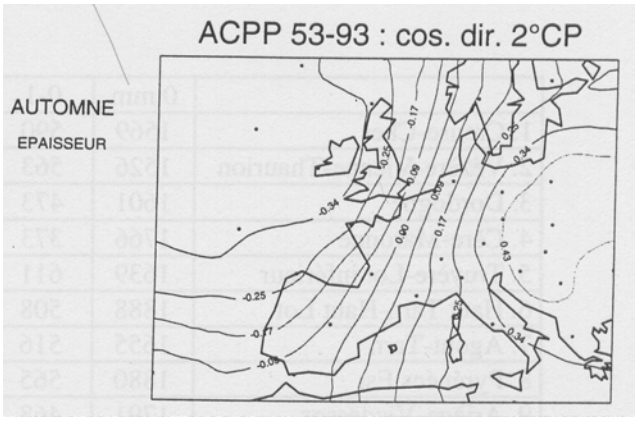


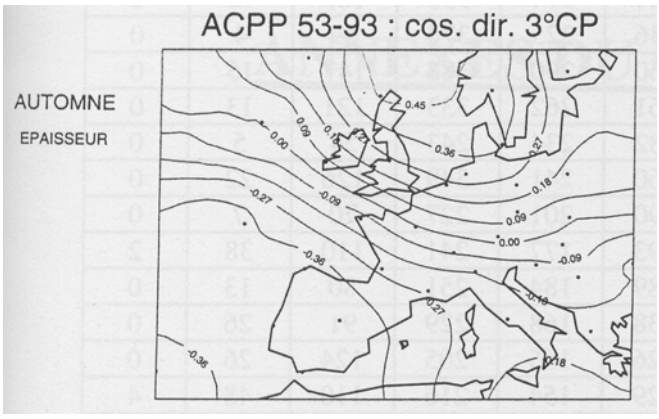


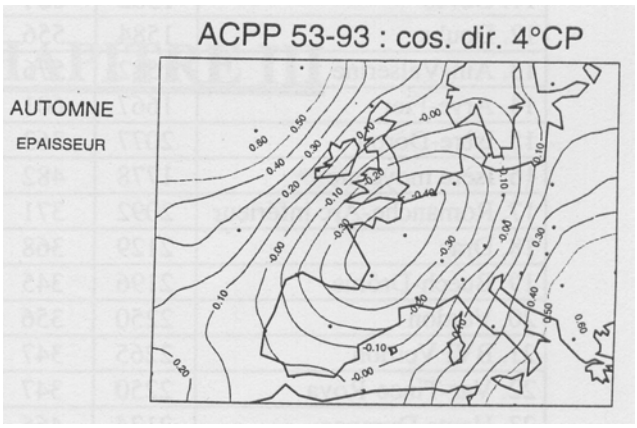
d) Cartes des 6 1ers cosinus directeurs de l'ACPP

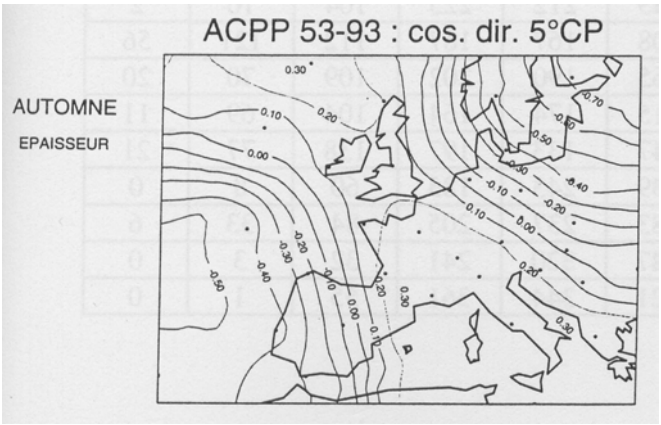
Epaisseur 700/1000 hPa, automne

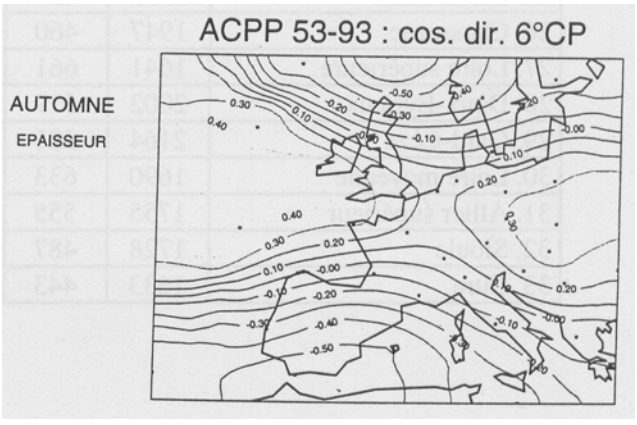












ANNEXE II-7 :

Climatologie des 33 groupements

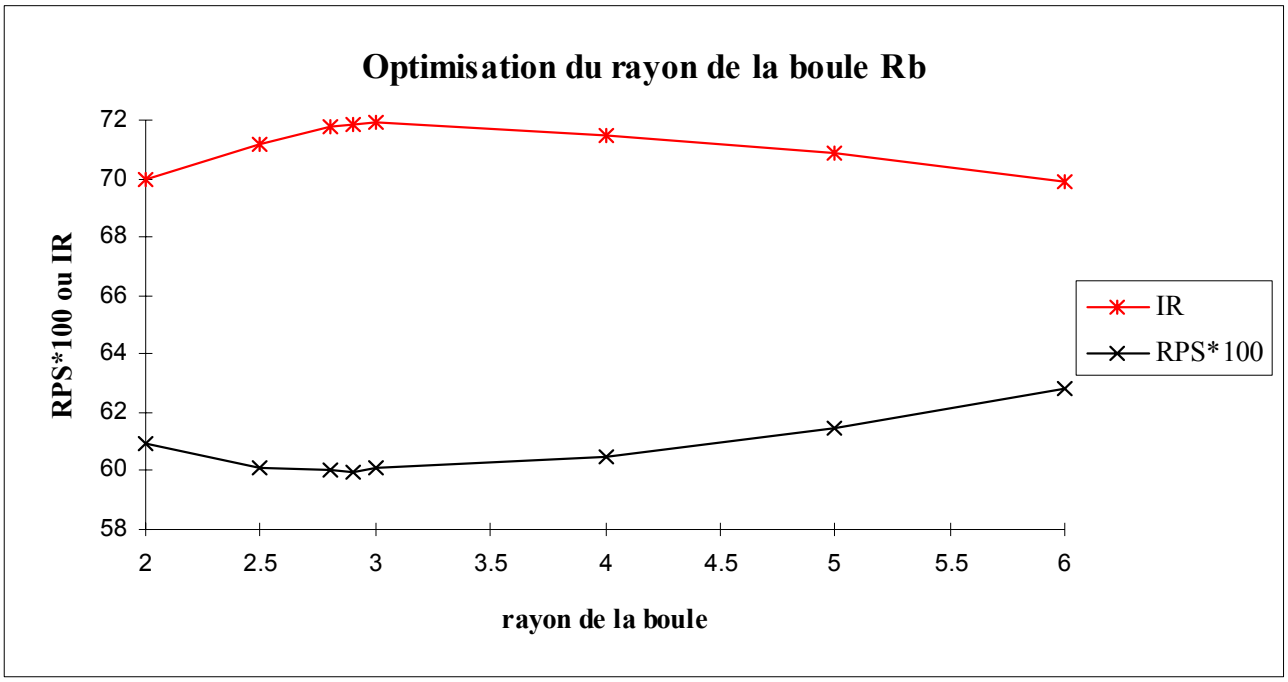
(nombre de jours par classe de pluie pour les automnes de 1953 à 1993, soit 3731 observations)

	0 mm	0-1	1-5	5-10	10-25	25-50	50-100	> 100
1. Creuse-Cher	1569	590	715	322	235	24	3	0
2. Vézère-Vienne-Thaurion	1526	563	590	366	335	71	5	2
3. Dordogne	1601	473	616	336	365	65	2	0
4. Cère-Maronne	1766	373	482	296	402	123	16	0
5. Truyère-Lot inférieur	1639	611	579	280	279	61	9	0
6. Haut Tarn-Haut Lot	1888	508	508	217	203	95	31	8
7. Agout-Tarn	1655	516	552	305	295	97	32	6
8. Pyrénées Est	1880	565	578	228	159	35	11	2
9. Ariège-Vicdessos	1791	468	545	307	281	58	7	1
10. Pique-Garonne-Salat	1577	579	556	324	326	86	10	0
11. Gaves	1382	684	624	314	338	106	10	0
12. Doubs	1584	556	546	325	374	69	4	0
13. Ain-Valserine	1582	576	460	290	388	147	15	0
14. Arve-Fier	1667	601	461	262	333	121	13	0
15. Isère-Doron	2077	363	482	234	243	54	5	0
16. Isère moyenne	1778	482	460	241	349	126	22	0
17. Romanche-Arc inférieur	2092	371	490	201	227	70	7	0
18. Drac	2129	368	393	177	241	110	38	2
19. Buech-Drôme	2196	345	389	184	251	80	13	0
20. Verdon	2250	356	338	168	229	91	26	0
21. BVI Verdon	2265	347	326	165	205	124	26	0
22. Var-Tinee-Roya	2250	347	329	154	216	110	48	4
23. Haute Durance	2124	466	394	182	206	74	12	0
24. Durance moyenne	2325	315	319	171	235	76	17	0
25. Mont Cenis	1999	457	443	212	225	104	16	2
26. Chassezac	1947	460	408	167	187	112	121	56
27. Loire supérieure	1641	661	565	190	202	109	70	20
28. Doux-Eyrieux	2002	519	415	174	164	104	69	11
29. Gard-Cèze	2164	381	347	133	197	138	77	21
30. Loire moyenne	1690	633	639	243	185	60	8	0
31. Allier supérieur	1755	555	583	237	205	84	33	6
32. Sioule	1728	487	647	320	241	32	3	0
33. Cure	1633	443	621	344	361	55	1	0

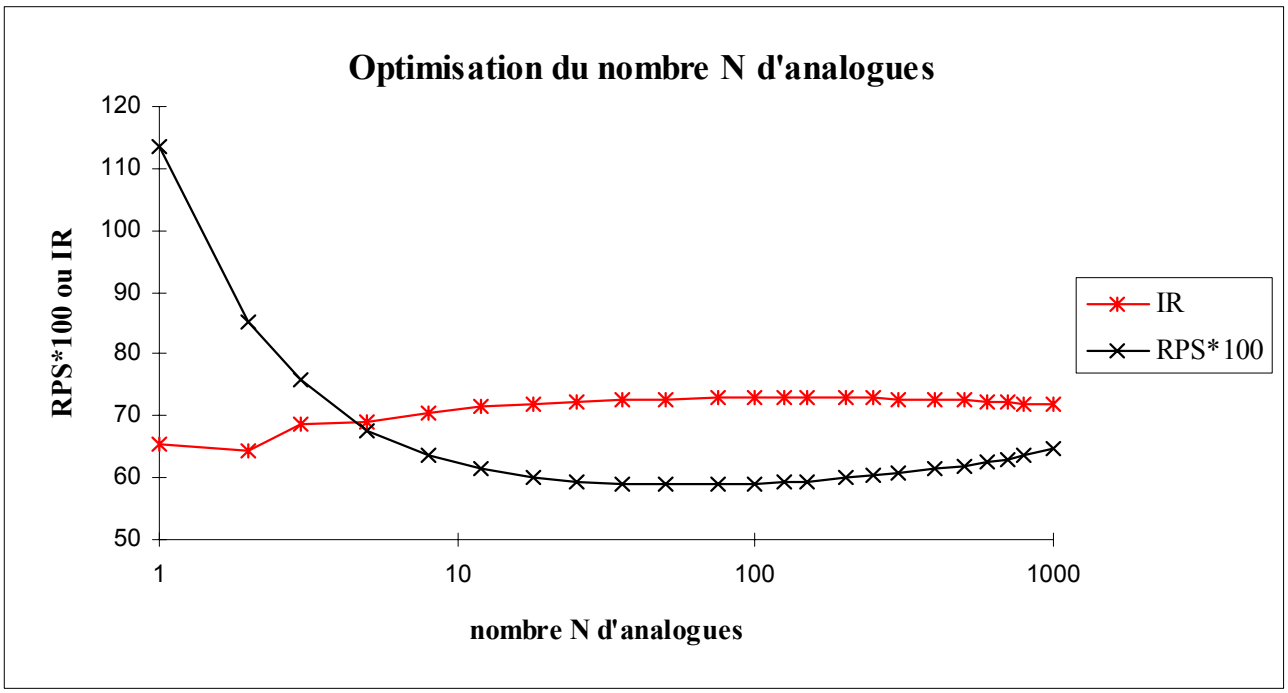
ANNEXES DU CHAPITRE III

ANNEXE III-1:

a) Optimisation du rayon de la boule Rb



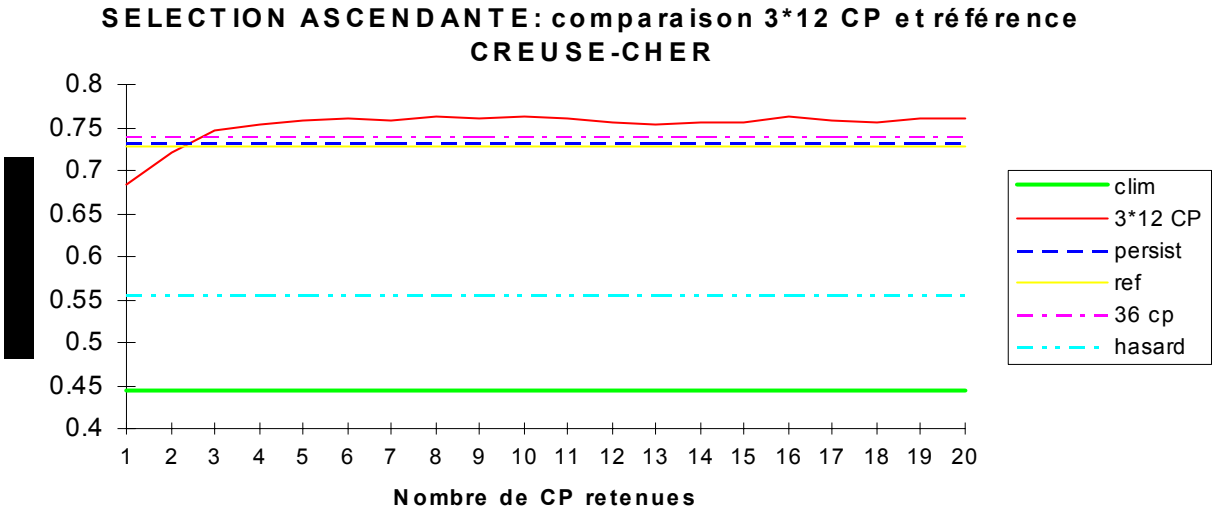
b) Optimisation du nombre N d'analogue

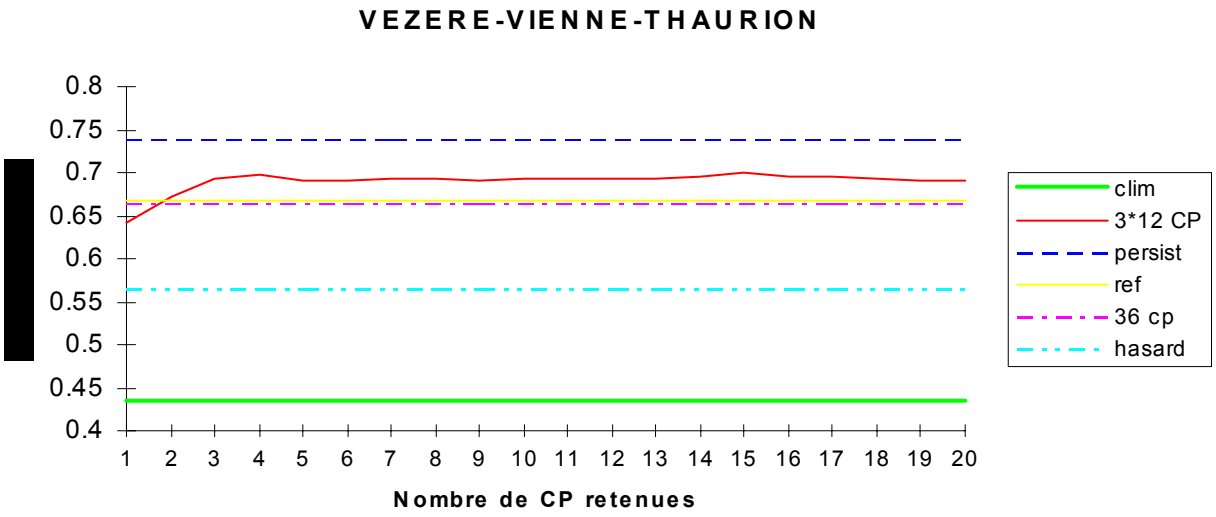


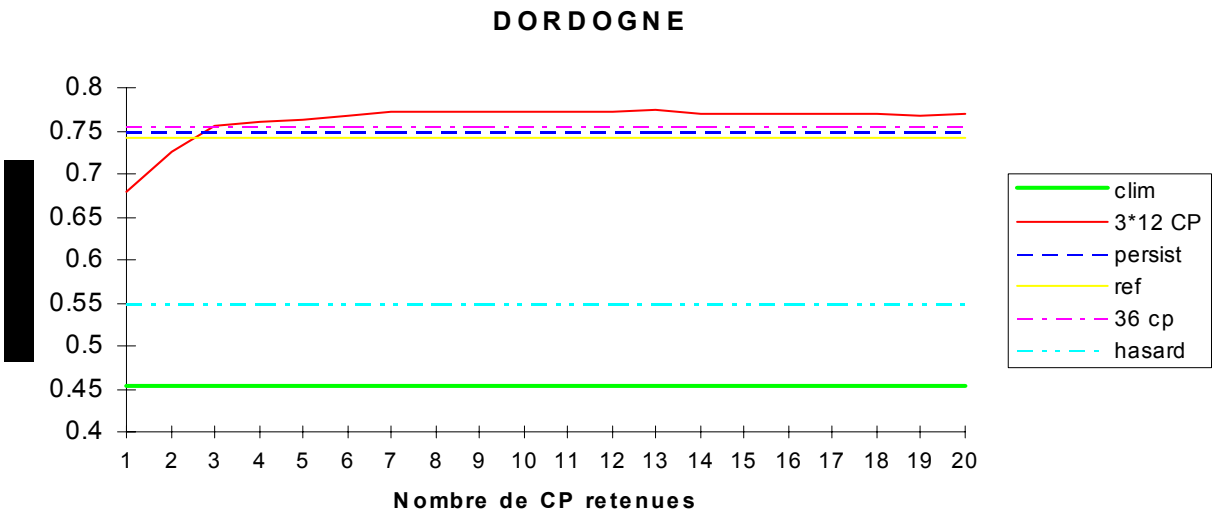
ANNEXE III-2 :

Sélection ascendante k=1 avec 36 CP

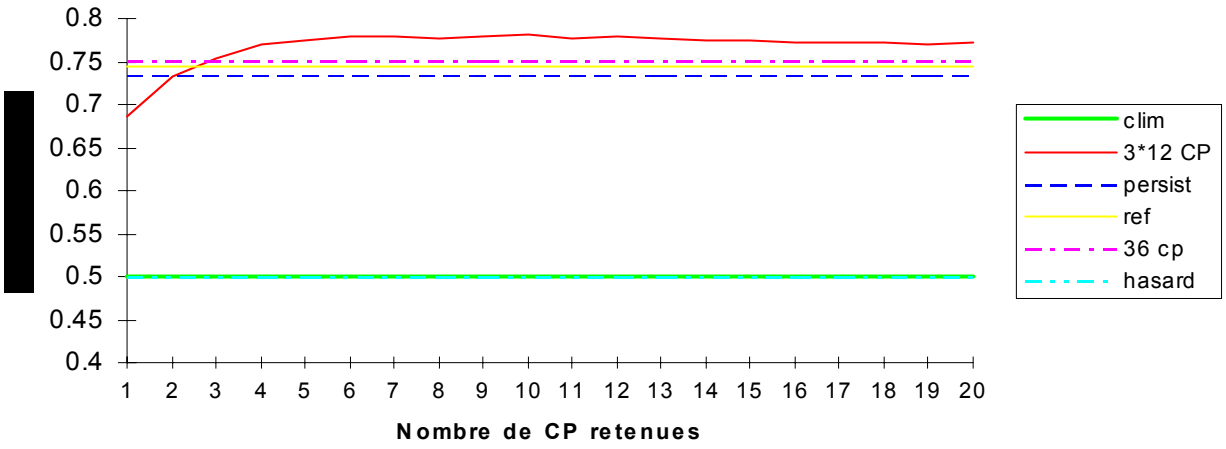
(comparaison avec les méthodes de référence)



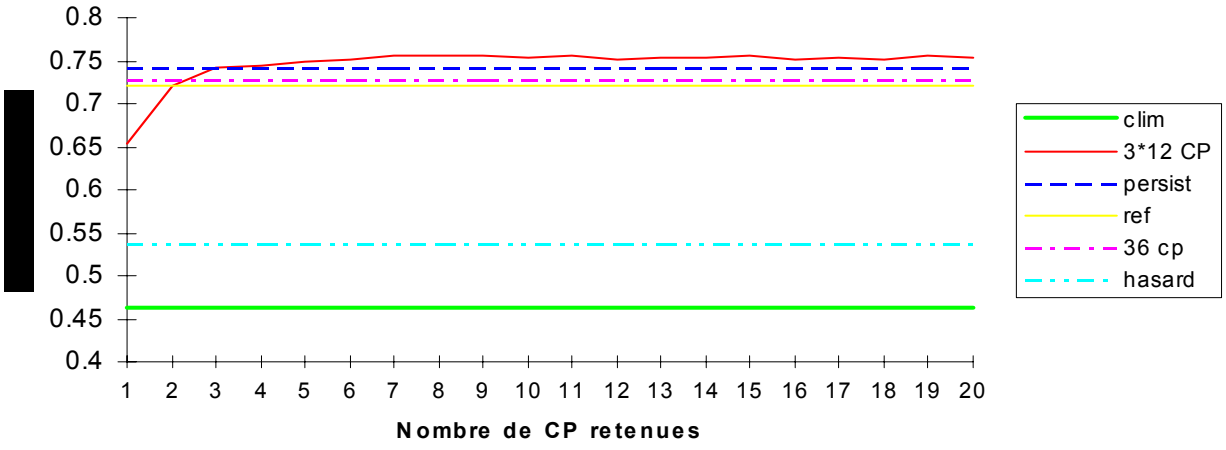




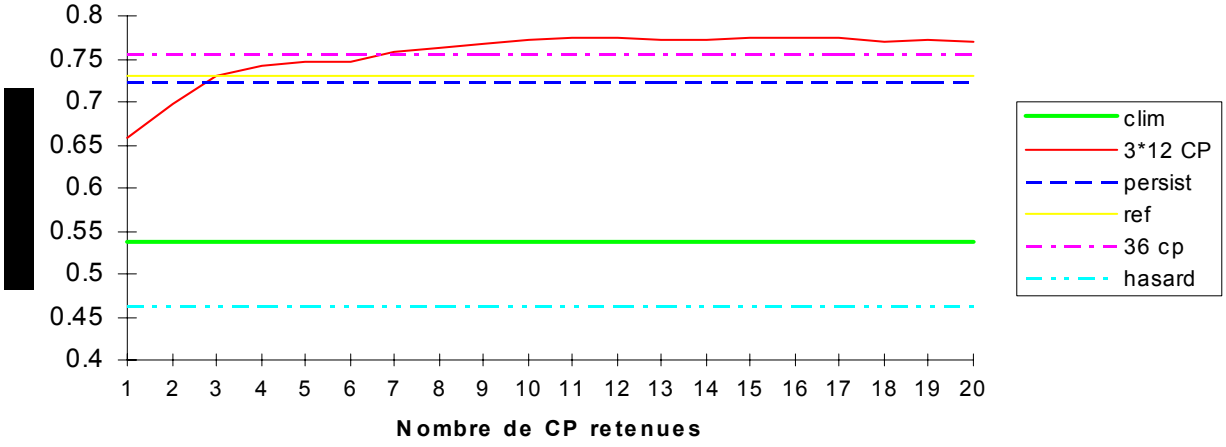
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
CERE-MARONNE



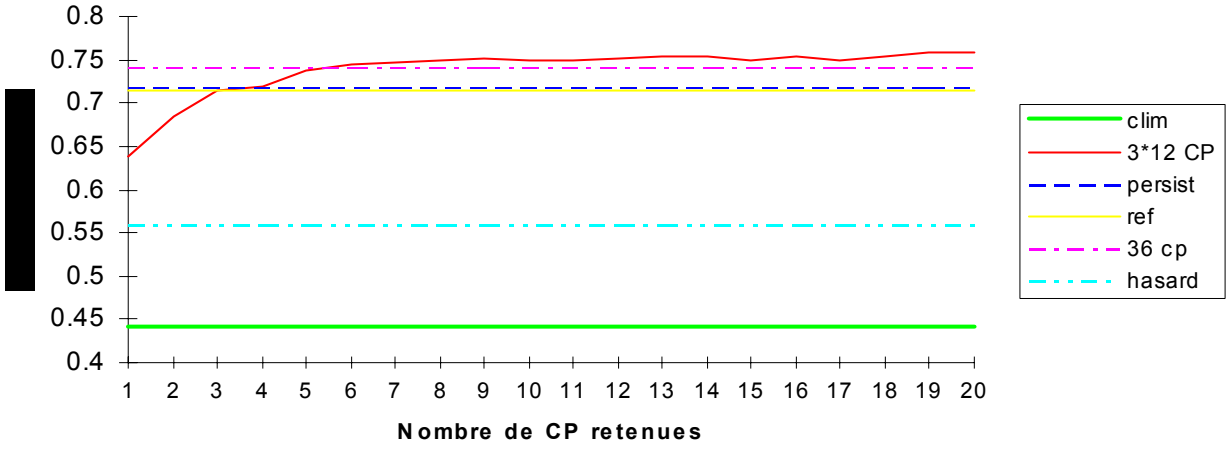
TRUYERE-LOT INFERIEUR



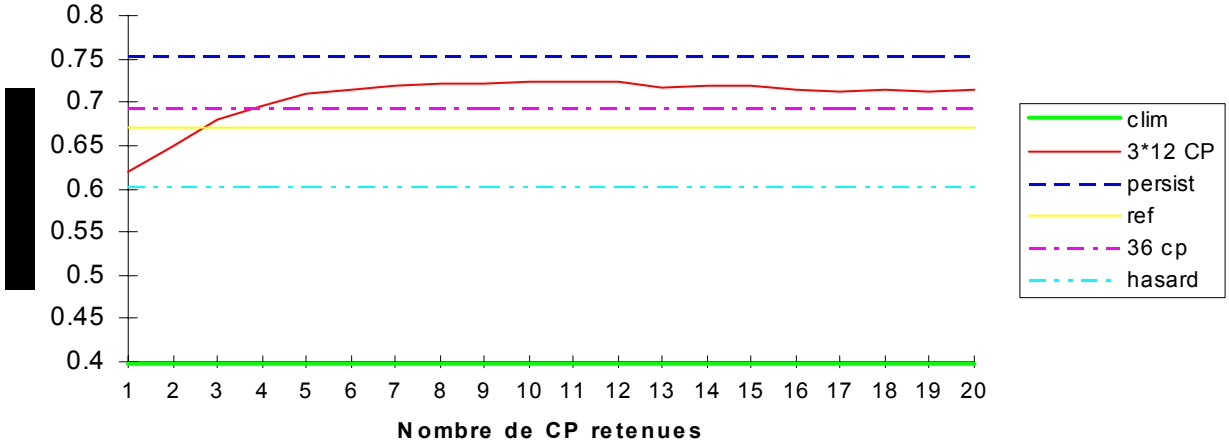
HAUT TARN-HAUT LOT



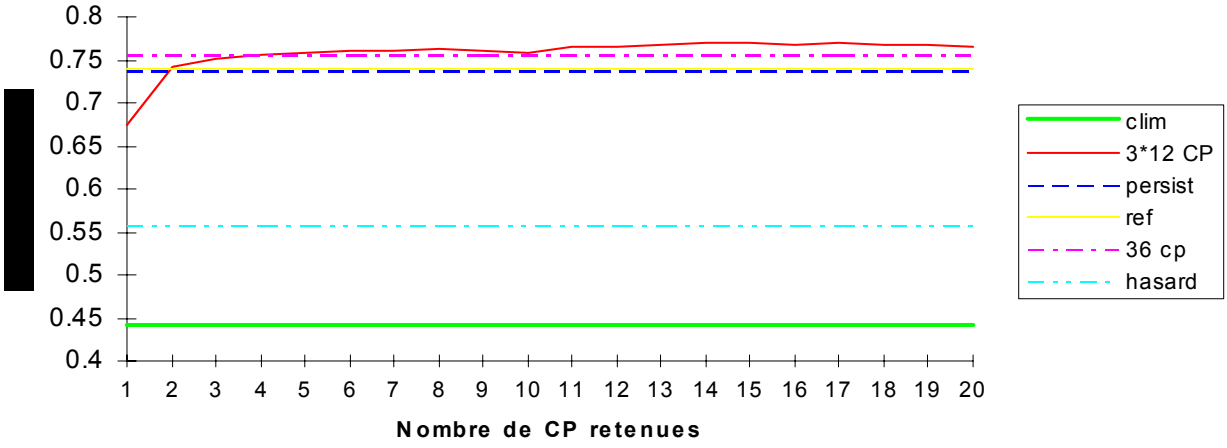
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
PIQUE-GARONNE-SALAT



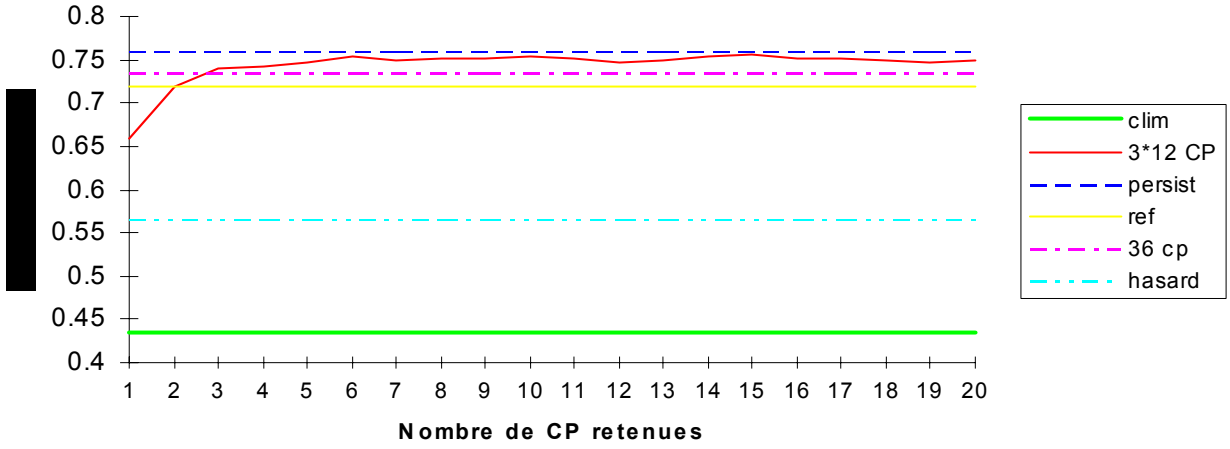
GAVES



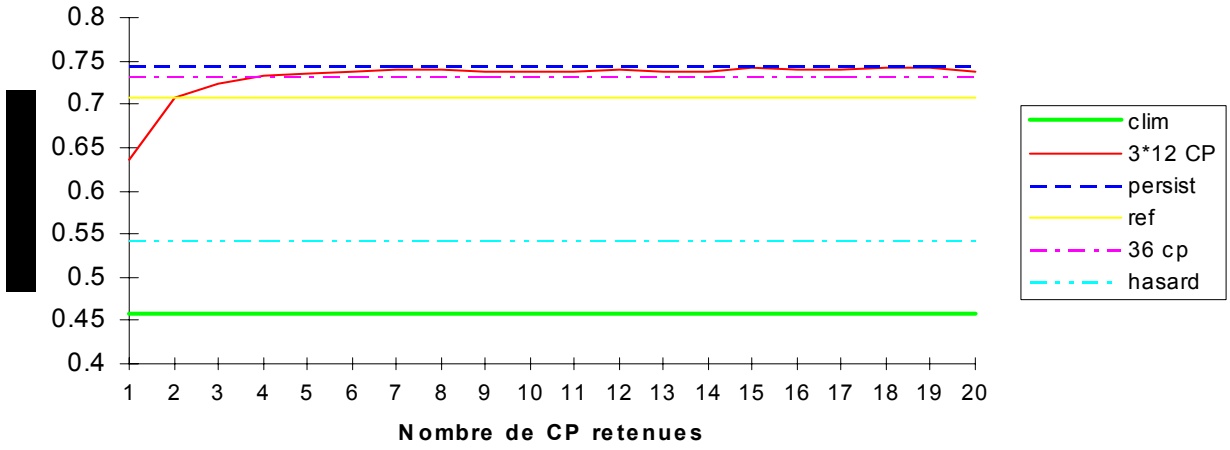
DOUBS



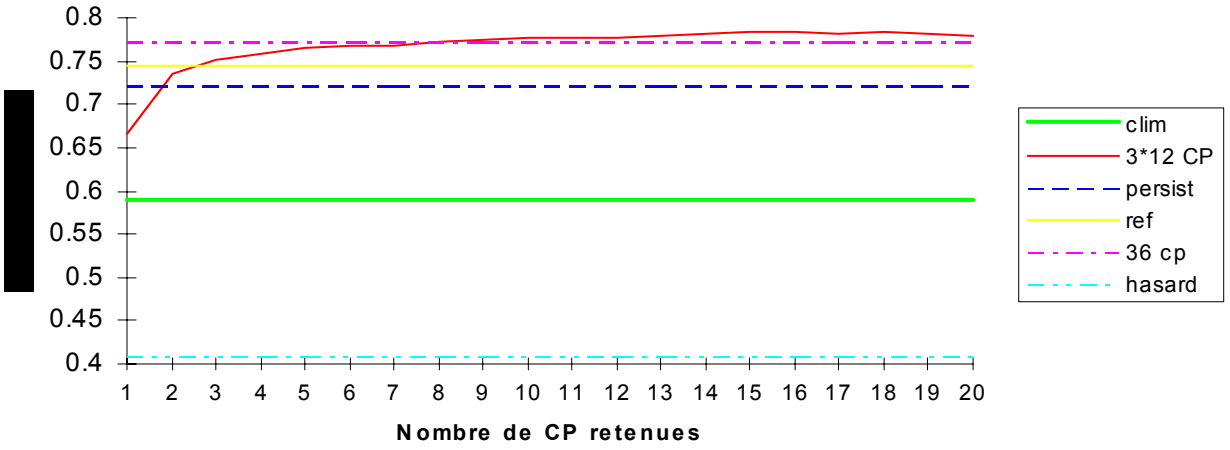
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
AIN-VALSERINE



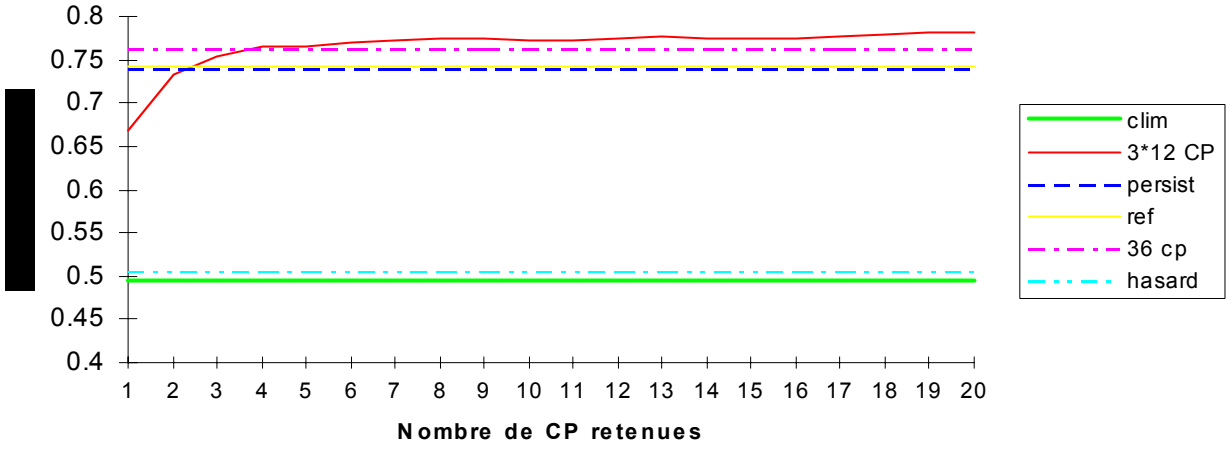
ARVE-FIER



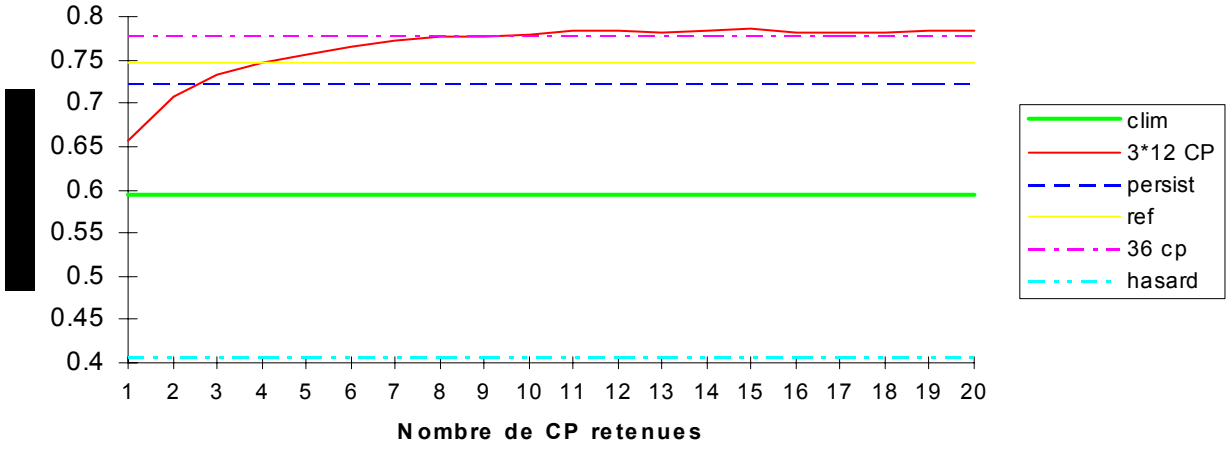
ISERE-DORON



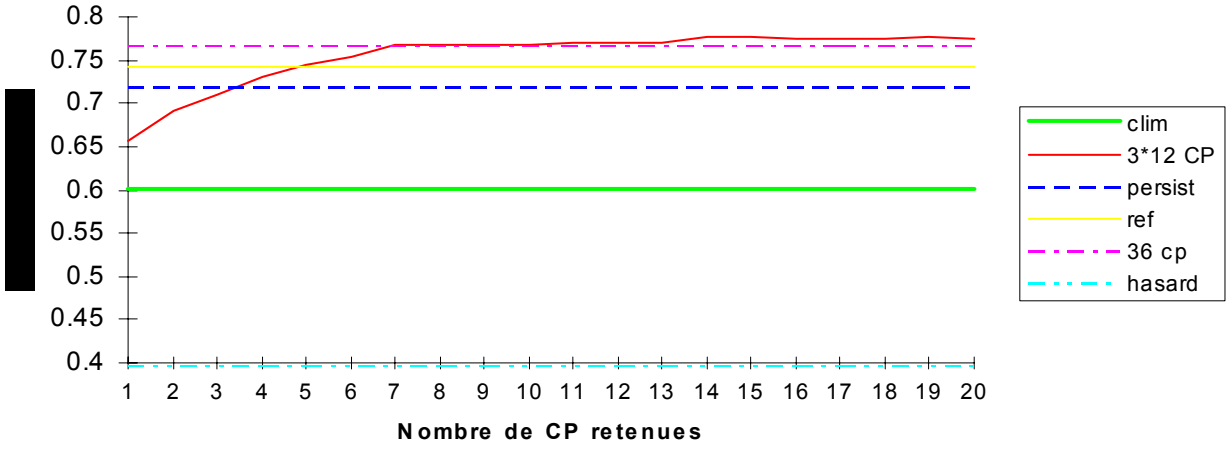
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
ISERE MOYENNE



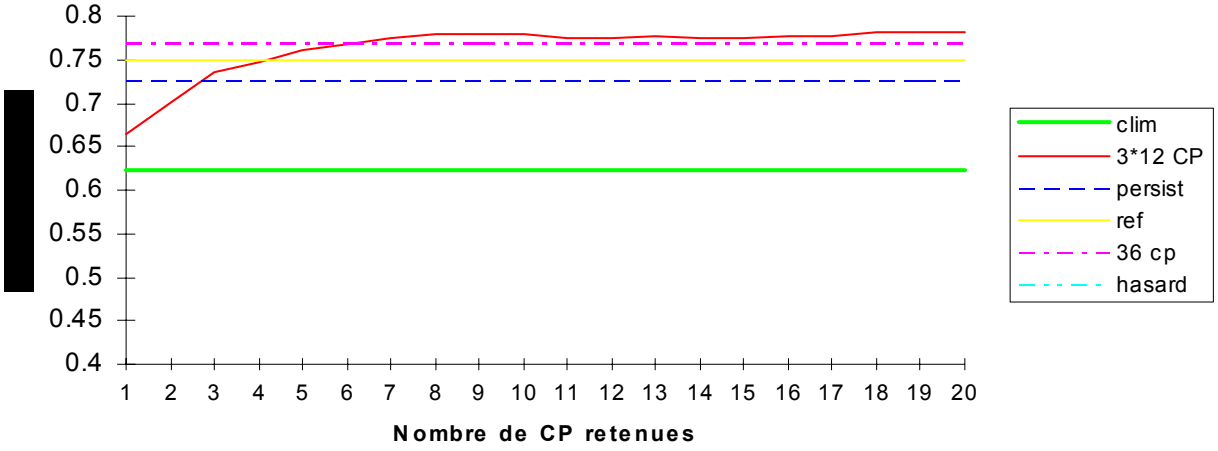
ROMANCHE-ARC INFERIEUR



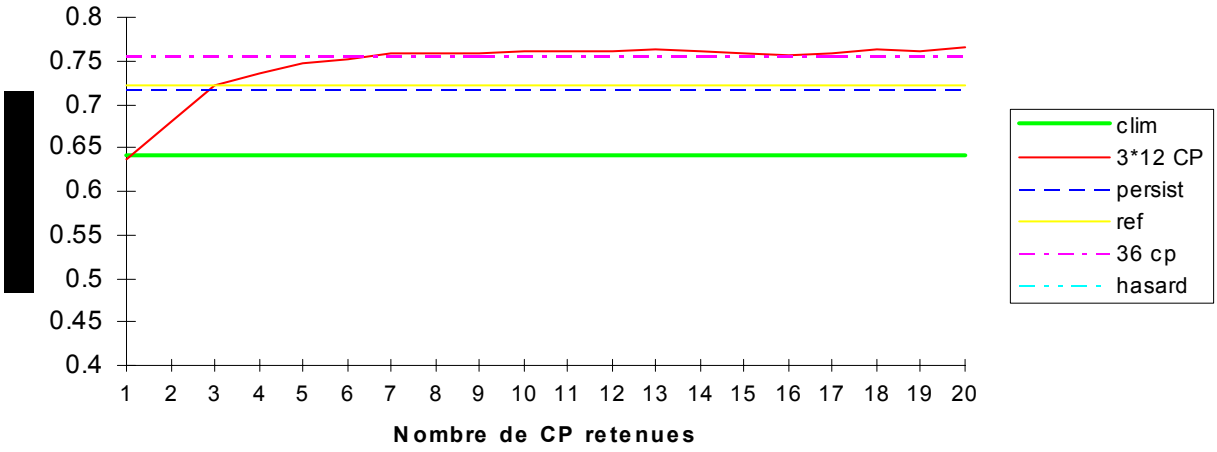
DRAC



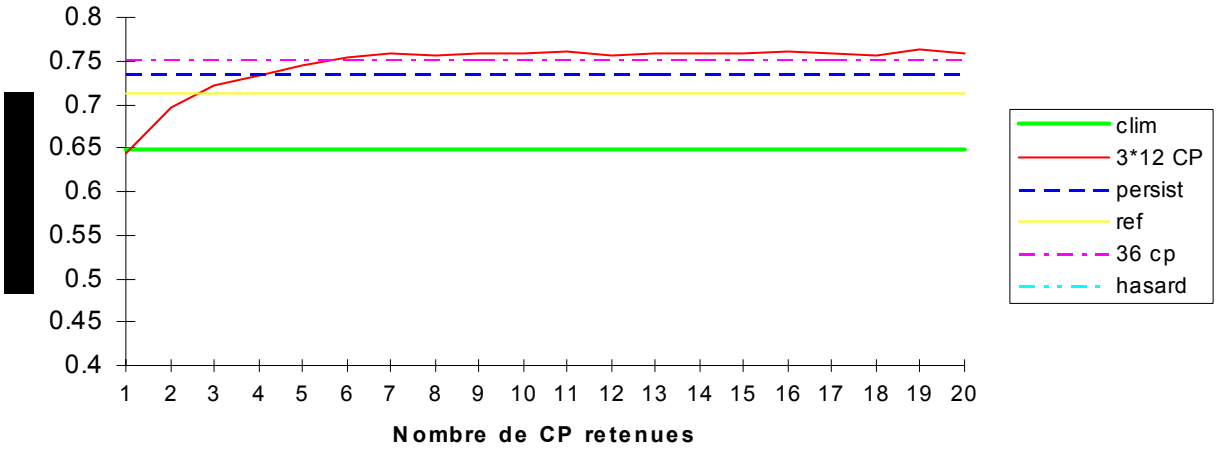
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
BUECH-DROME



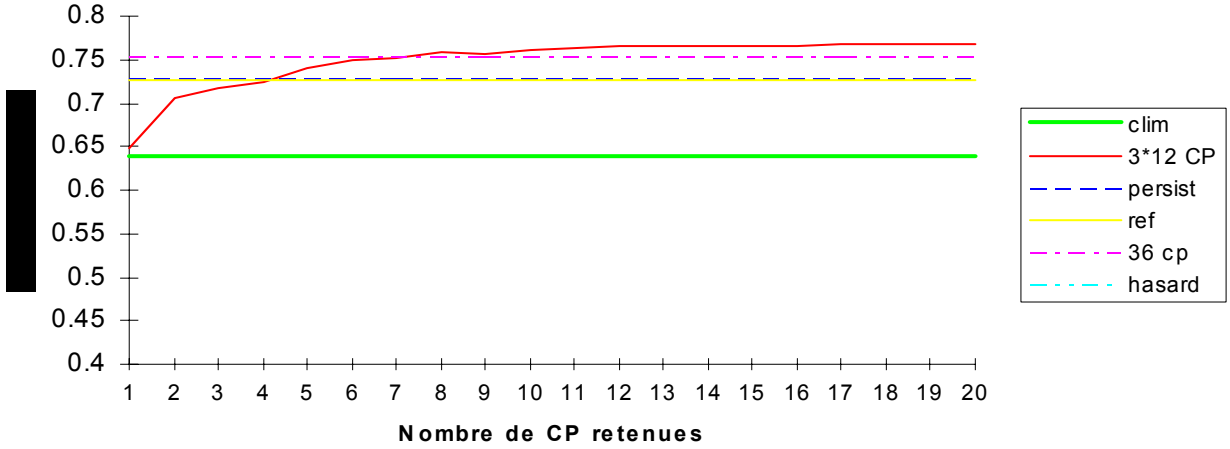
VERDON



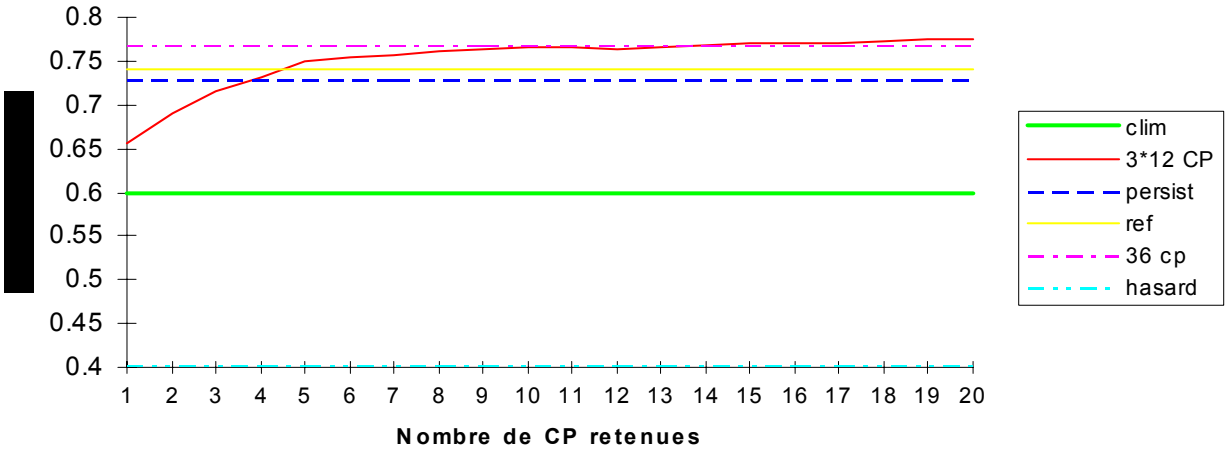
B.V.I. VERDON



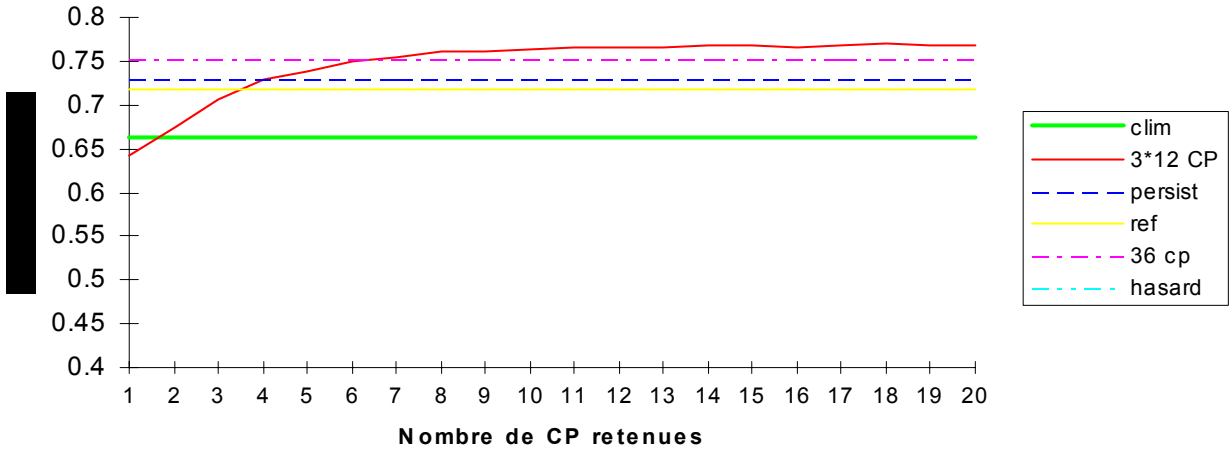
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
VAR-TINEE-ROYA



HAUTE DURANCE

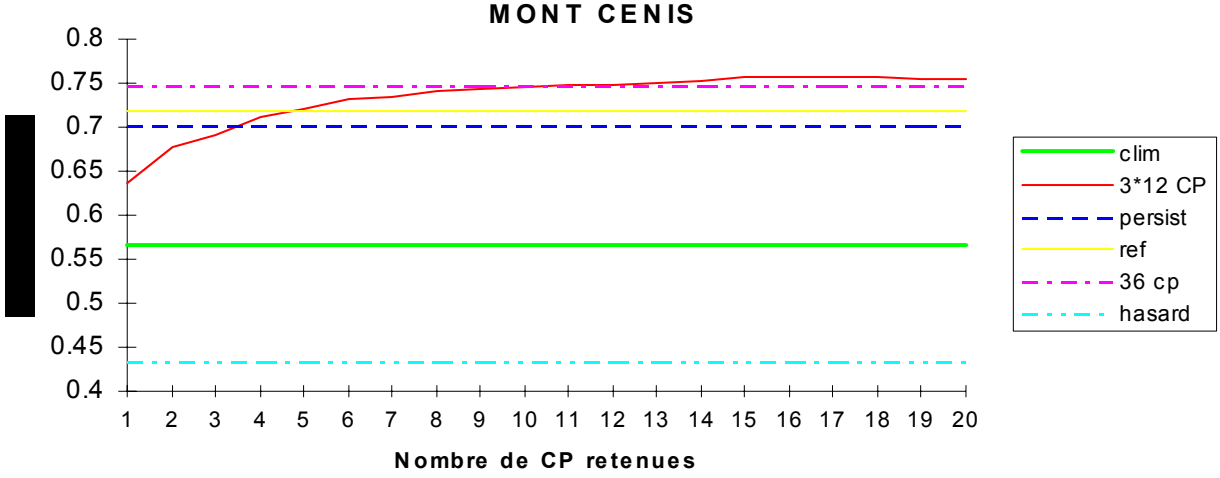


DURANCE MOYENNE

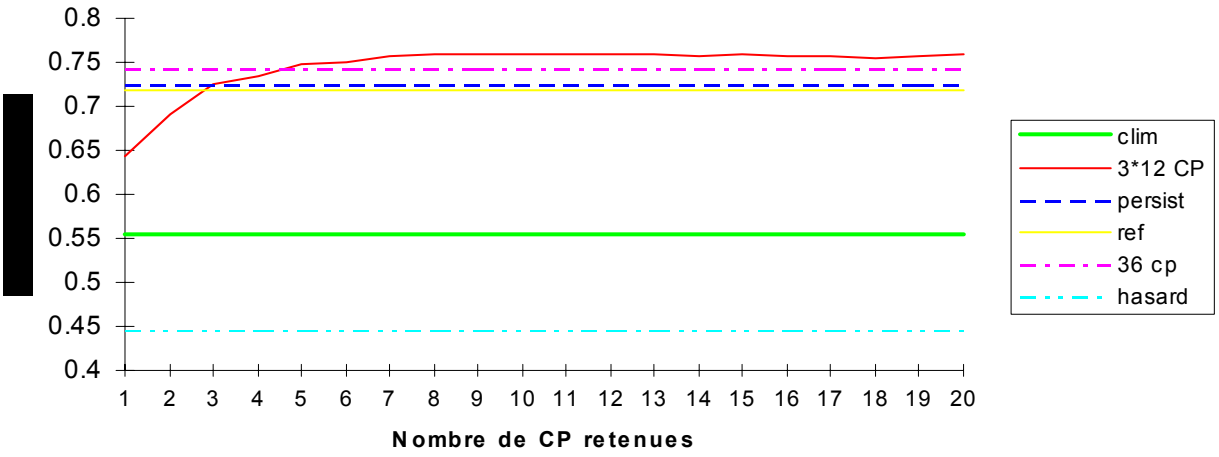


SELECTION ASCENDANTE: comparaison 3*12 CP et référence

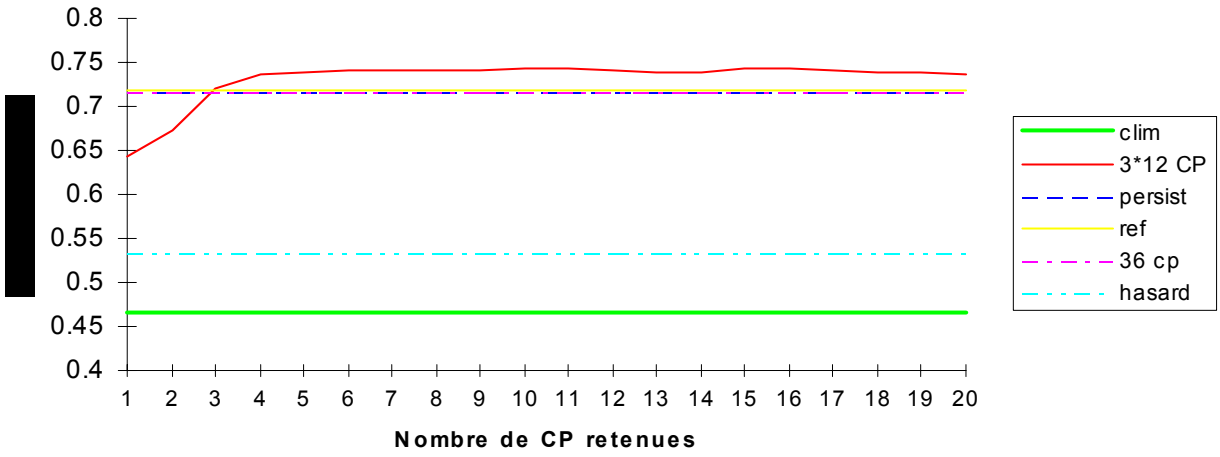
MONT CENIS



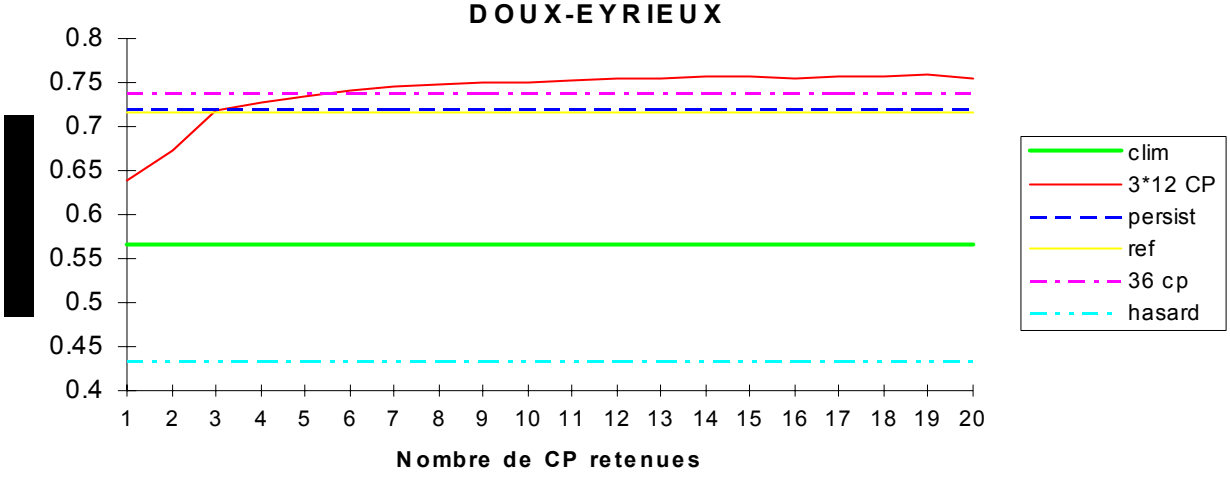
CHASSEZAC



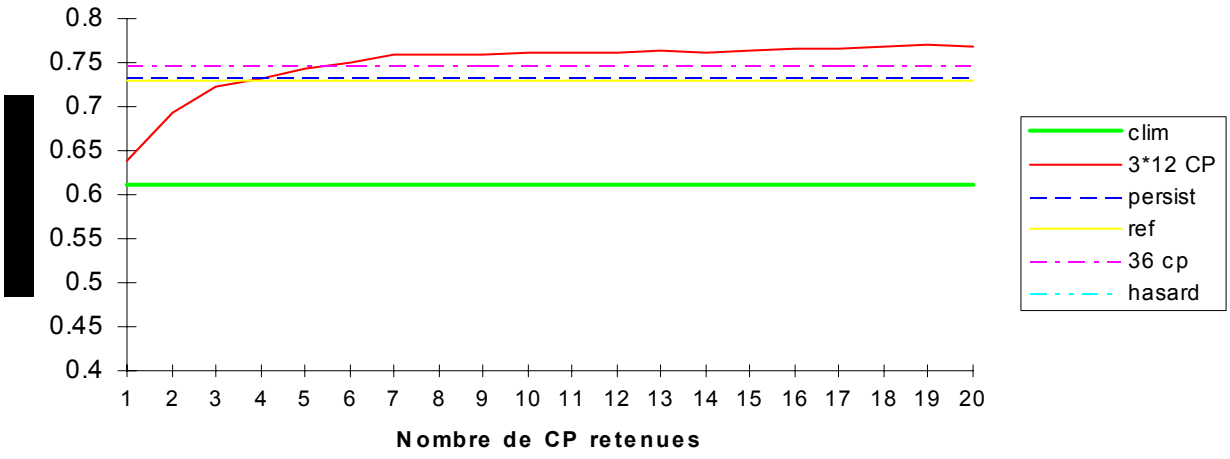
LOIRE SUPERIEURE



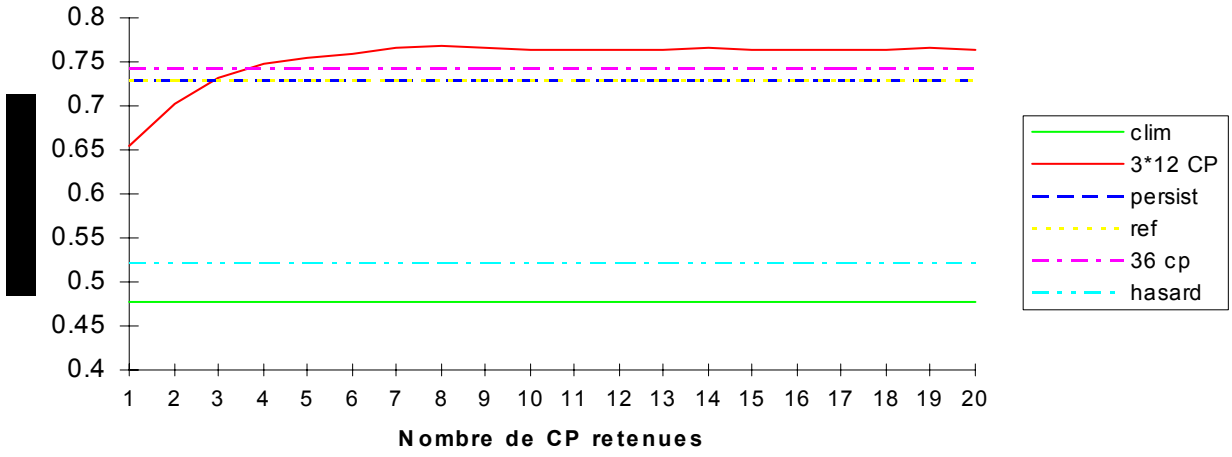
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
DOUX-EYRIEUX



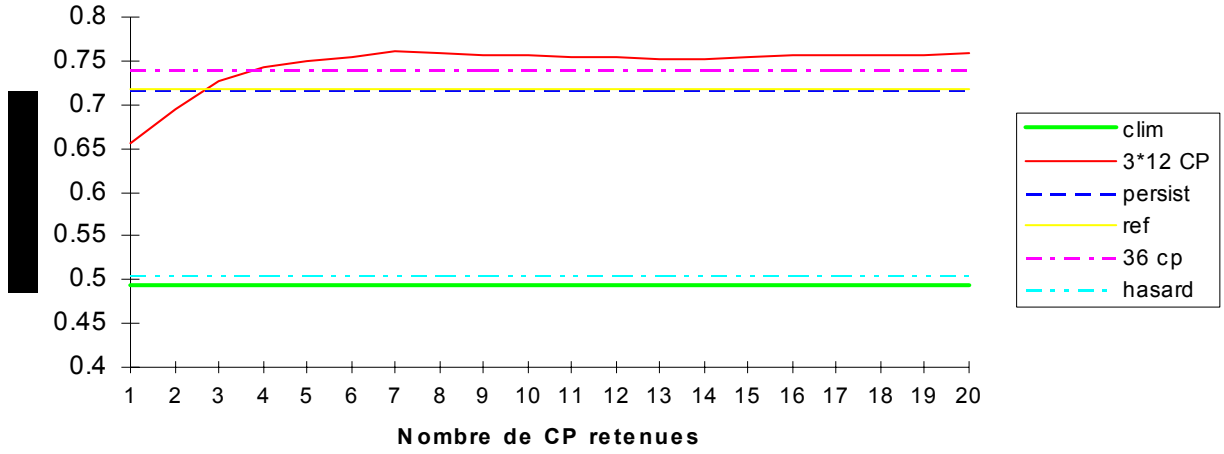
GARD-CEZE



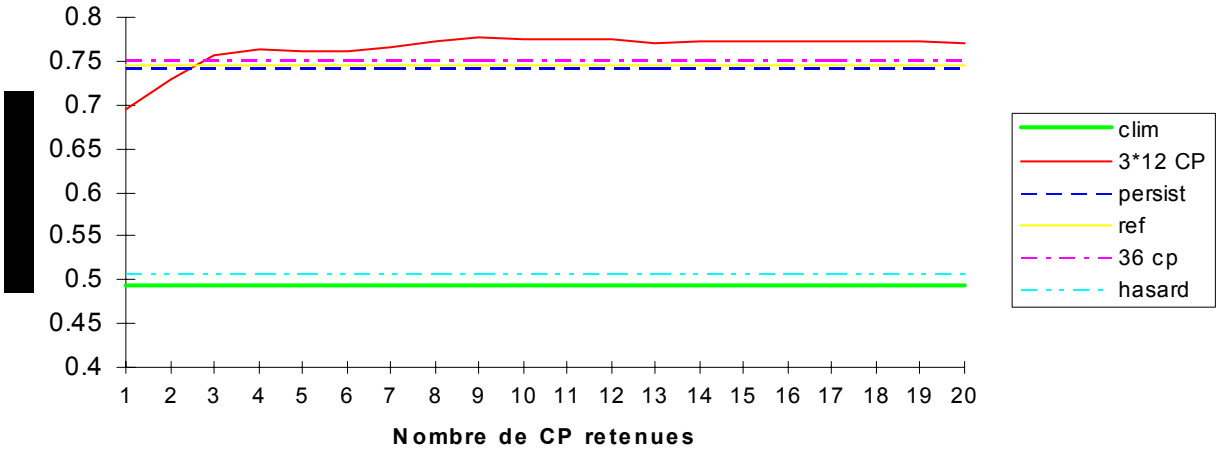
LOIRE MOYENNE



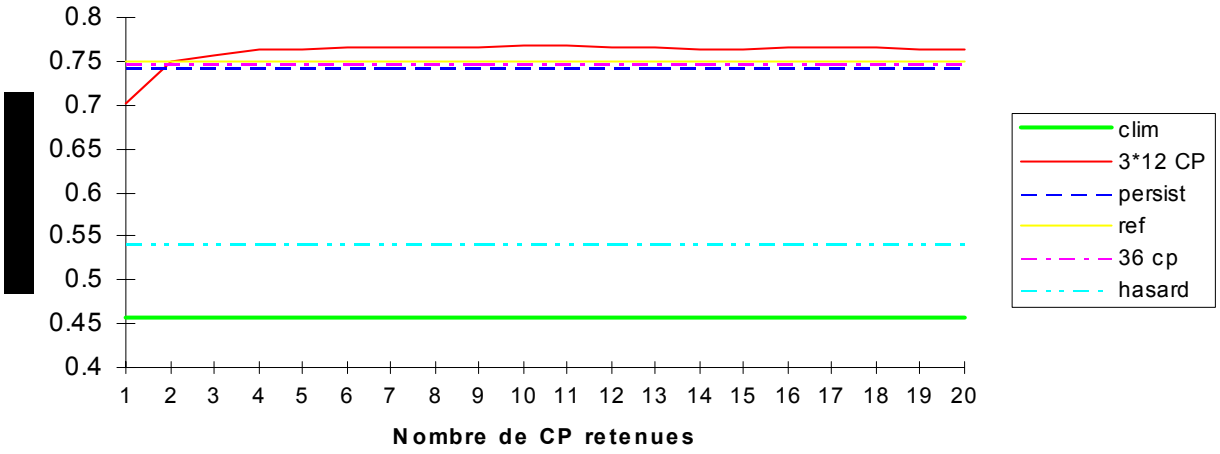
SELECTION ASCENDANTE: comparaison 3*12 CP et référence
ALLIER SUPERIEUR



SIOULE



CURE



ANNEXE III-3:

Comparaison sélection ascendante k=1 avec 36 et 24 CP

	groupements	36 CP	24 CP	écart
1	Creuse-Cher	76.33	76.09	-0.24
2	Vézère-Vienne-Thaurion	69.31	69.31	0.00
3	Dordogne	77.11	76.74	-0.37
4	Cère-Maronne	77.78	77.78	0.00
5	Truyère-Lot inférieur	75.50	75.58	0.08
6	Haut Tarn-Haut Lot	76.28	76.28	0.00
7	Agout-Tarn	72.39	72.34	-0.05
8	Pyrénées Est	73.81	74.91	1.10
9	Ariège-Vicdessos	72.39	72.63	0.24
10	Pique-Garonne-Salat	74.86	75.18	0.32
11	Gaves	72.15	71.24	-0.91
12	Doubs	76.17	76.36	0.19
13	Ain-Valserine	75.07	74.54	-0.53
14	Arve-Fier	73.95	73.87	-0.08
15	Isère-Doron	77.11	76.36	-0.75
16	Isère moyenne	77.57	77.57	0.00
17	Romanche-Arc inférieur	77.70	77.43	-0.27
18	Drac	76.68	77.11	0.43
19	Buech-Drôme	77.86	78.29	0.43
20	Verdon	75.82	75.82	0.00
21	BVI Verdon	75.64	76.28	0.64
22	Var-Tinee-Roya	75.93	75.93	0.00
23	Haute Durance	76.17	76.49	0.32
24	Durance moyenne	76.01	76.28	0.27
25	Mont Cenis	74.14	75.10	0.96
26	Chassezac	75.80	75.61	-0.19
27	Loire supérieure	74.19	74.35	0.16
28	Doux-Eyrieux	74.86	74.86	0.00
29	Gard-Cèze	75.85	75.85	0.00
30	Loire moyenne	76.76	75.90	-0.86
31	Allier supérieur	75.88	75.88	0.00
32	Sioule	77.22	77.35	0.13
33	Cure	76.47	76.07	-0.40
	moyenne	**75.48**	**75.50**	**0.02**

ANNEXE III-4 :

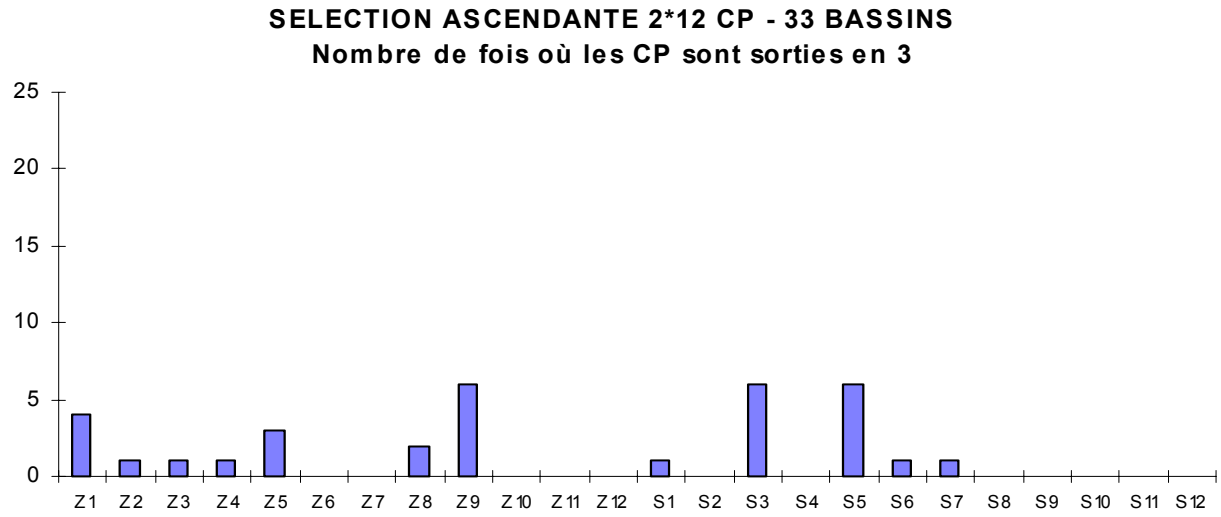
Matrice de corrélation Z (700 hPa) / S (1000 hPa)

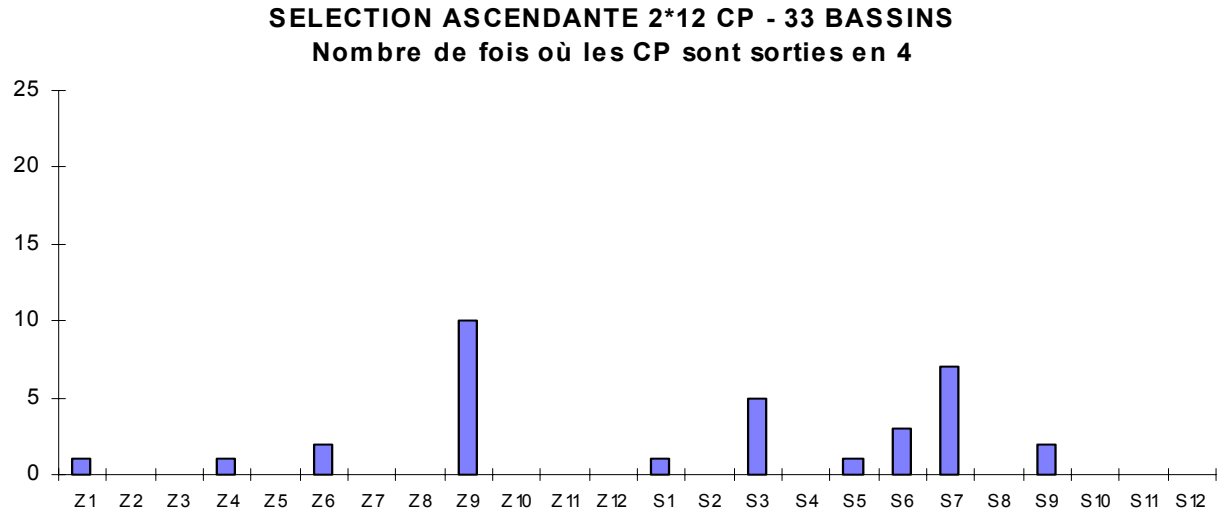
	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}
S_1	**0.86**	0.36	-0.03	0.05	0.09	0.10	-0.02	0.01	0.01	0.04	-0.01	0.02
S_2	0.20	**-0.74**	-0.43	0.18	0.32	0.01	0.01	0.03	-0.02	0.03	-0.02	0.01
S_3	-0.07	0.35	**-0.83**	-0.16	-0.05	-0.20	-0.01	0.01	0.01	-0.03	-0.03	-0.02
S_4	-0.06	0.12	-0.06	**0.85**	-0.21	-0.02	0.19	0.13	-0.07	-0.01	-0.07	0.00
S_5	0.01	0.13	0.14	0.06	**0.64**	-0.36	0.34	-0.16	0.22	-0.05	0.02	-0.06
S_6	-0.01	-0.04	0.04	-0.01	-0.22	**-0.67**	0.02	-0.13	-0.10	0.16	-0.00	-0.02
S_7	0.02	-0.04	-0.05	-0.15	-0.12	0.13	**0.68**	0.12	-0.04	0.11	0.33	0.12
S_8	0.03	-0.04	0.03	-0.09	0.05	-0.11	-0.09	**0.70**	0.14	-0.01	-0.10	0.14
S_9	-0.04	0.00	-0.05	0.11	-0.09	0.03	-0.10	-0.08	**0.68**	0.03	0.11	-0.13
S_{10}	0.02	0.02	0.01	0.00	-0.02	-0.01	-0.01	0.04	0.04	**0.68**	-0.06	-0.03
S_{11}	-0.09	0.10	0.01	0.04	0.15	0.05	0.01	-0.13	-0.01	0.06	**0.31**	**0.43**
S_{12}	-0.01	0.03	0.01	0.11	0.08	-0.09	-0.18	0.03	0.02	0.04	**0.44**	**0.33**

ANNEXE III-5:

Histogrammes des CP sorties en 3ème et 4ème positions

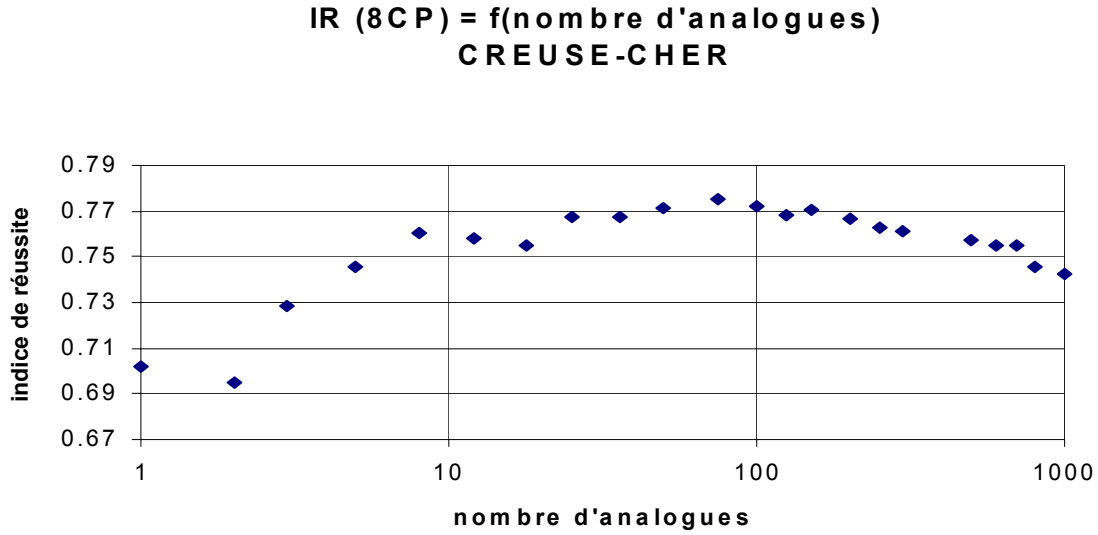
(Sélection ascendante k=1 sur 24 CP)

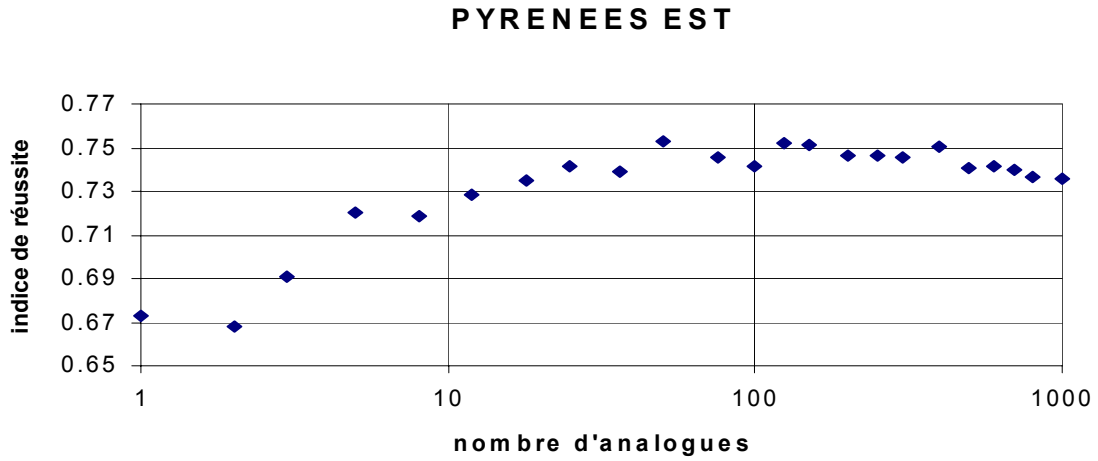


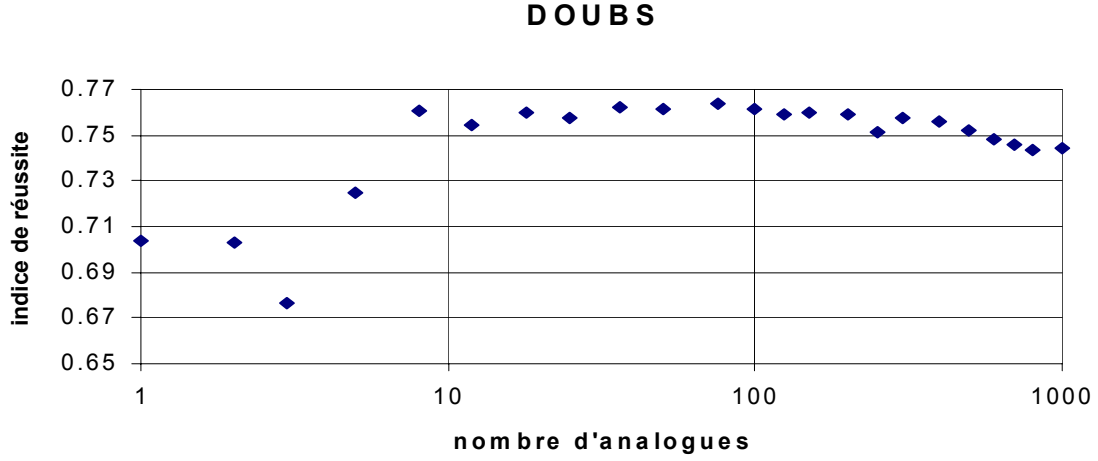


ANNEXE III-6 :

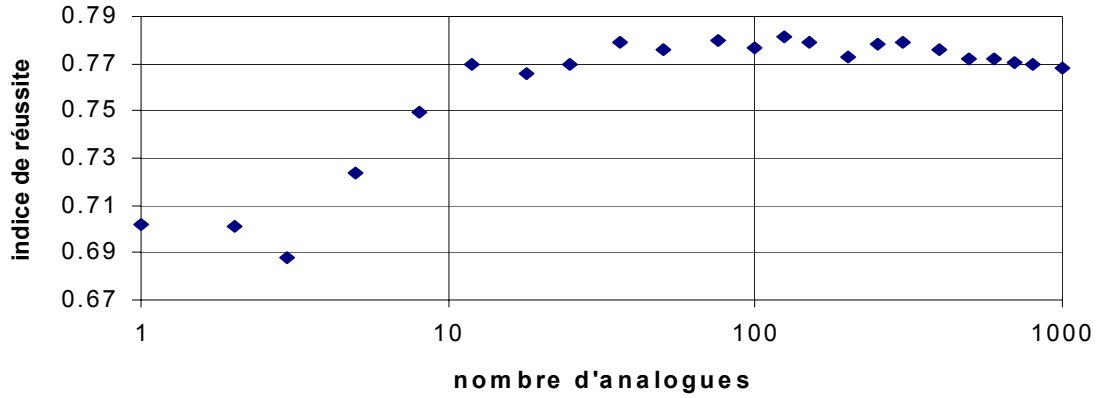
Optimisation du nombre d'analogues (sélection ascendante k=1 avec 24 CP)



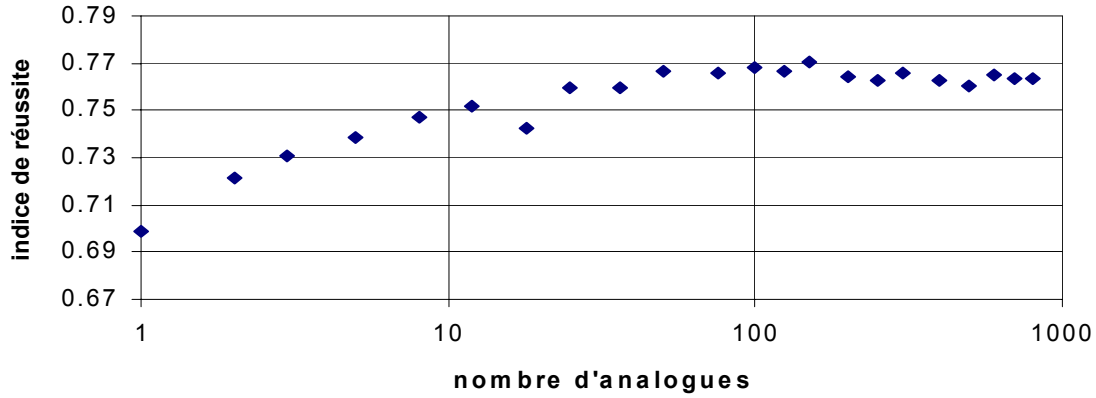




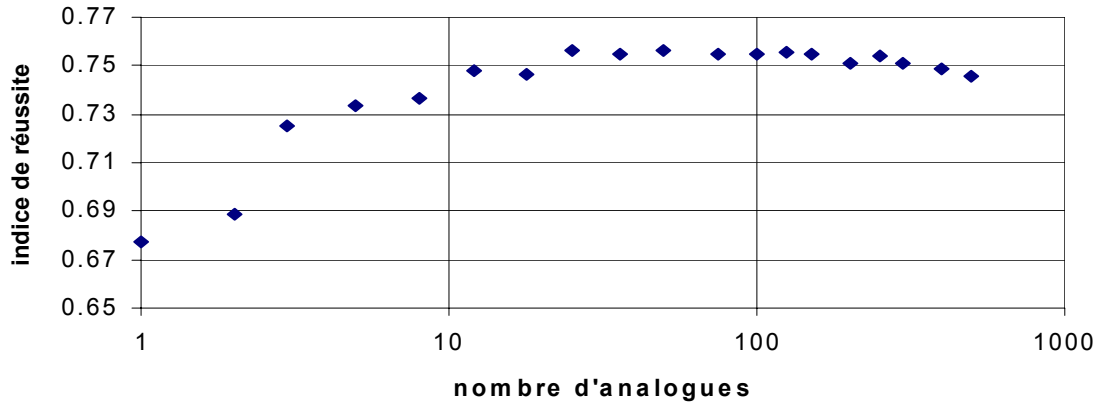
IR (8CP) = f(nombre d'analogues)
ISERE MOYENNE



VAR-TINEE-ROYA



CHASSEZAC



ANNEXE III-7:

Comparaison sélections ascendantes k=1 / k=2 (avec 24 CP)

	groupements	K=1	K=2	écart
1	Creuse-Cher	76.09	76.17	0.08
2	Vézère-Vienne-Thaurion	69.31	69.63	0.32
3	Dordogne	76.74	76.98	0.24
4	Cère-Maronne	77.78	77.89	0.11
5	Truyère-Lot inférieur	75.58	76.09	0.51
6	Haut Tarn-Haut Lot	76.28	76.44	0.16
7	Agout-Tarn	72.34	72.82	0.48
8	Pyrénées Est	74.91	74.91	0.00
9	Ariège-Vicdessos	72.63	72.66	0.03
10	Pique-Garonne-Salat	75.18	75.18	0.00
11	Gaves	71.24	71.86	0.62
12	Doubs	76.36	77.27	0.91
13	Ain-Valserine	74.54	74.70	0.16
14	Arve-Fier	73.87	74.00	0.13
15	Isère-Doron	76.36	77.91	1.55
16	Isère moyenne	77.57	77.75	0.18
17	Romanche-Arc inférieur	77.43	78.08	0.65
18	Drac	77.11	77.14	0.03
19	Buech-Drôme	78.29	78.10	-0.19
20	Verdon	75.82	75.82	0.00
21	BVI Verdon	76.28	76.28	0.00
22	Var-Tinee-Roya	75.93	75.80	-0.13
23	Haute Durance	76.49	76.49	0.00
24	Durance moyenne	76.28	76.60	0.32
25	Mont Cenis	75.10	75.31	0.21
26	Chassezac	75.61	75.61	0.00
27	Loire supérieure	74.35	74.19	-0.16
28	Doux-Eyrieux	74.86	75.10	0.24
29	Gard-Cèze	75.85	75.85	0.00
30	Loire moyenne	75.90	75.98	0.08
31	Allier supérieur	75.88	75.88	0.00
32	Sioule	77.35	77.35	0.00
33	Cure	76.07	76.57	0.50
	moyenne	**75.50**	**75.71**	**0.21**

ANNEXE III-8:

Comparaison méthode de référence / S-8CP

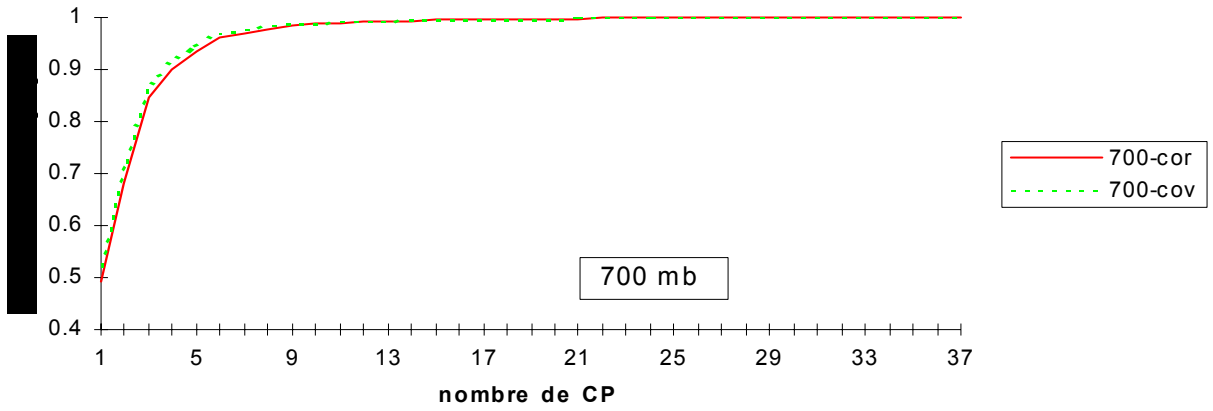
	groupements	référence	S-8CP	écart
1	Creuse-Cher	72.50	76.17	3.67
2	Vézère-Vienne-Thaurion	65.75	69.63	3.88
3	Dordogne	74.19	76.98	2.79
4	Cère-Maronne	74.03	77.89	3.86
5	Truyère-Lot inférieur	72.05	76.09	4.04
6	Haut Tarn-Haut Lot	72.90	76.44	3.54
7	Agout-Tarn	67.94	72.82	4.88
8	Pyrénées Est	68.88	74.91	6.03
9	Ariège-Vicdessos	67.35	72.66	5.31
10	Pique-Garonne-Salat	70.20	75.18	4.98
11	Gaves	66.79	71.86	5.07
12	Doubs	74.30	77.27	2.97
13	Ain-Valserine	72.66	74.70	2.04
14	Arve-Fier	71.64	74.00	2.36
15	Isère-Doron	74.38	77.91	3.53
16	Isère moyenne	74.62	77.75	3.13
17	Romanche-Arc inférieur	74.97	78.08	3.11
18	Drac	74.91	77.14	2.23
19	Buech-Drôme	73.95	78.10	4.15
20	Verdon	73.09	75.82	2.73
21	BVI Verdon	72.39	76.28	3.89
22	Var-Tinee-Roya	73.17	75.80	2.63
23	Haute Durance	74.24	76.49	2.25
24	Durance moyenne	72.61	76.60	3.99
25	Mont Cenis	72.23	75.31	3.08
26	Chassezac	71.43	75.61	4.18
27	Loire supérieure	71.78	74.19	2.41
28	Doux-Eyrieux	71.24	75.10	3.86
29	Gard-Cèze	72.82	75.85	3.03
30	Loire moyenne	73.79	75.98	2.19
31	Allier supérieur	72.37	75.88	3.51
32	Sioule	74.16	77.35	3.19
33	Cure	73.81	76.57	2.76
	moyenne	**72.22**	**75.71**	**3.49**

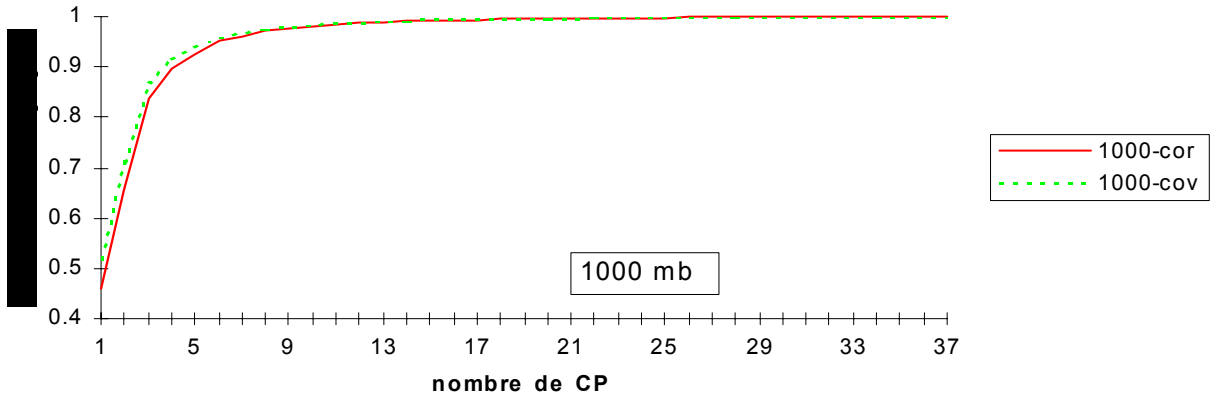
ANNEXE III-9:

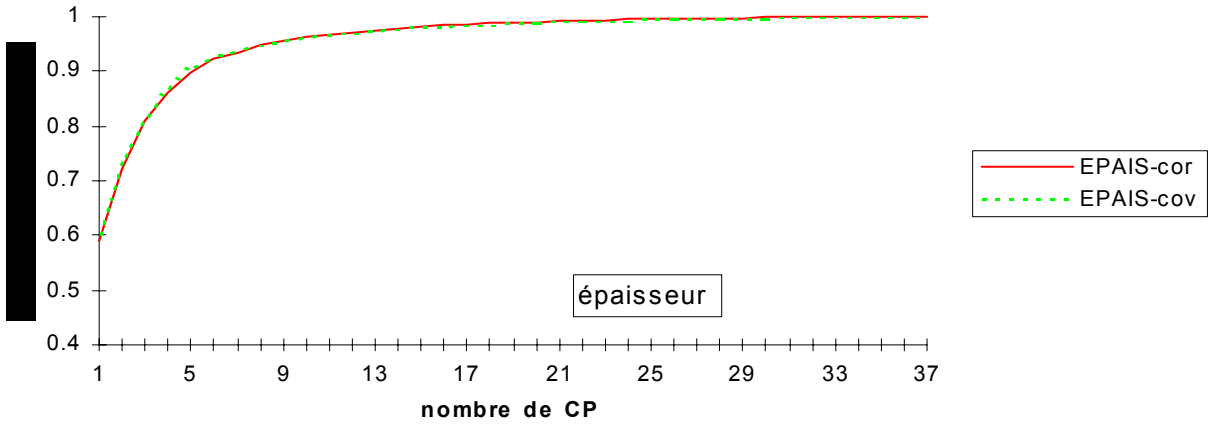
Comparaison ACPP 1) avec la matrice de covariance (cov)

2) avec la matrice de corrélation (cor)

Automnes 1953-1993
ACPP 59 covariance / corrélation

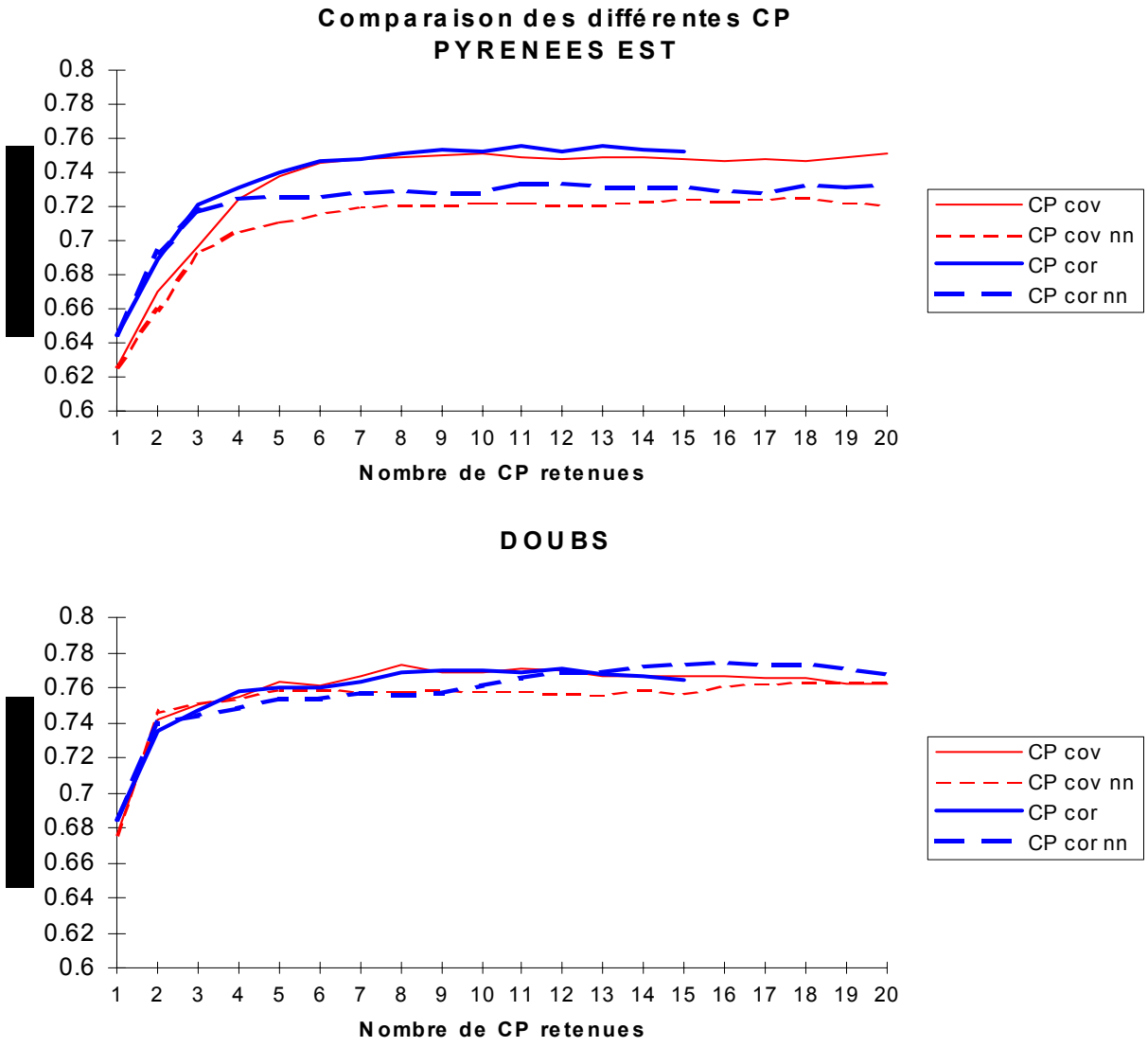


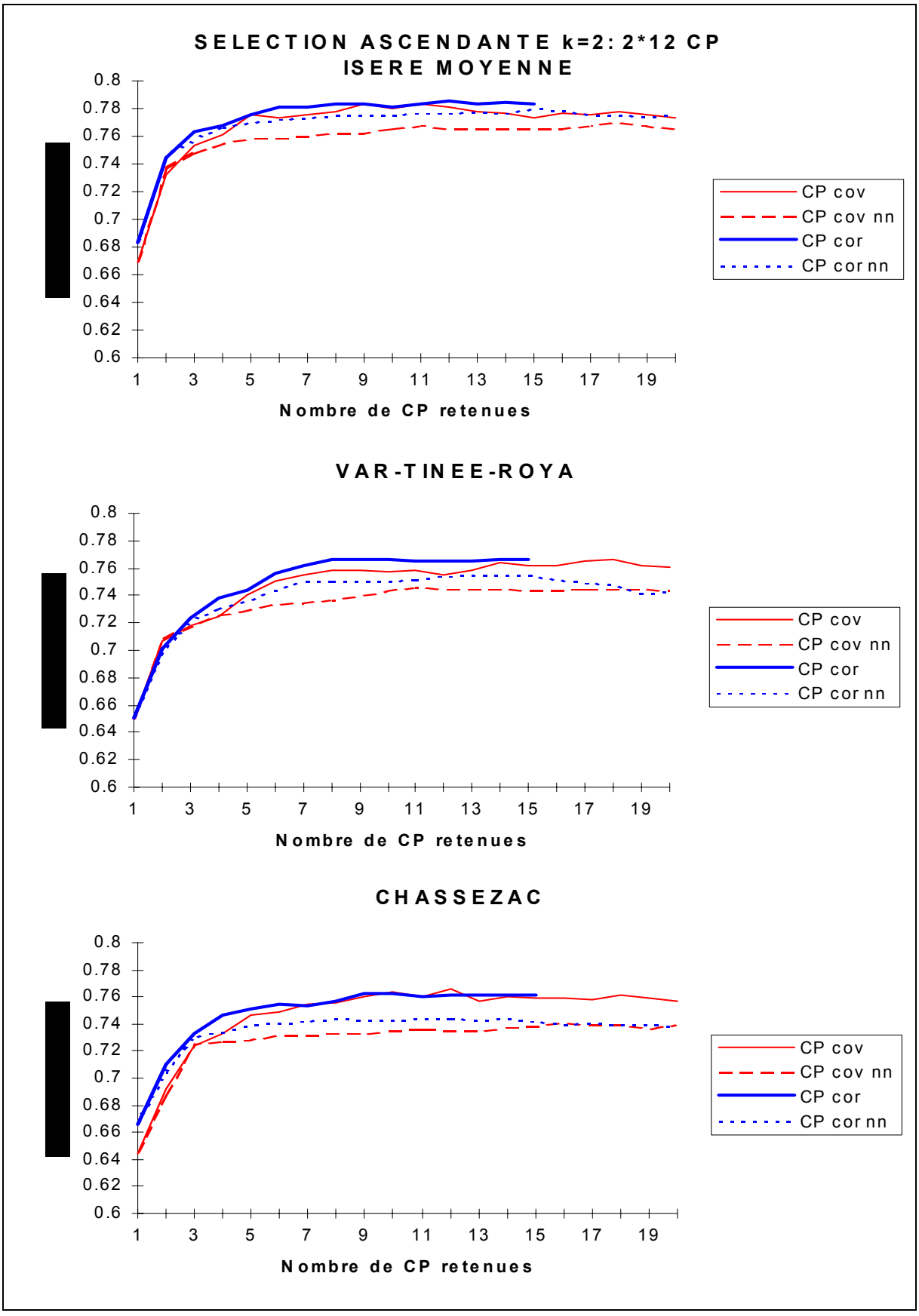




ANNEXE III-10:

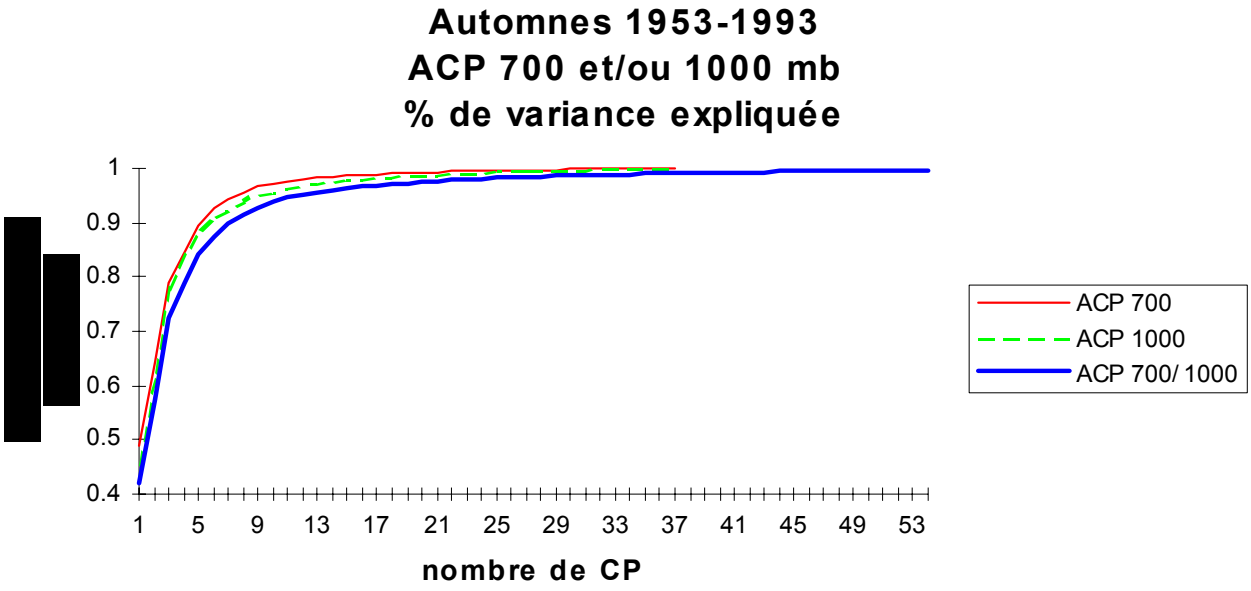
Courbes de l'indice de réussite avec différents types de CP




SELECTION ASCENDANTE k=2: 2*12 CP
ISERE MOYENNE
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
CP cov
CP cov nn
CP cor
CP cor nn
VAR-TINEE-ROYA
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
CP cov
CP cov nn
CP cor
CP cor nn
CHASSEZAC
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
CP cov
CP cov nn
CP cor
CP cor nn

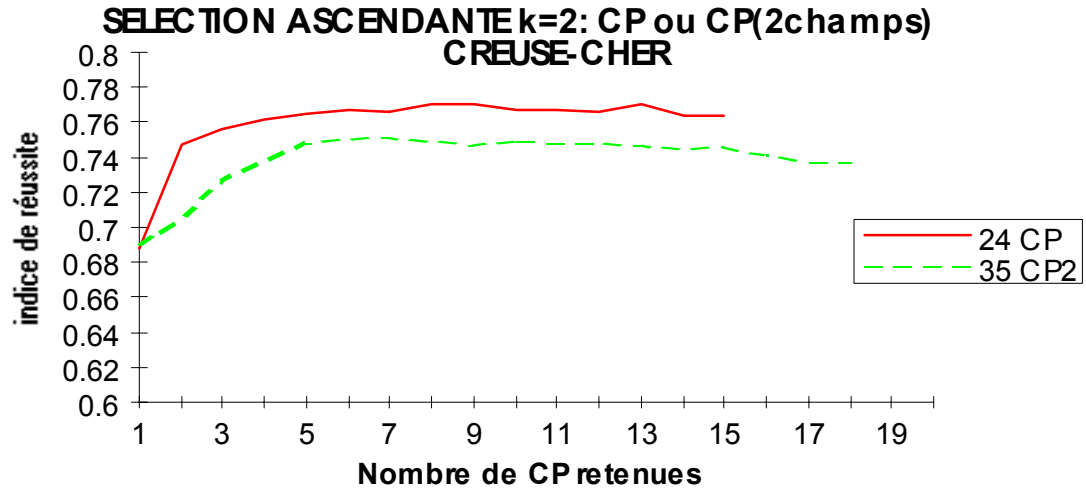
ANNEXE III-11:

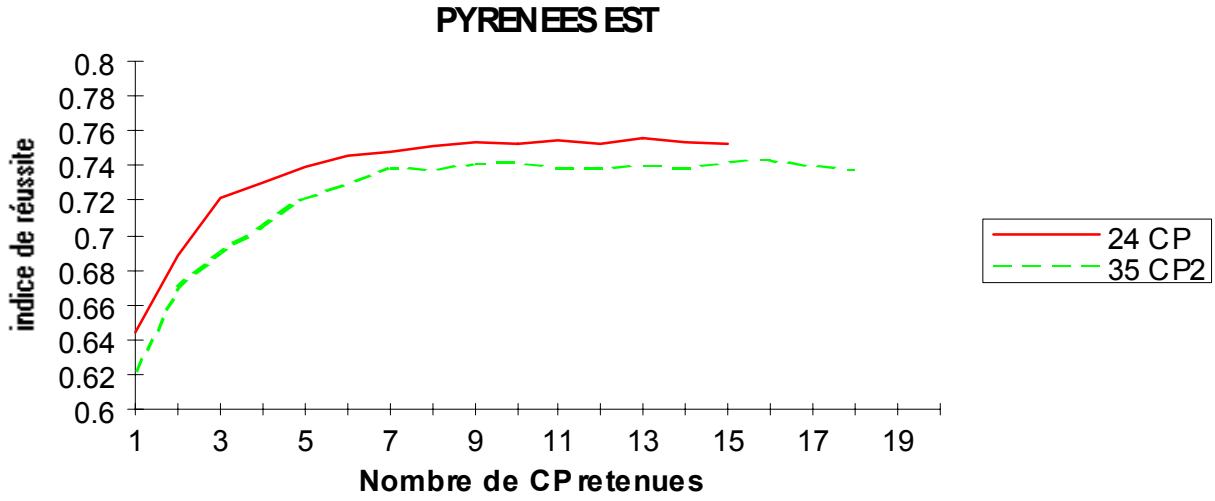
Pourcentage de variance expliquée en fonction du nombre de CP retenues pour l'ACP combinée des champs 700 et 1000 hPa

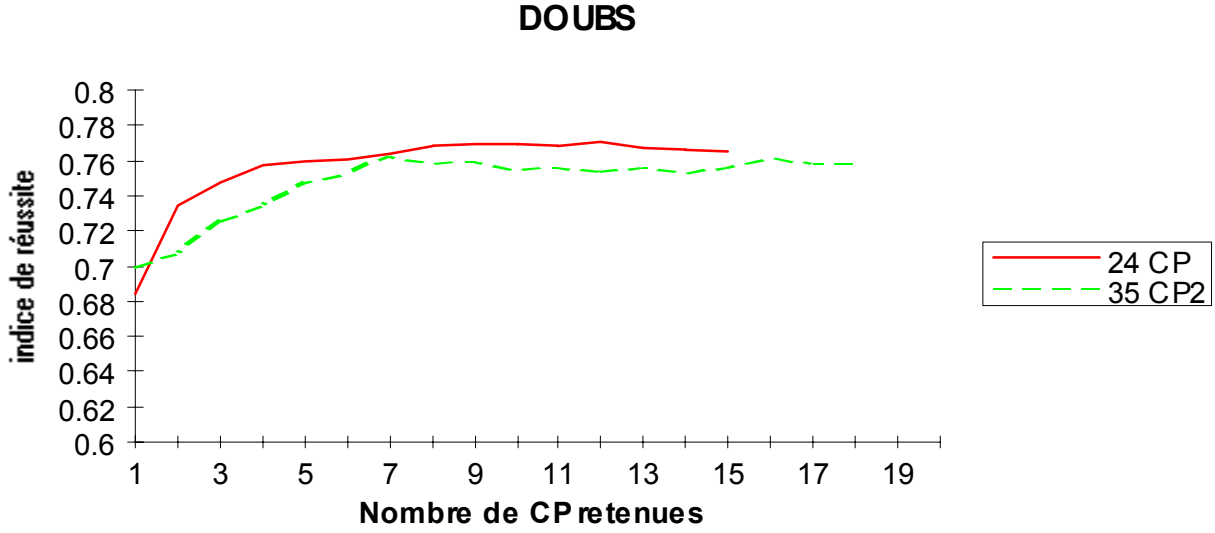


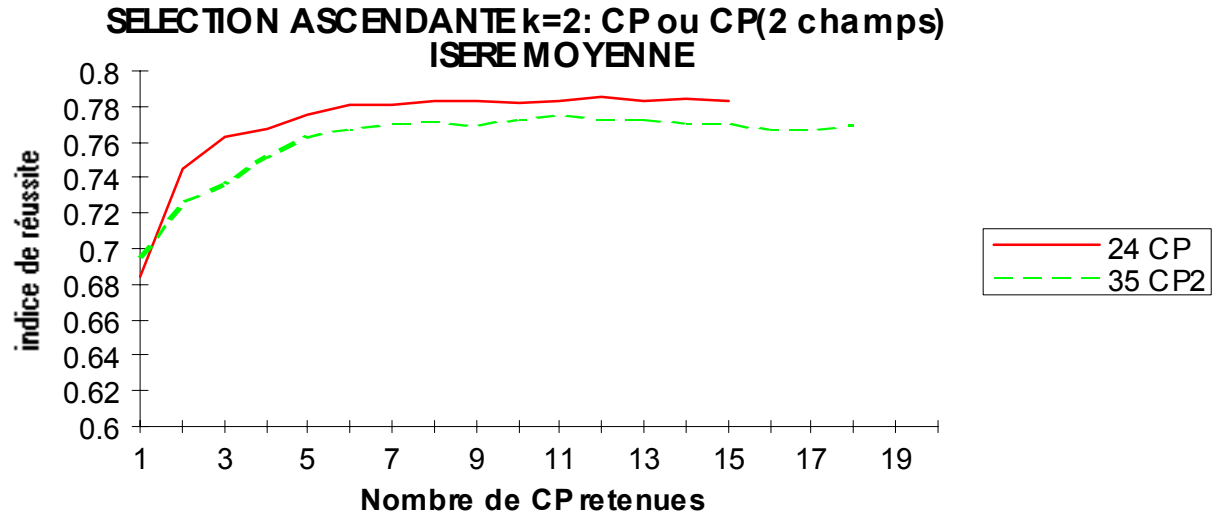
ANNEXE III-12 :

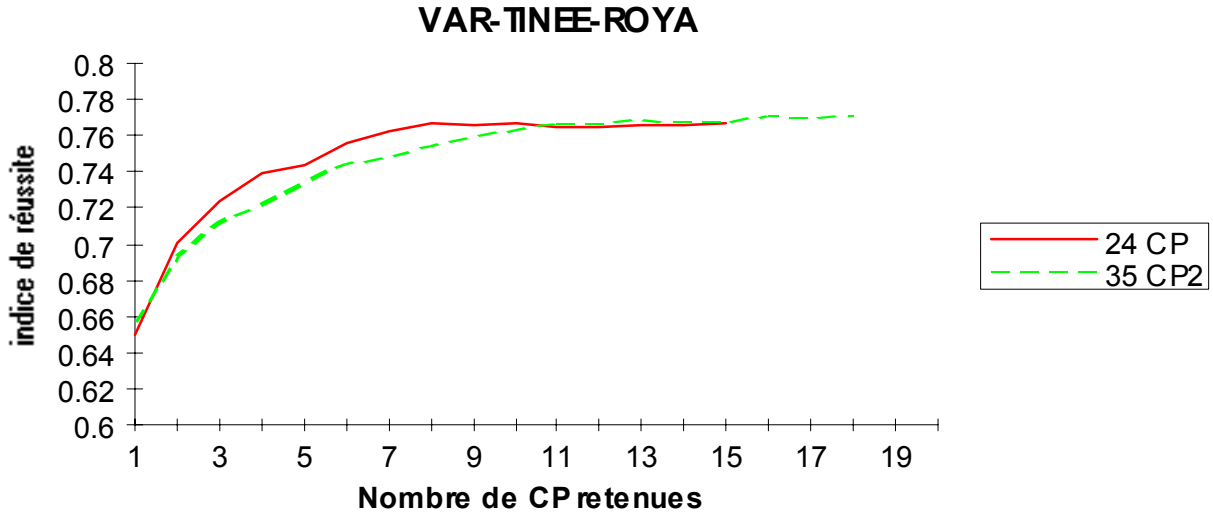
Courbes d'évolution de IR pour les CP combinées et les 6 groupements témoins

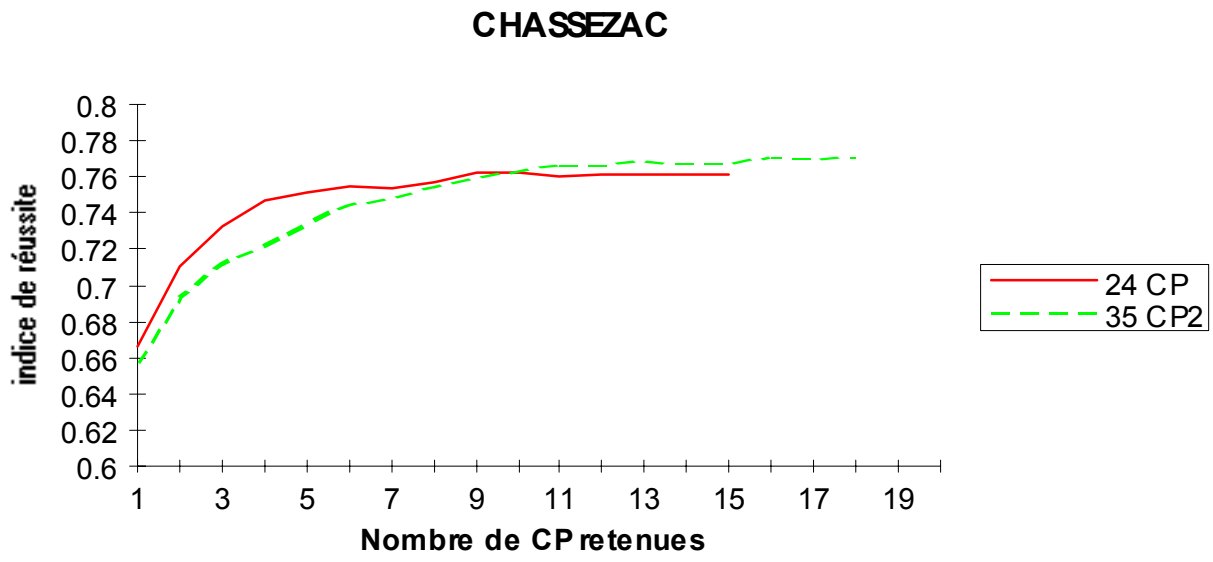






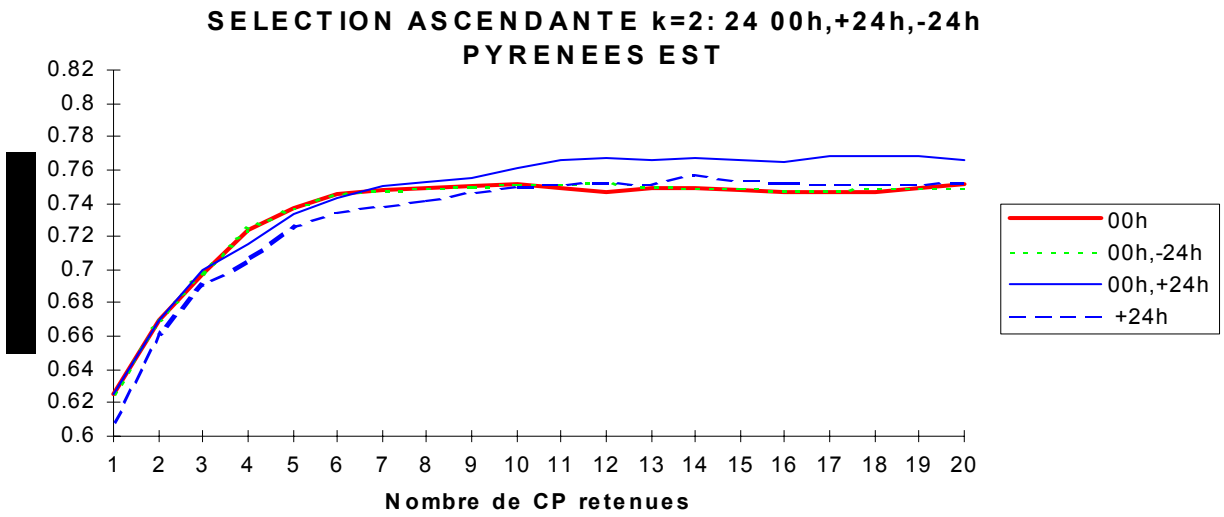

SELECTION ASCENDANTE k=2: CP ou CP(2 champs)
ISERE MOYENNE
indice de réussite
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
24 CP
35 CP2

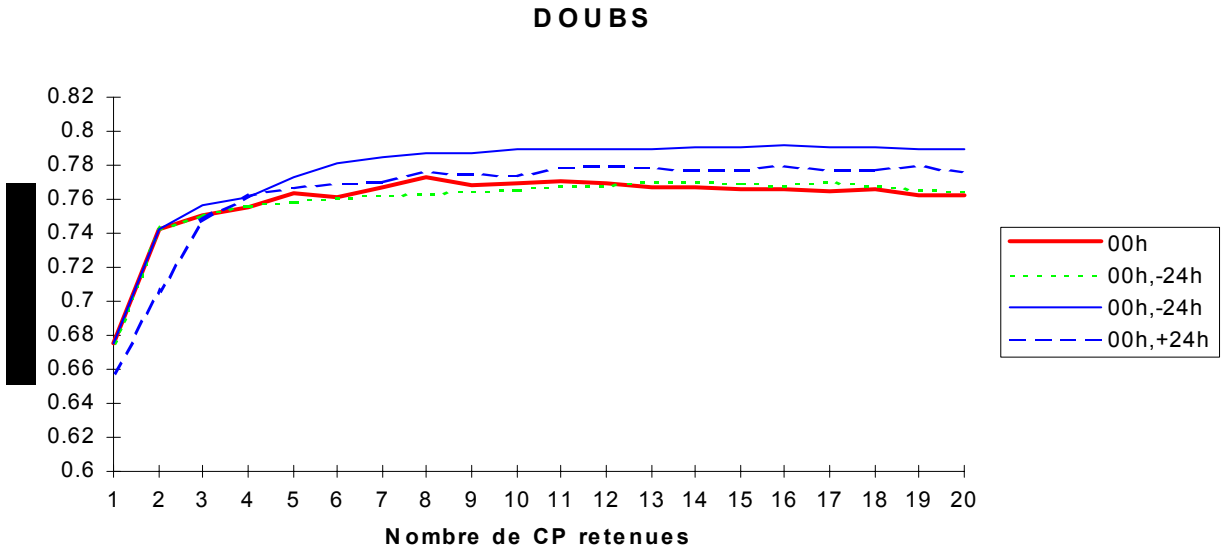

VAR-TINEE-ROYA
indice de réussite
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
24 CP
35 CP2


CHASSEZAC
indice de réussite
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
24 CP
35 CP2

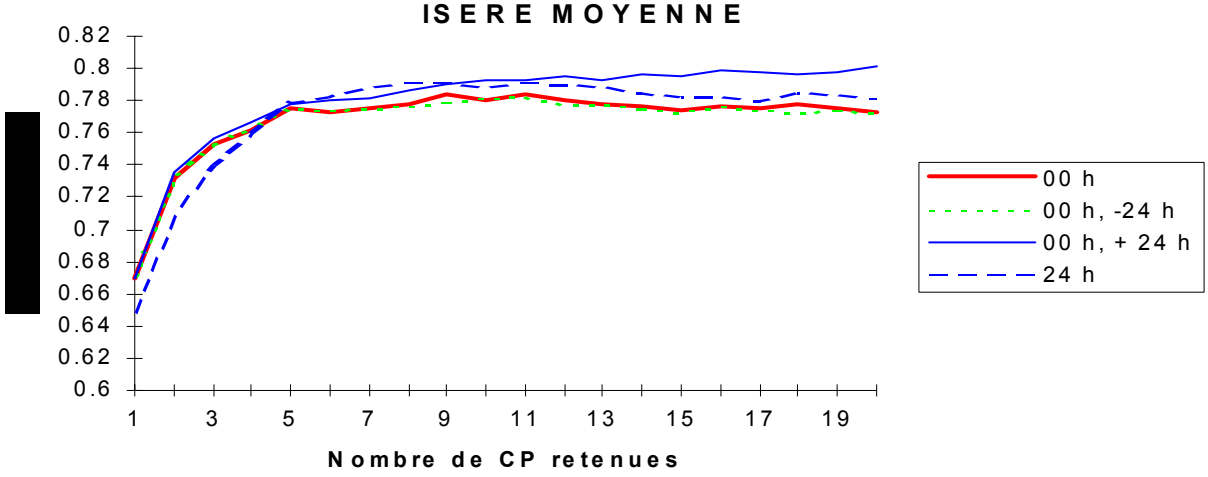
ANNEXE III-13 :

Courbes d'évolution de IR pour les CP aux différentes échéances

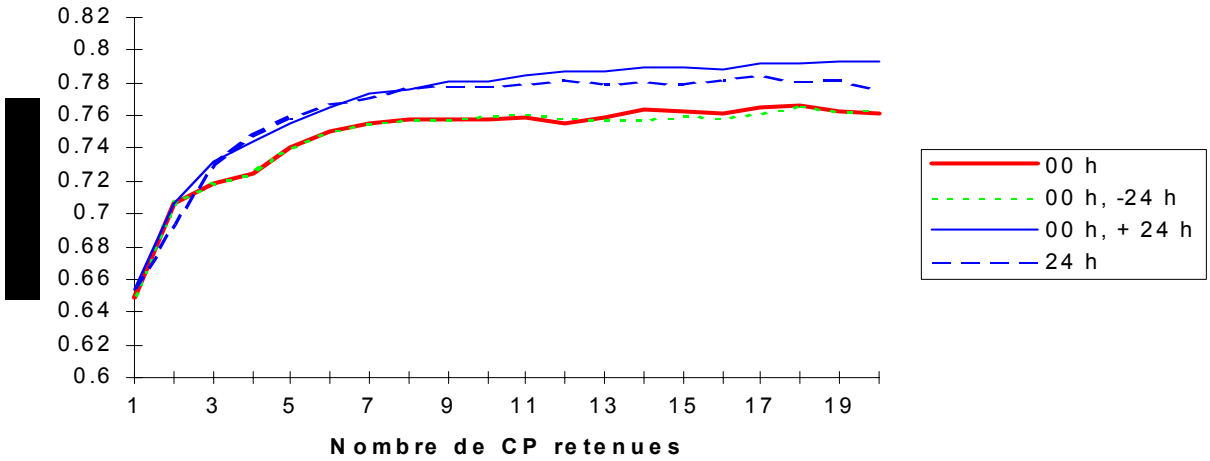




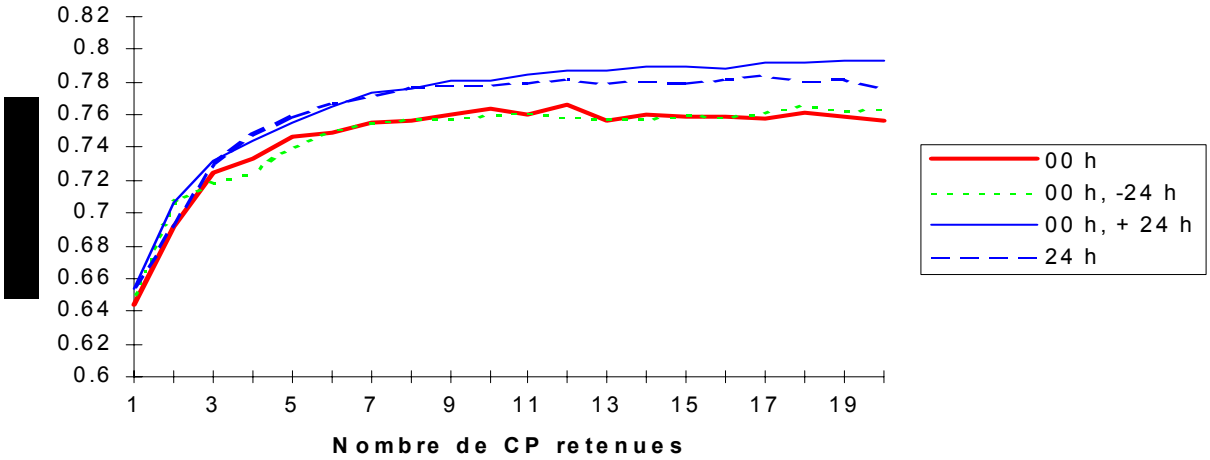
SELECTION ASCENDANTE k=2: CP 00h, +24h, -24h
ISERE MOYENNE



VAR-TINEE-ROYA

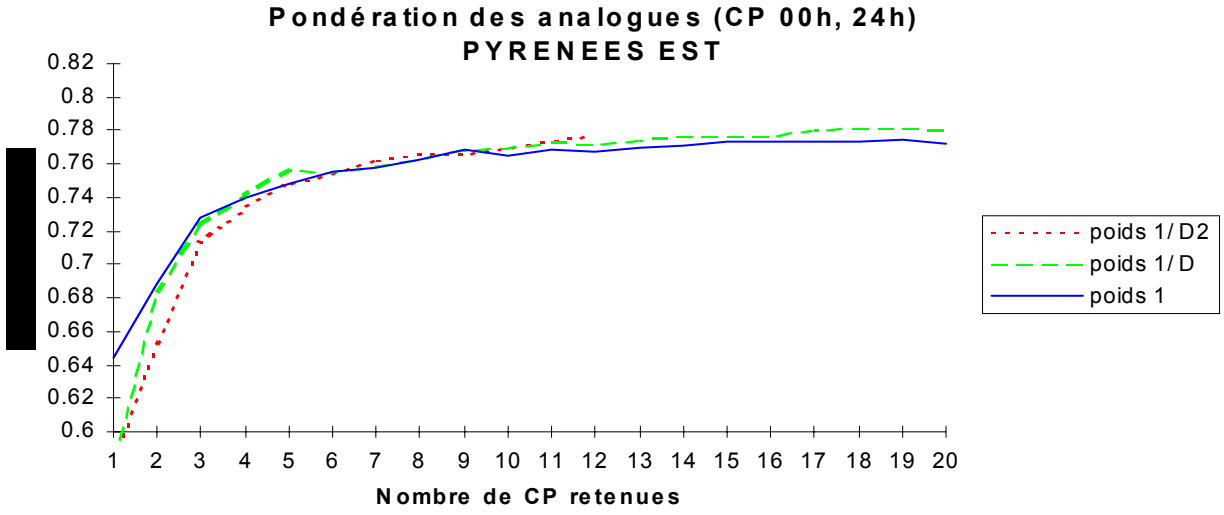


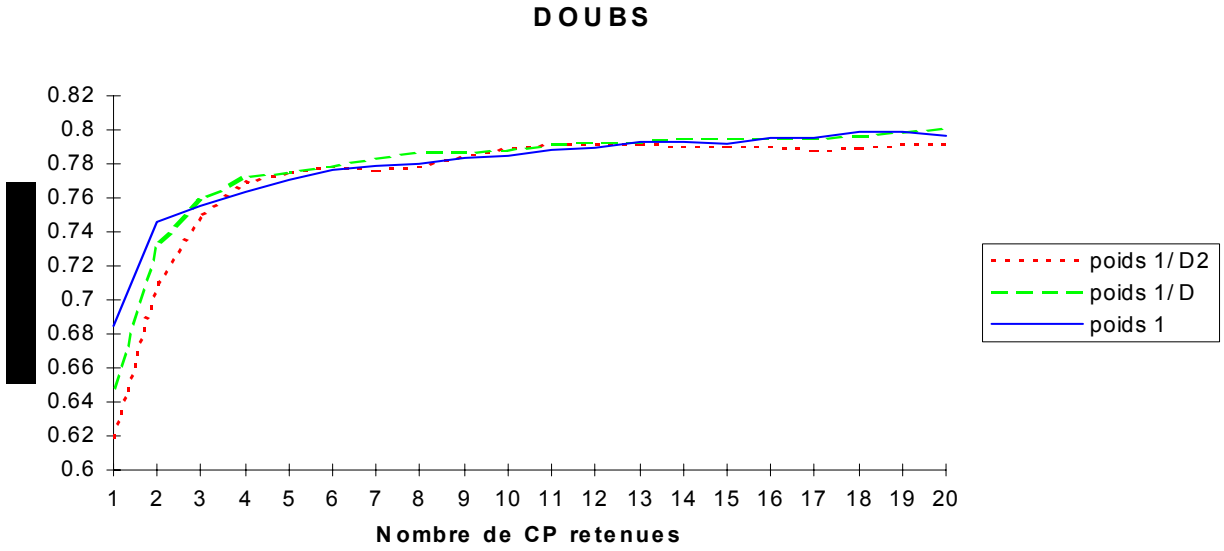
CHASSEZAC



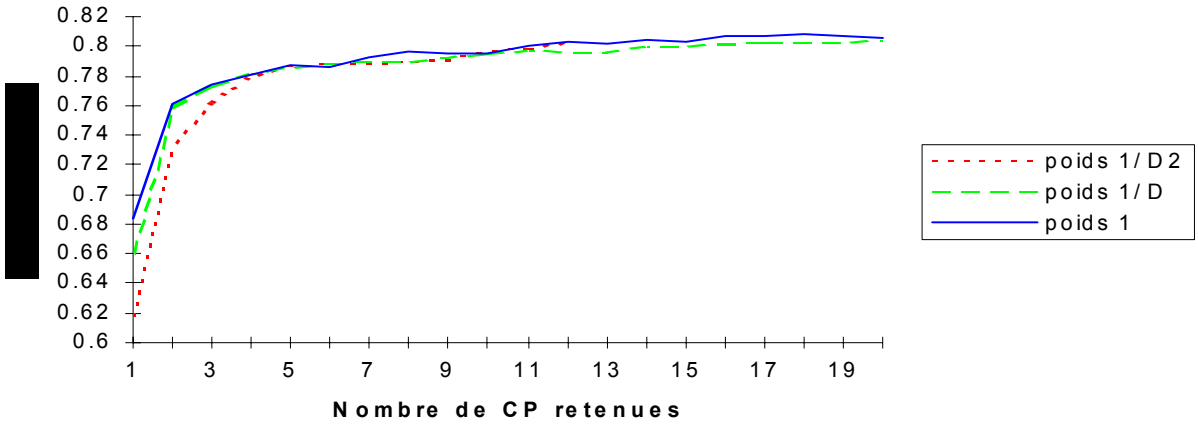
ANNEXE III-14:

Pondération des analogues

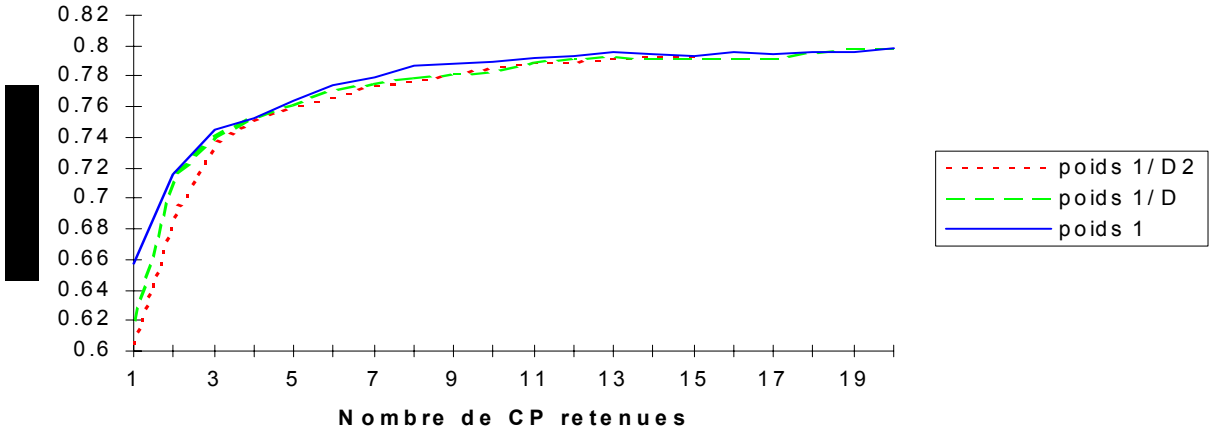




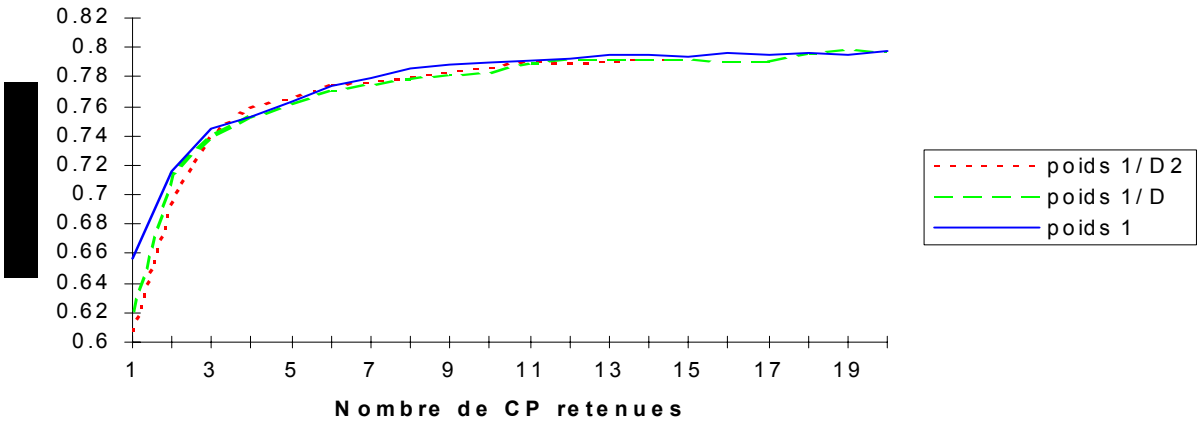
Pondérations des analogues (CP 00h, 24h)
ISERE MOYENNE



VAR-TINEE-ROYA



CHASSEZAC



ANNEXE III-15:

Comparaison des méthodes S-8CP/S-12CP

	groupements	S-8CP	S-12CP	écart
1	Creuse-Cher	76.17	80.22	4.05
2	Vézère-Vienne-Thaurion	69.63	72.37	2.74
3	Dordogne	76.98	80.65	3.67
4	Cère-Maronne	77.89	81.40	3.51
5	Truyère-Lot inférieur	76.09	79.25	3.16
6	Haut Tarn-Haut Lot	76.44	80.03	3.59
7	Agout-Tarn	72.82	75.05	2.23
8	Pyrénées Est	74.91	77.41	2.50
9	Ariège-Vicdessos	72.66	75.93	3.27
10	Pique-Garonne-Salat	75.18	78.42	3.24
11	Gaves	71.86	74.30	2.44
12	Doubs	77.27	79.17	1.90
13	Ain-Valserine	74.7	77.49	2.79
14	Arve-Fier	74	76.31	2.31
15	Isère-Doron	77.91	81.05	3.14
16	Isère moyenne	77.75	79.84	2.09
17	Romanche-Arc inférieur	78.08	80.76	2.68
18	Drac	77.14	79.60	2.46
19	Buech-Drôme	78.1	79.87	1.77
20	Verdon	75.82	79.23	3.41
21	BVI Verdon	76.28	79.01	2.73
22	Var-Tinee-Roya	75.8	79.20	3.40
23	Haute Durance	76.49	79.87	3.38
24	Durance moyenne	76.6	78.99	2.39
25	Mont Cenis	75.31	78.34	3.03
26	Chassezac	75.61	77.89	2.28
27	Loire supérieure	74.19	76.98	2.79
28	Doux-Eyrieux	75.1	78.53	3.43
29	Gard-Cèze	75.85	79.58	3.73
30	Loire moyenne	75.98	79.15	3.17
31	Allier supérieur	75.88	78.42	2.54
32	Sioule	77.35	81.56	4.21
33	Cure	76.57	78.80	2.23
	moyenne	**75.71**	**78.63**	**2.92**

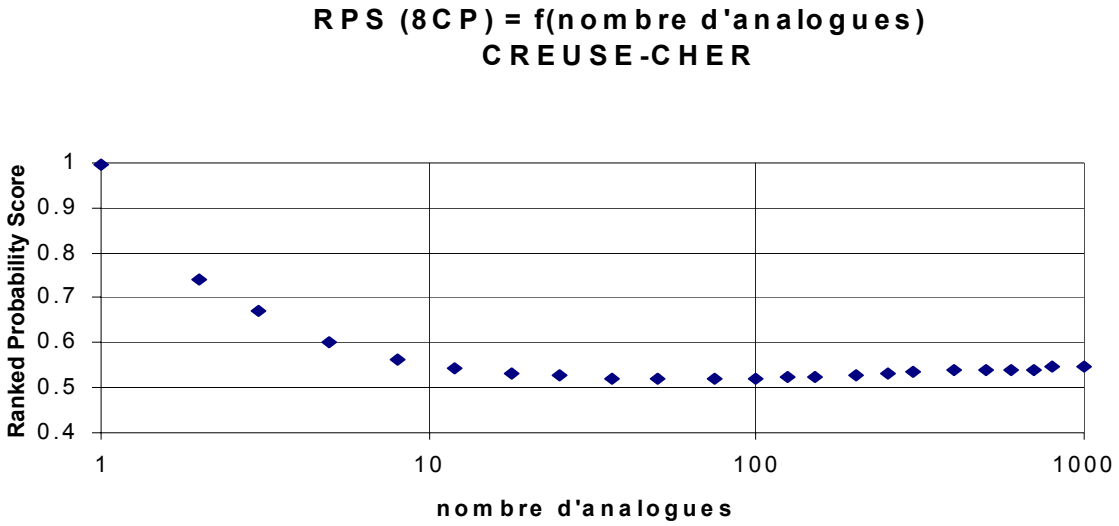
ANNEXE III-16:

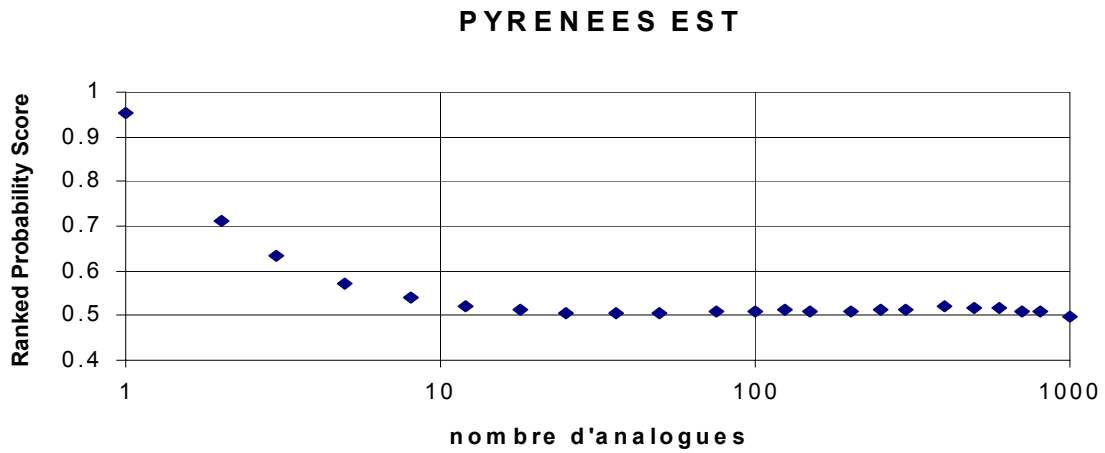
Comparaison sélection (S-12CP)/pondération

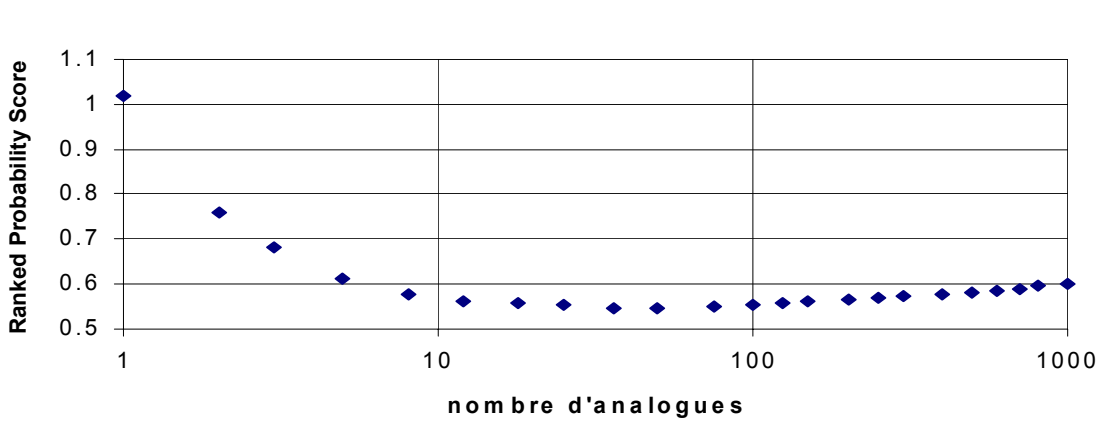
	groupements	S-12CP	pondération	écart
1	Creuse-Cher	80.22	80.70	0.48
2	Vézère-Vienne-Thaurion	72.37	73.09	0.72
3	Dordogne	80.65	81.18	0.53
4	Cère-Maronne	81.40	81.94	0.54
5	Truyère-Lot inférieur	79.25	79.31	0.06
6	Haut Tarn-Haut Lot	80.03	80.92	0.89
7	Agout-Tarn	75.05	75.61	0.56
8	Pyrénées Est	77.41	78.56	1.15
9	Ariège-Vicdessos	75.93	76.71	0.78
10	Pique-Garonne-Salat	78.42	78.77	0.35
11	Gaves	74.30	75.36	1.06
12	Doubs	79.17	79.71	0.54
13	Ain-Valserine	77.49	78.16	0.67
14	Arve-Fier	76.31	76.31	0.00
15	Isère-Doron	81.05	82.04	0.99
16	Isère moyenne	79.84	80.46	0.62
17	Romanche-Arc inférieur	80.76	81.53	0.77
18	Drac	79.60	80.81	1.21
19	Buech-Drôme	79.87	81.08	1.21
20	Verdon	79.23	80.84	1.61
21	BVI Verdon	79.01	80.65	1.64
22	Var-Tinee-Roya	79.20	79.84	0.64
23	Haute Durance	79.87	81.45	1.58
24	Durance moyenne	78.99	79.60	0.61
25	Mont Cenis	78.34	78.58	0.24
26	Chassezac	77.89	79.04	1.15
27	Loire supérieure	76.98	78.24	1.26
28	Doux-Eyrieux	78.53	78.88	0.35
29	Gard-Cèze	79.58	80.46	0.88
30	Loire moyenne	79.15	79.55	0.40
31	Allier supérieur	78.42	79.28	0.86
32	Sioule	81.56	81.96	0.40
33	Cure	78.80	79.04	0.24
	moyenne	**78.63**	**79.38**	**0.76**

ANNEXE III-17:

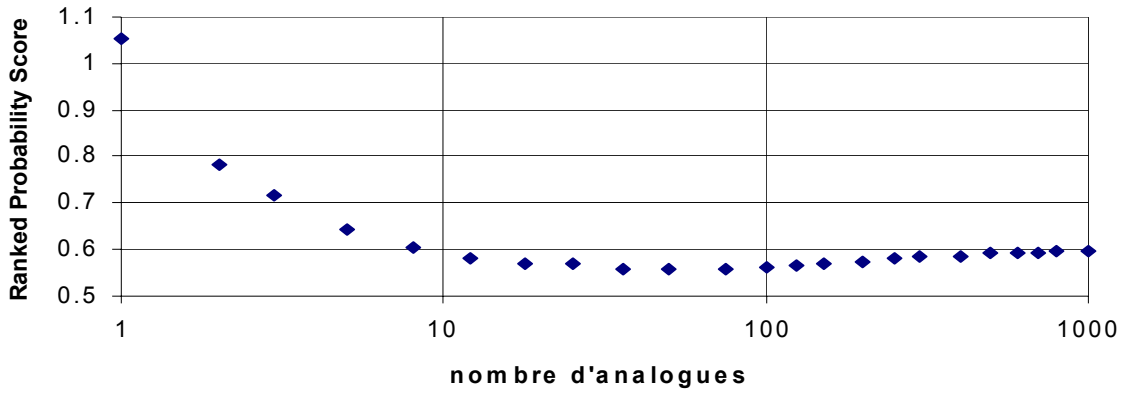
Optimisation du nombre d'analogue (prévision en classes)



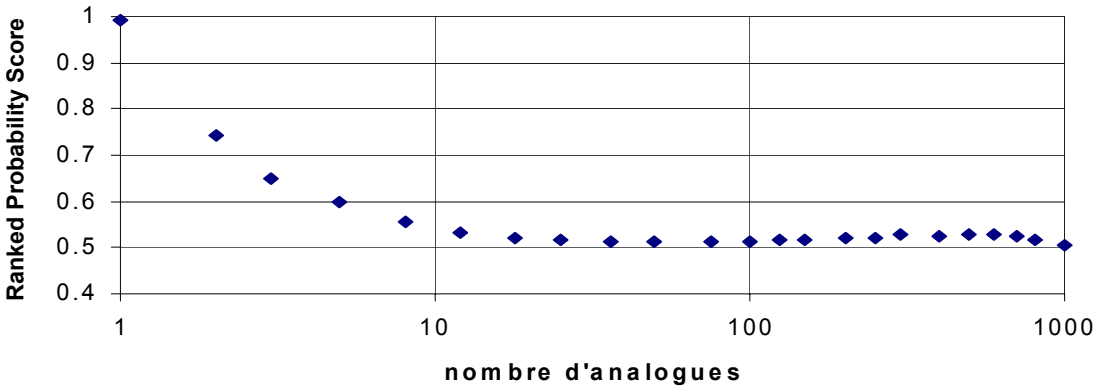




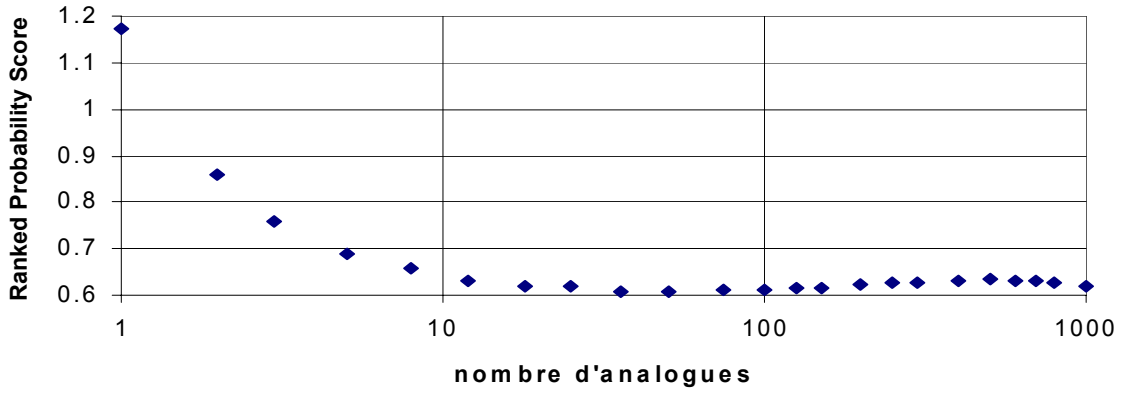
RPS (8CP) = f(nombre d'analogues)
ISERE MOYENNE



VAR-TINEE-ROYA

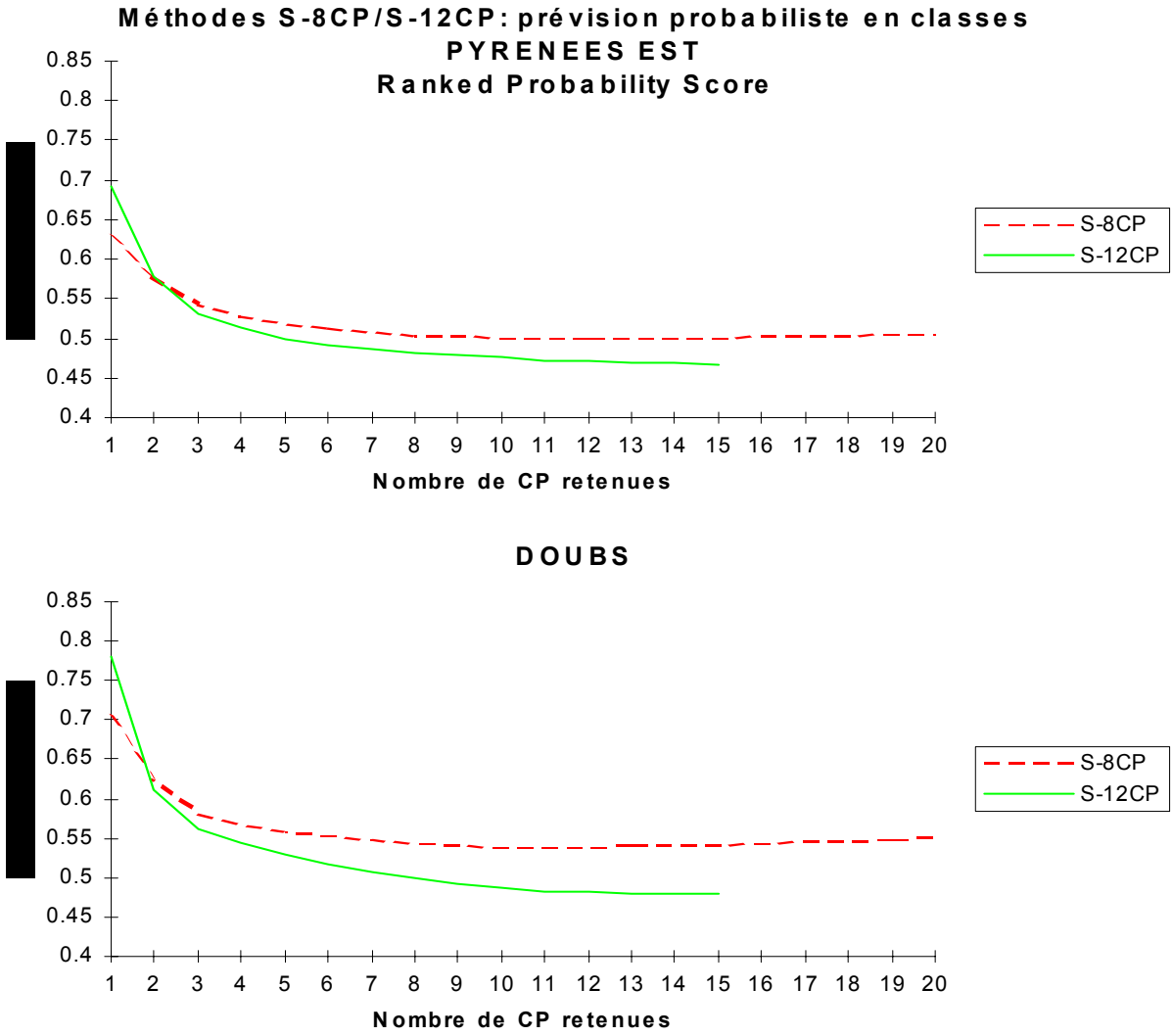


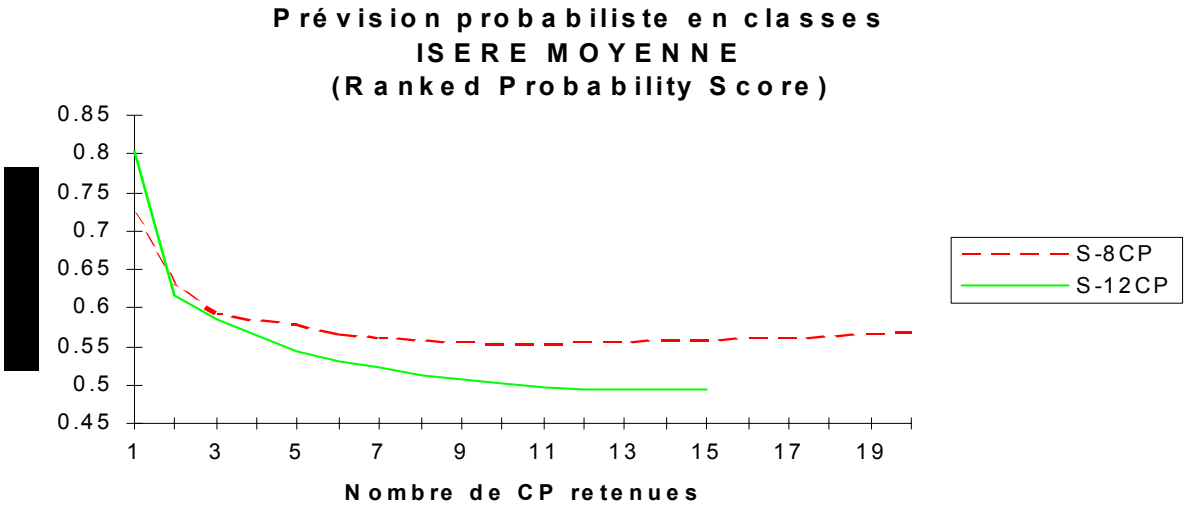
CHASSEZAC

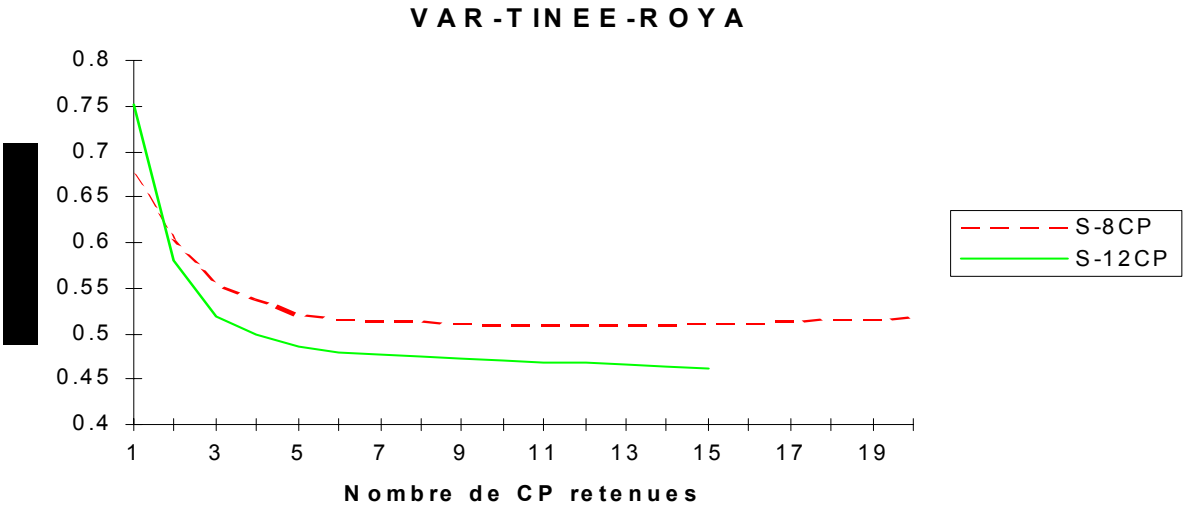


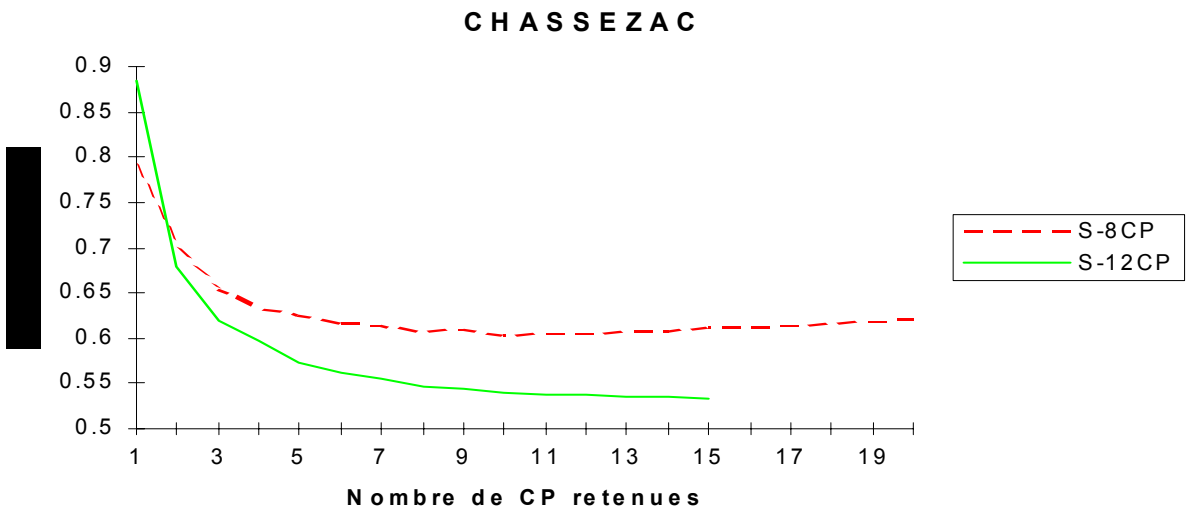
ANNEXE III-18:

Méthodes S-8CP/S-12CP en prévision probabiliste




Prévision probabiliste en classes
ISERE MOYENNE
(Ranked Probability Score)
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
S-8CP
S-12CP


VAR-TINEE-ROYA
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
S-8CP
S-12CP


CHASSEZAC
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1 3 5 7 9 11 13 15 17 19
Nombre de CP retenues
S-8CP
S-12CP

ANNEXE III-19 :

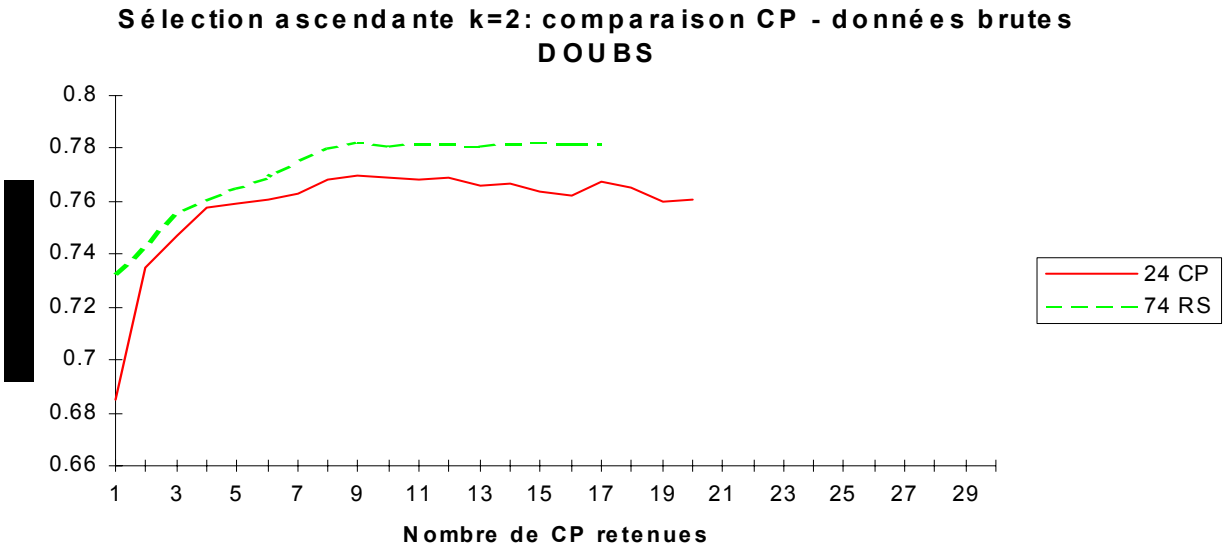
Les différentes méthodes en prévision probabiliste en classes

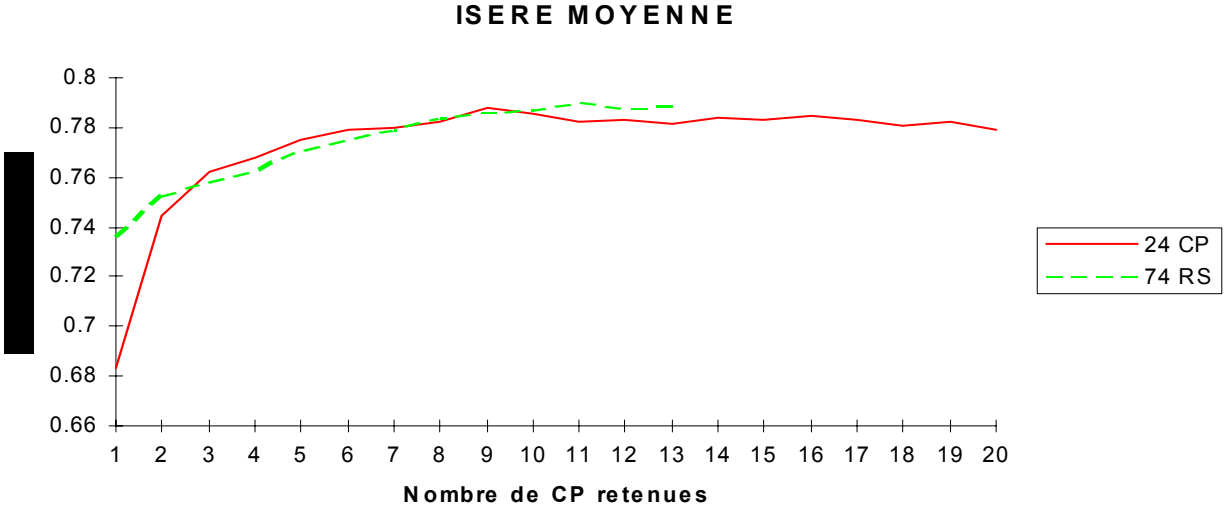
groupements		persistance	clim	hasard	référence	S-12CP	pond
1	Creuse-Cher	0.9400	0.7070	0.7216	0.5696	0.4503	0.4417
2	Vézère-Vienne-Thaurion	1.0260	0.7989	0.8165	0.7186	0.6365	0.6292
3	Dordogne	0.9869	0.7946	0.8099	0.6069	0.4680	0.4578
4	Cère-Maronne	1.0807	0.8557	0.8724	0.6551	0.5060	0.4915
5	Truyère-Lot inférieur	0.9552	0.7457	0.7583	0.5926	0.4710	0.4584
6	Haut Tarn-Haut Lot	0.9748	0.7422	0.7532	0.5905	0.4803	0.4663
7	Agout-Tarn	1.0456	0.8164	0.8306	0.6949	0.5767	0.5615
8	Pyrénées Est	0.9124	0.6478	0.6588	0.5648	0.4706	0.4618
9	Ariège-Vicdessos	1.0308	0.7465	0.7582	0.6509	0.5332	0.5241
10	Pique-Garonne-Salat	1.0710	0.7931	0.8076	0.6668	0.5357	0.5241
11	Gaves	0.9954	0.8103	0.8260	0.6763	0.5409	0.5327
12	Doubs	1.0381	0.7994	0.8125	0.5964	0.4816	0.4715
13	Ain-Valserine	1.0791	0.8583	0.8731	0.6306	0.5191	0.5045
14	Arve-Fier	1.0528	0.8089	0.8174	0.6131	0.5029	0.4889
15	Isère-Doron	0.9263	0.6868	0.6963	0.5279	0.4272	0.4166
16	Isère moyenne	1.0788	0.8257	0.8359	0.6153	0.4941	0.4833
17	Romanche-Arc inférieur	0.9496	0.7046	0.7143	0.5374	0.4574	0.4413
18	Drac	1.0113	0.7459	0.7567	0.5657	0.4691	0.4530
19	Buech-Drôme	0.9539	0.6912	0.7004	0.5334	0.4497	0.4365
20	Verdon	0.9416	0.6838	0.6943	0.5422	0.4510	0.4438
21	BVI Verdon	0.8990	0.6927	0.7032	0.5536	0.4626	0.4502
22	Var-Tinee-Roya	0.9207	0.7180	0.7277	0.5541	0.4671	0.4552
23	Haute Durance	0.8799	0.6648	0.6728	0.5144	0.4327	0.4195
24	Durance moyenne	0.8955	0.6553	0.6664	0.5277	0.4440	0.4323
25	Mont Cenis	0.9555	0.7271	0.7358	0.5846	0.4904	0.4745
26	Chassezac	1.0801	0.8774	0.8918	0.6696	0.5377	0.5240
27	Loire supérieure	1.0421	0.8084	0.8220	0.6293	0.5227	0.5116
28	Doux-Eyrieux	1.0038	0.7497	0.7618	0.5942	0.4891	0.4744
29	Gard-Cèze	0.9799	0.7807	0.7941	0.5977	0.4879	0.4782
30	Loire moyenne	0.9228	0.6925	0.7042	0.5581	0.4603	0.4478
31	Allier supérieur	0.9895	0.7464	0.7610	0.5947	0.4952	0.4839
32	Sioule	0.9327	0.7117	0.7254	0.5677	0.4503	0.4402
33	Cure	0.9954	0.7897	0.8055	0.6032	0.5041	0.4946
moyenne		**0.9863**	**0.7539**	**0.7662**	**0.5969**	**0.4899**	**0.4780**

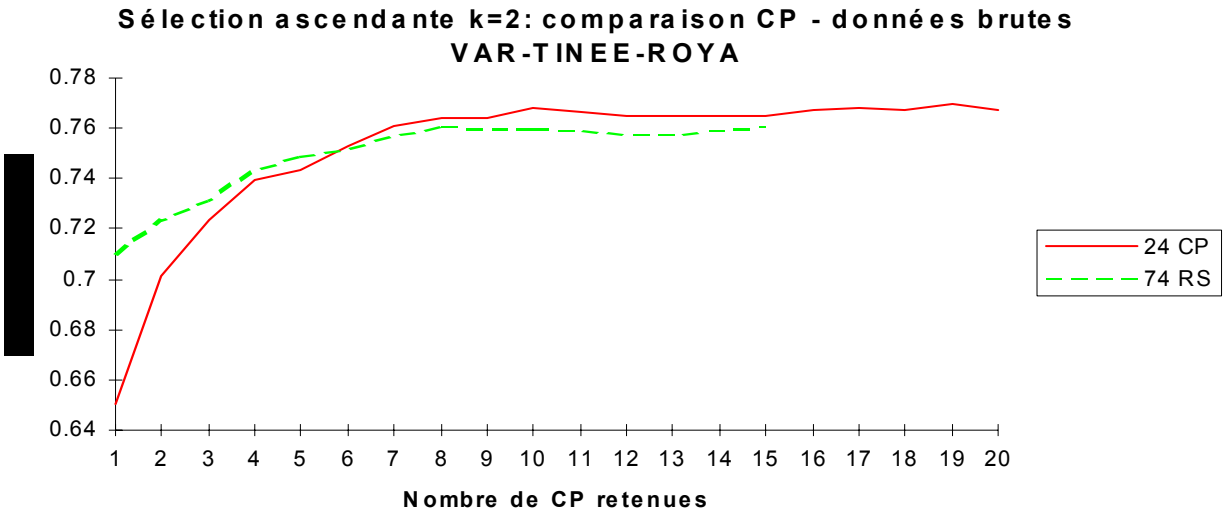
ANNEXES DU CHAPITRE IV

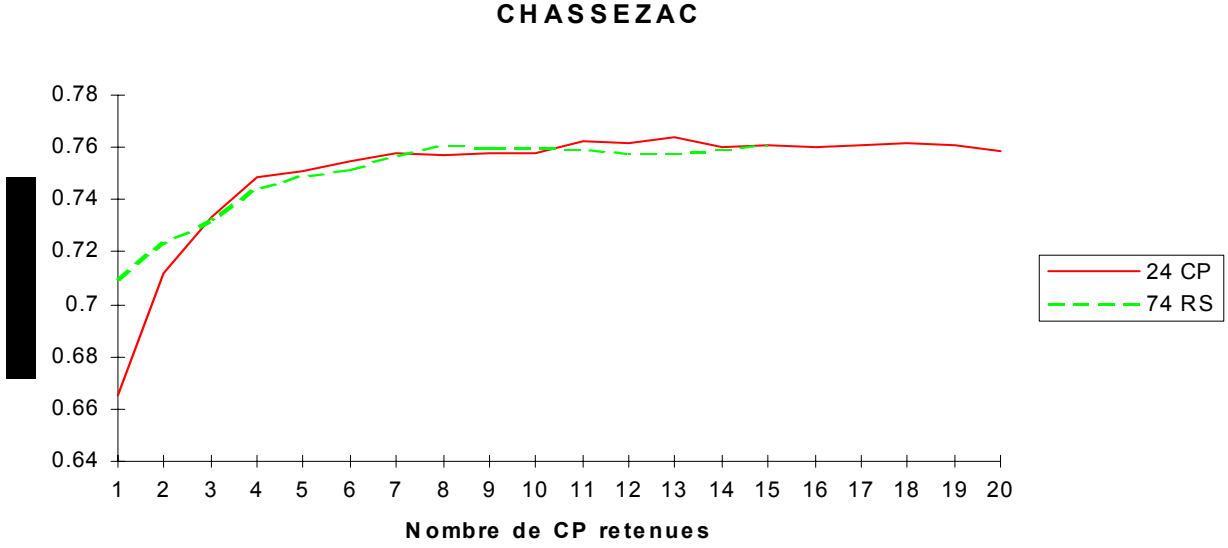
ANNEXE IV-1:

Comparaison Sélection ascendante K=2, CP / RS






Sélection ascendante k=2: comparaison CP - données brutes
VAR-TINEE-ROYA
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de CP retenues
24 CP
74 RS


CHASSEZAC
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de CP retenues
24 CP
74 RS

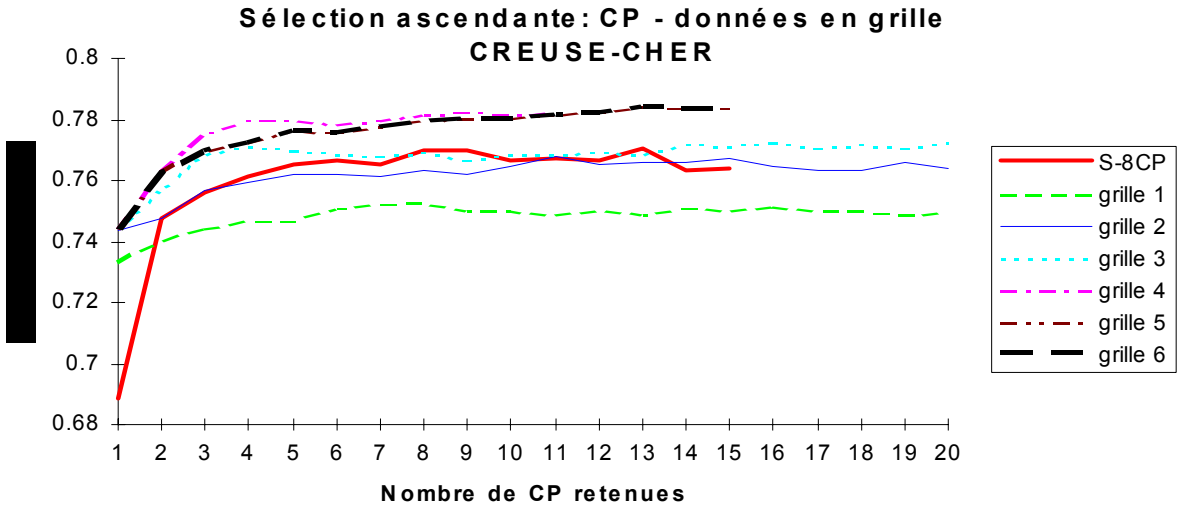
ANNEXE IV-2 :

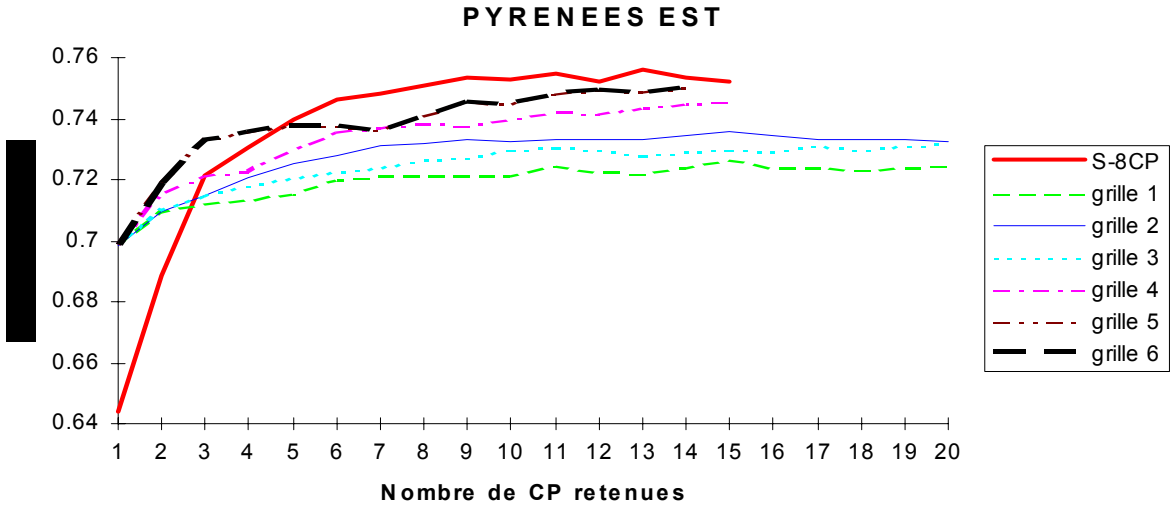
Comparaison des méthodes S-12CP / S-12RS

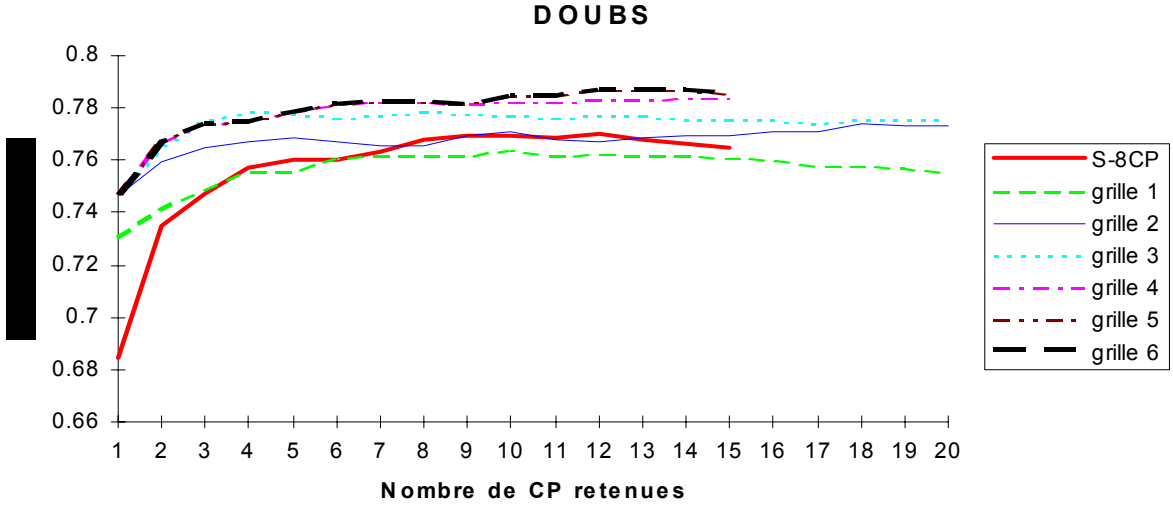
		pluie/non pluie			prévision en classes		
groupements		**S-12CP**	**S-12RS**	**écart**	**S-12CP**	**S-12RS**	**écart**
1	Creuse-Cher	80.22	81.45	1.23	0.4503	0.4388	-0.01
2	Vézère-Vienne-Thaurion	72.37	73.04	0.67	0.6365	0.6285	-0.01
3	Dordogne	80.65	81.56	0.91	0.4680	0.4513	-0.02
4	Cère-Maronne	81.40	81.48	0.08	0.5060	0.4807	-0.03
5	Truyère-Lot inférieur	79.25	79.92	0.67	0.4710	0.4512	-0.02
6	Haut Tarn-Haut Lot	80.03	80.70	0.67	0.4803	0.4693	-0.01
7	Agout-Tarn	75.05	76.74	1.69	0.5767	0.5518	-0.02
8	Pyrénées Est	77.41	76.55	-0.86	0.4706	0.4577	-0.01
9	Ariège-Vicdessos	75.93	75.69	-0.24	0.5332	0.5217	-0.01
10	Pique-Garonne-Salat	78.42	78.32	-0.10	0.5357	0.5211	-0.01
11	Gaves	74.30	76.23	1.93	0.5409	0.5200	-0.02
12	Doubs	79.17	80.68	1.51	0.4816	0.4682	-0.01
13	Ain-Valserine	77.49	78.34	0.85	0.5192	0.5036	-0.02
14	Arve-Fier	76.31	77.00	0.69	0.5029	0.4825	-0.02
15	Isère-Doron	81.05	80.94	-0.11	0.4272	0.4235	0.00
16	Isère moyenne	79.84	80.97	1.13	0.4941	0.4829	-0.01
17	Romanche-Arc inférieur	80.76	81.45	0.69	0.4574	0.4348	-0.02
18	Drac	79.60	78.64	-0.96	0.4691	0.4522	-0.02
19	Buech-Drôme	79.87	79.36	-0.51	0.4497	0.4343	-0.02
20	Verdon	79.23	79.60	0.37	0.4510	0.4343	-0.02
21	BVI Verdon	79.01	79.44	0.43	0.4626	0.4359	-0.03
22	Var-Tinee-Roya	79.20	79.47	0.27	0.4671	0.4403	-0.03
23	Haute Durance	79.87	80.65	0.78	0.4327	0.4098	-0.02
24	Durance moyenne	78.99	78.99	0.00	0.4440	0.4259	-0.02
25	Mont Cenis	78.34	78.18	-0.16	0.4904	0.4702	-0.02
26	Chassezac	77.89	79.44	1.55	0.5377	0.5108	-0.03
27	Loire supérieure	76.98	78.21	1.23	0.5227	0.5066	-0.02
28	Doux-Eyrieux	78.53	78.37	-0.16	0.4891	0.4663	-0.02
29	Gard-Cèze	79.58	81.18	1.60	0.4879	0.4593	-0.03
30	Loire moyenne	79.15	79.92	0.77	0.4603	0.4507	-0.01
31	Allier supérieur	78.42	79.23	0.81	0.4952	0.4736	-0.02
32	Sioule	81.56	82.12	0.56	0.4503	0.4339	-0.02
33	Cure	78.80	80.01	1.21	0.5041	0.4798	-0.02
moyenne		**78.63**	**79.21**	**0.58**	**0.4899**	**0.4719**	**-0.02**

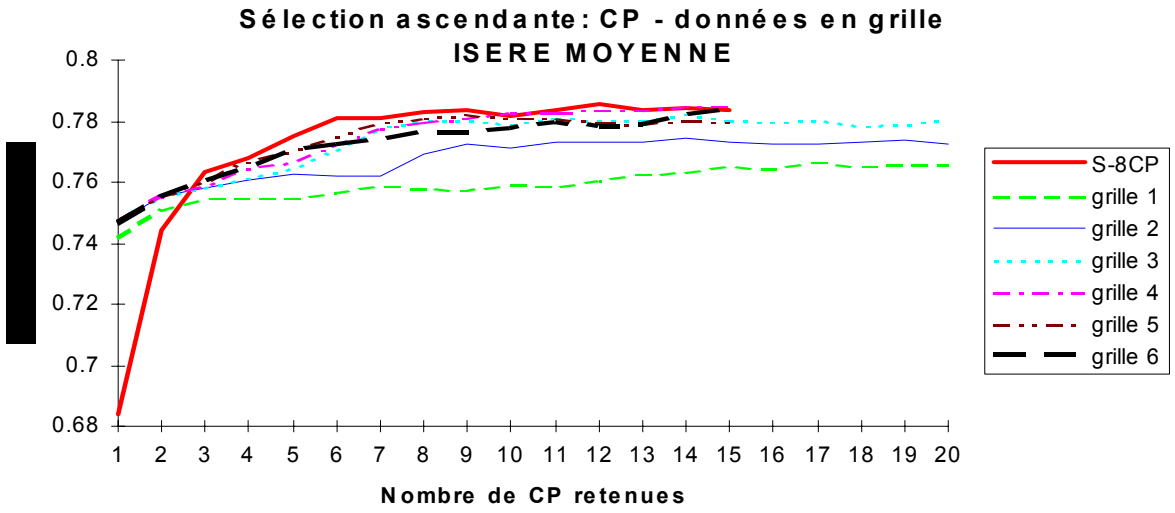
ANNEXE IV-3:

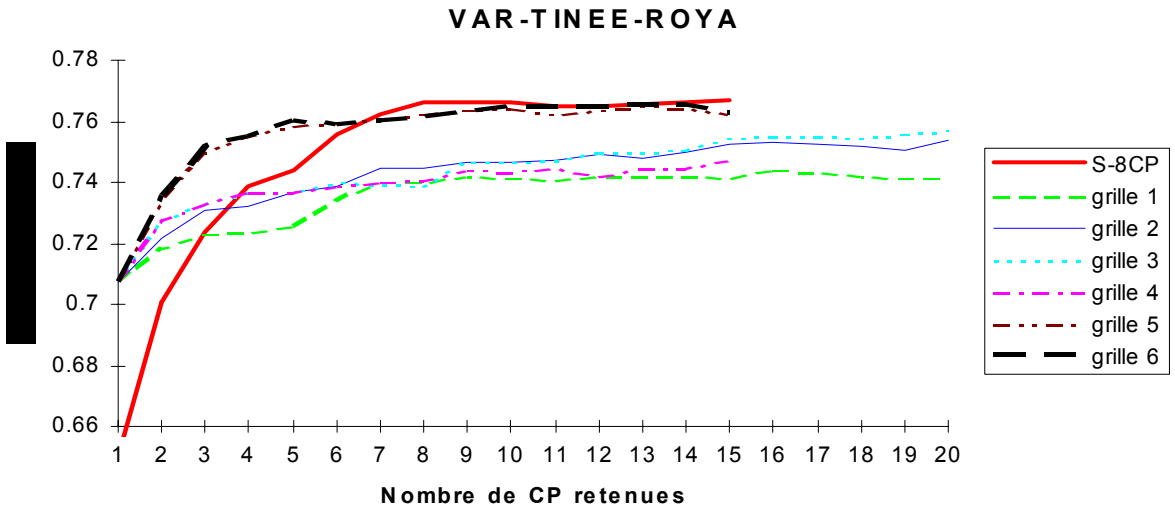
Comparaison sélection ascendante avec CP (S-8CP) et points de grille

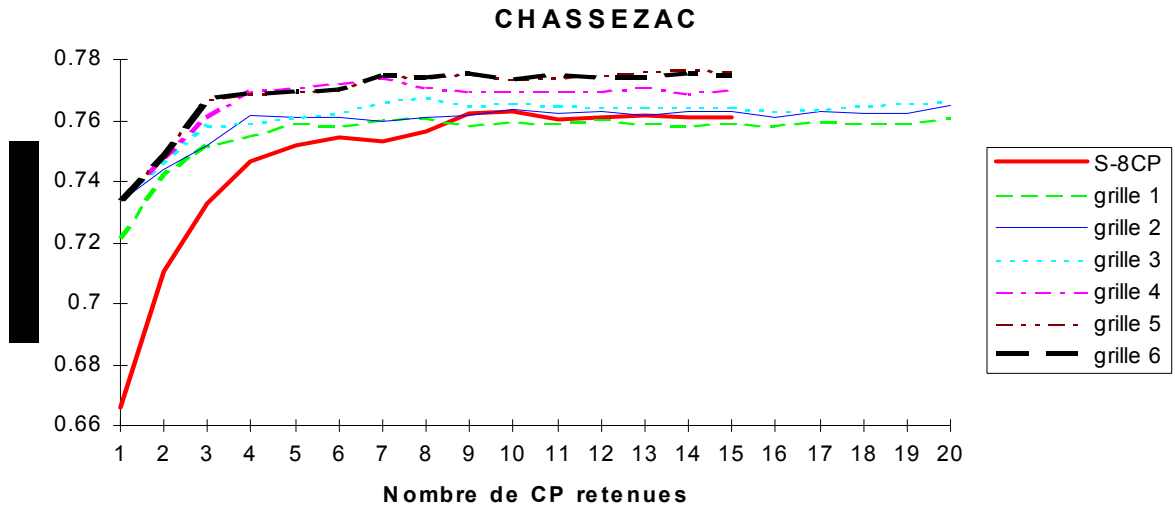






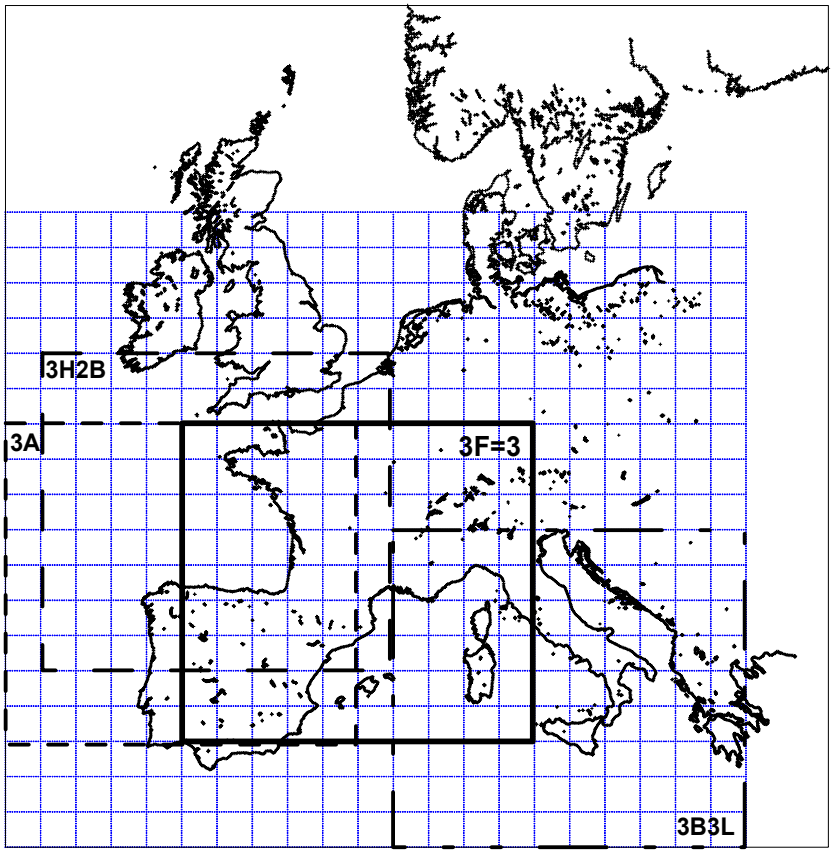

Sélection ascendante: CP - données en grille
ISERE MOYENNE
0.8
0.78
0.76
0.74
0.72
0.7
0.68
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de CP retenues
S-8CP
grille 1
grille 2
grille 3
grille 4
grille 5
grille 6


VAR-TINEE-ROYA
0.78
0.76
0.74
0.72
0.7
0.68
0.66
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de CP retenues
S-8CP
grille 1
grille 2
grille 3
grille 4
grille 5
grille 6


CHASSEZAC
0.78
0.76
0.74
0.72
0.7
0.68
0.66
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Nombre de CP retenues
S-8CP
grille 1
grille 2
grille 3
grille 4
grille 5
grille 6

ANNEXE IV-4 :

Localisation optimale de la grille n°3



	pluie / non pluie IR	classes RPS * 100
3A	77.08	50.45
3B	77.73	49.10
3C	78.57	47.55
3D	79.26	46.18
3E	79.90	45.38
3F = 3	**80.07**	**45.05**
3G	79.96	45.12
3H	79.45	45.67
3I	79.06	46.21
3J	78.26	47.54
3K	77.64	49.76
3L	77.11	49.77

	pluie / non pluie IR	classes RPS * 100
H1A	76.95	
H1B	77.59	
H1C	78.44	
H1D	79.05	46.98
H1E	79.70	46.01
H1F	**79.84**	**45.68**
H1G	79.75	45.78
H1H	79.38	46.22
H1I	78.78	46.85
H1J	78.13	
H1K	77.45	
H1L	76.97	

	pluie / non pluie IR	classes RPS * 100
B1A		
B1B		
B1C		
B1D	79.09	46.26
B1E	79.52	45.51
B1F	**79.66**	**45.24**
B1G	79.56	45.28
B1H	79.17	45.81
B1I	78.72	46.54
B1J		
B1K		
B1L		

	pluie / non pluie IR	classes RPS * 100
H2A	76.49	
H2B	77.21	
H2C	77.94	
H2D	78.69	47.90
H2E	79.17	47.09
H2F	**79.39**	**46.74**
H2G	79.27	46.77
H2H	78.69	47.34
H2I	78.15	47.86
H2J	77.40	
H2K	76.79	
H2L	76.31	

	pluie / non pluie IR	classes RPS * 100
B2A		
B2B		
B2C		
B2D	79.08	46.10
B2E	79.34	45.55
B2F	**79.38**	**45.39**
B2G	79.19	45.59
B2H	78.88	46.14
B2I	78.50	46.74
B2J		
B2K		
B2L		

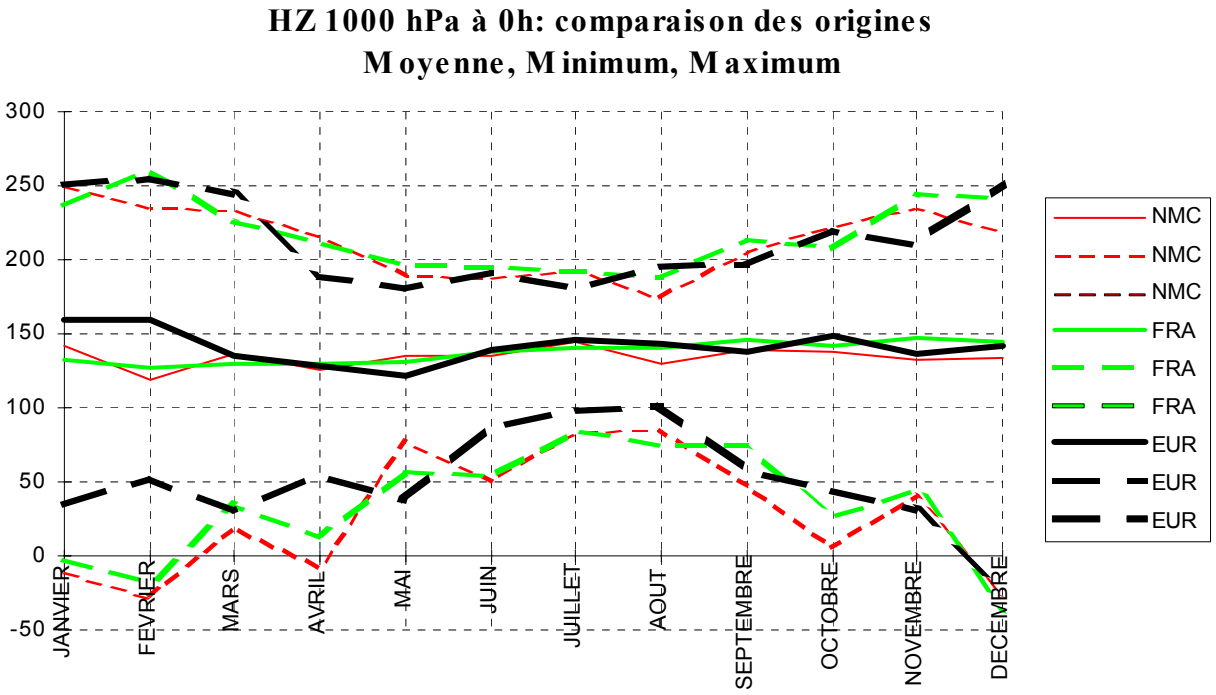
	pluie / non pluie IR	classes RPS * 100
H3A	75.98	
H3B	76.68	
H3C	77.29	
H3D	77.79	49.40
H3E	78.46	48.49
H3F	**78.67**	**48.05**
H3G	78.54	48.09
H3H	77.92	48.55
H3I	77.47	49.12
H3J		
H3K		
H3L		

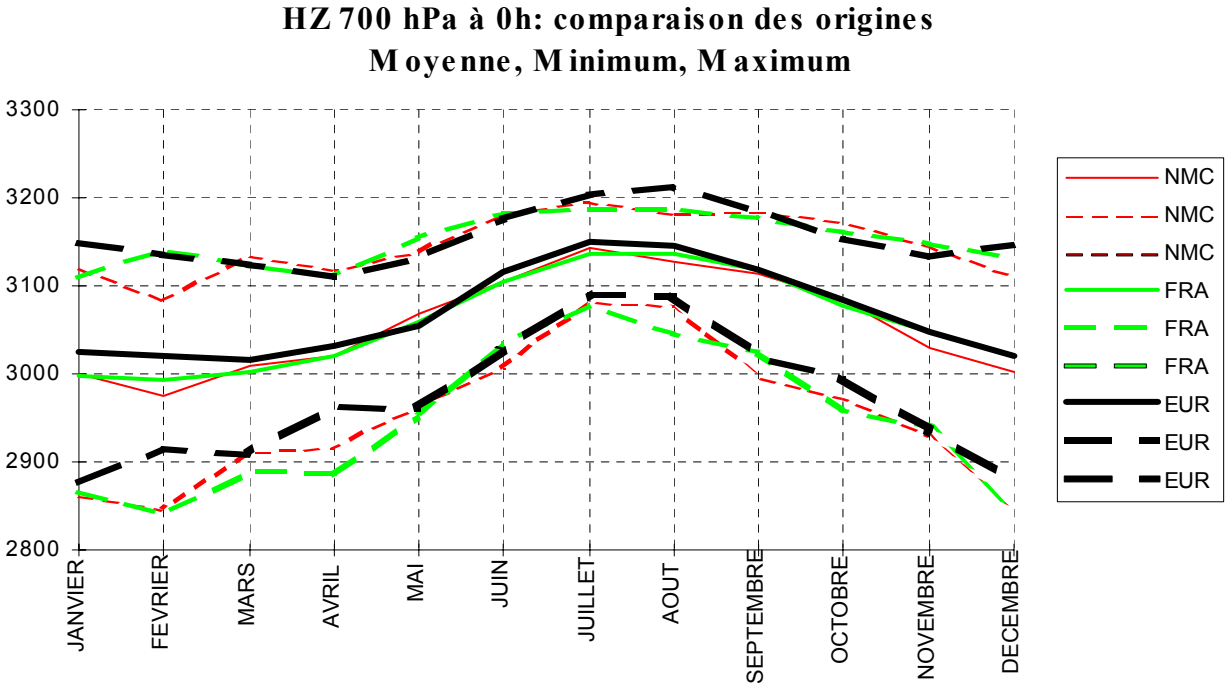
	pluie / non pluie IR	classes RPS * 100
B3A		
B3B		
B3C		
B3D	78.29	47.04
B3E	78.69	46.41
B3F	**78.77**	**46.15**
B3G	78.63	46.31
B3H	78.28	46.92
B3I	77.86	47.53
B3J		
B3K		
B3L		

ANNEXES DU CHAPITRE V

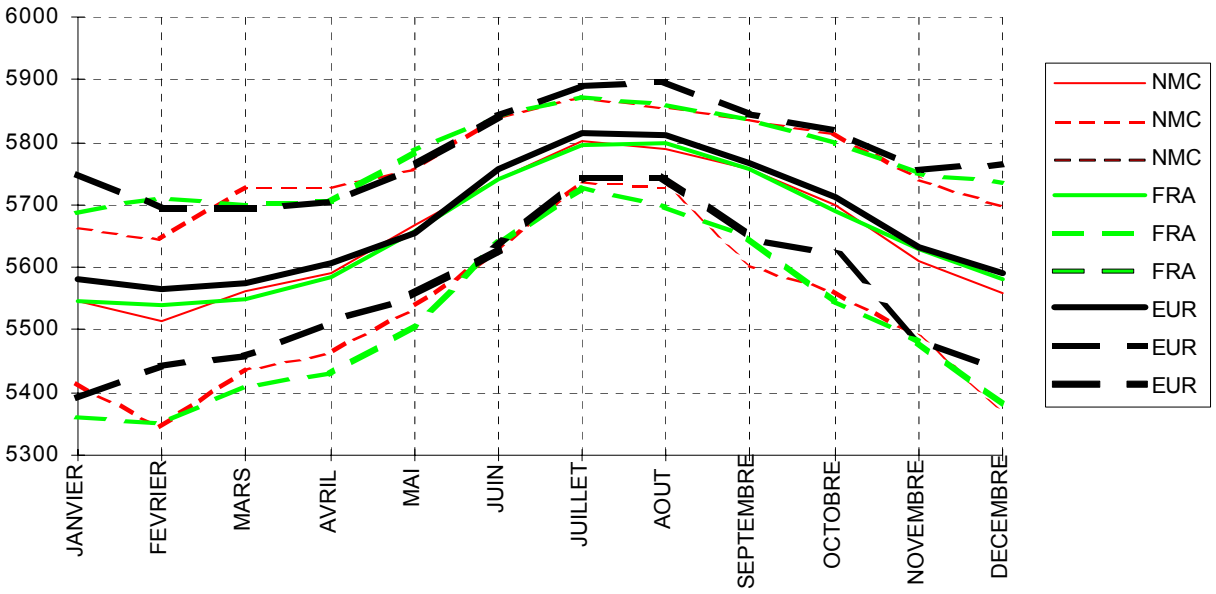
ANNEXE V-1 :

Moyennes mensuelles des différents champs à 00 h pour les 3 types d'archive

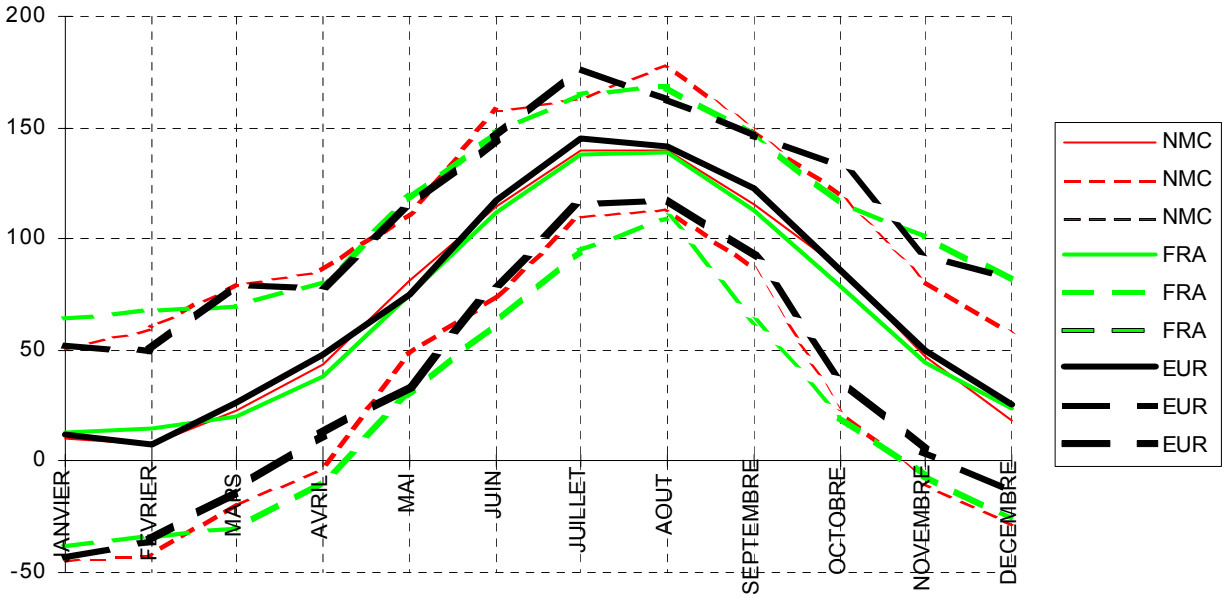


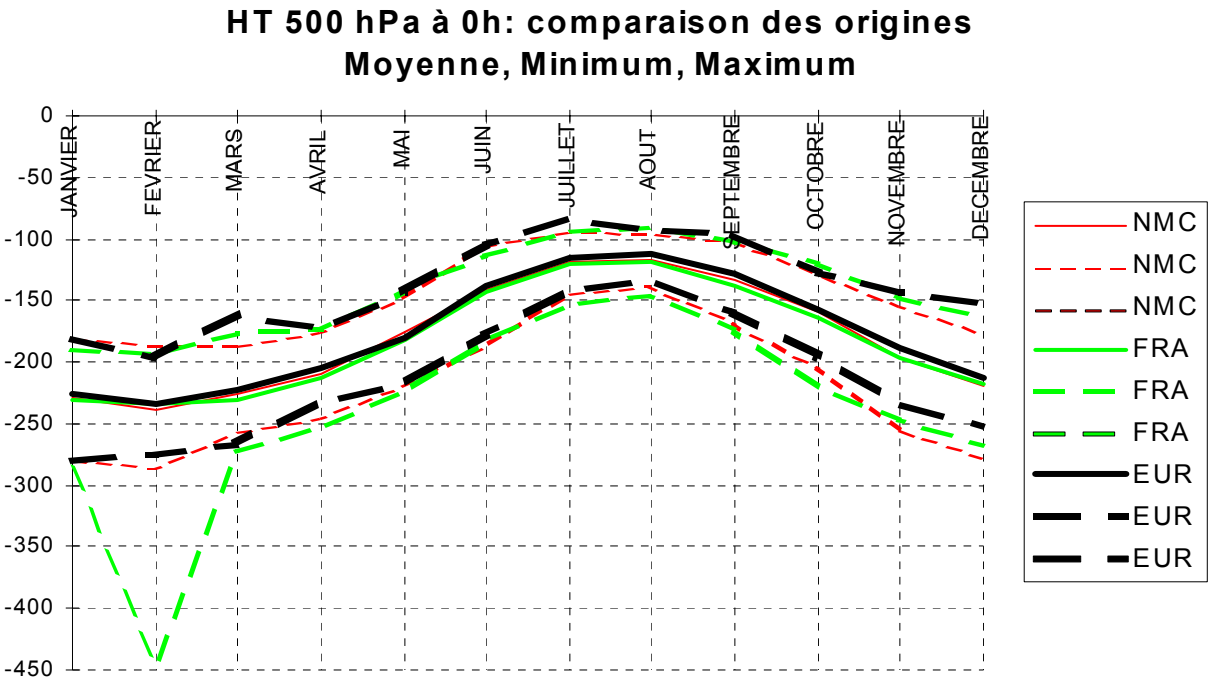


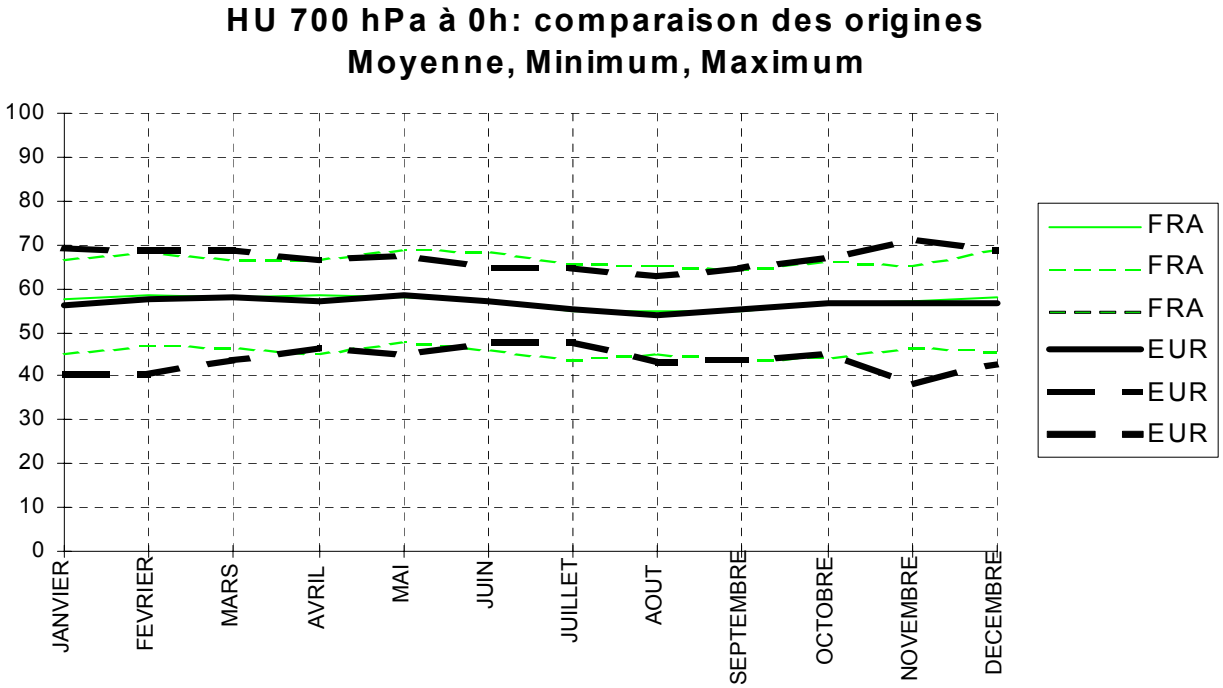
HZ 500 hPa à 0h: comparaison des origines
Moyenne, Minimum, Maximum

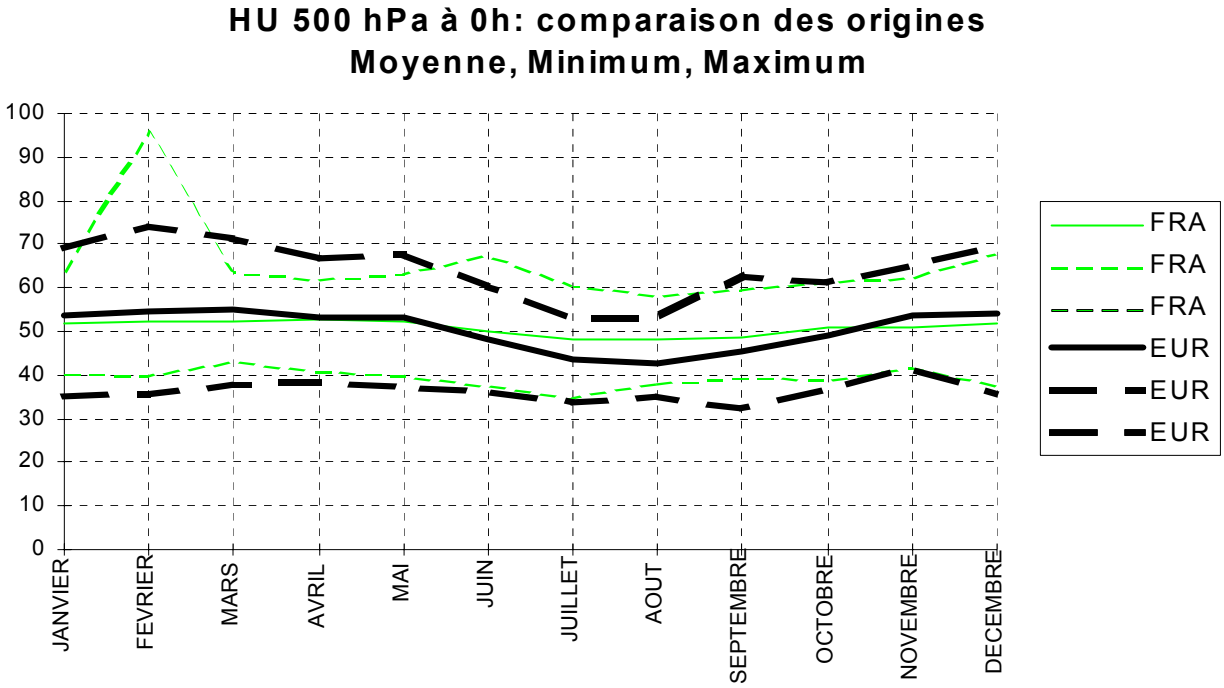


HT 850 hPa à 0h: comparaison des origines
Moyenne, Minimum, Maximum




HT 500 hPa à 0h: comparaison des origines
Moyenne, Minimum, Maximum
0
-50
-100
-150
-200
-250
-300
-350
-400
-450
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
NMC
NMC
NMC
FRA
FRA
FRA
EUR
EUR
EUR


HU 700 hPa à 0h: comparaison des origines
Moyenne, Minimum, Maximum
100
90
80
70
60
50
40
30
20
10
0
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
FRA
FRA
FRA
EUR
EUR
EUR


HU 500 hPa à 0h: comparaison des origines
Moyenne, Minimum, Maximum
100
90
80
70
60
50
40
30
20
10
0
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
FRA
FRA
FRA
EUR
EUR
EUR

ANNEXE V-2:

Qualité des données synoptiques

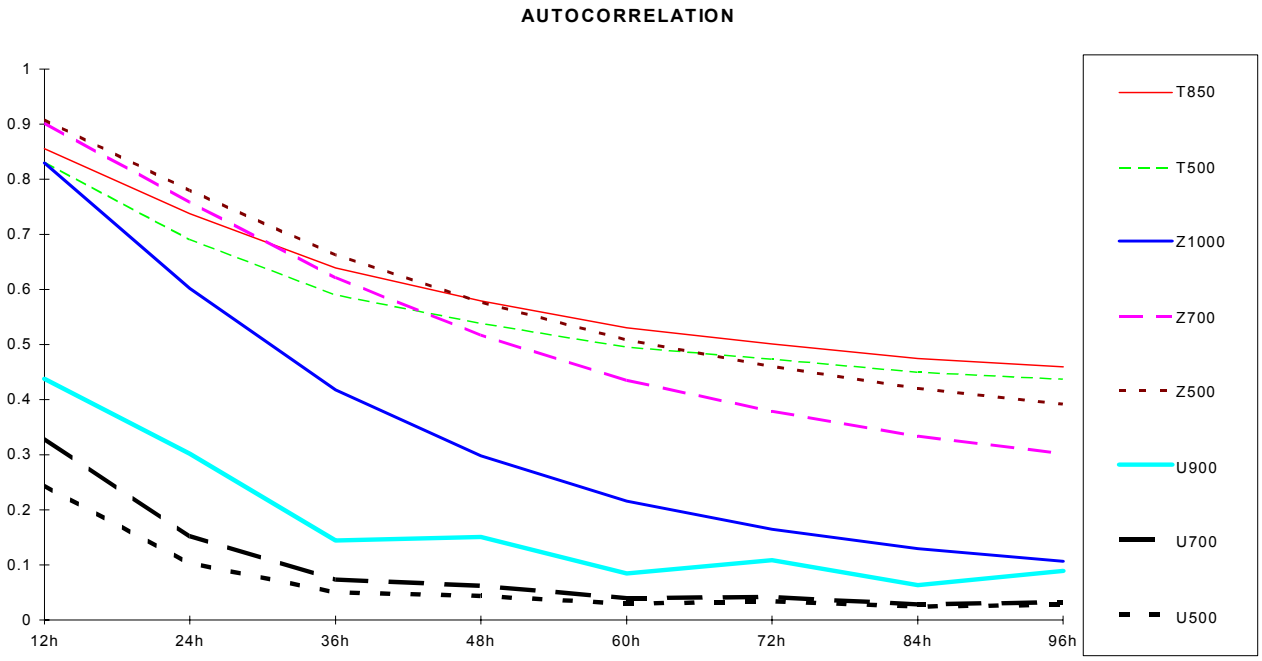
Qualité des données de Météo-France (en %):

	correcte	douteuse	fausse
Z1000 00H	99.0	0	1.0
12H	99.0	0	1.0
Z700 00H	98.6	0	1.4
12H	99.0	0	1.0
Z500 00H	97.8	0	2.2
12H	98.6	0	1.4
T850 00H	98.9	0	1.1
12H	99.2	0	0.8
T500 00H	99.1	0	0.9
12H	99.3	0	0.7
U900 00H	61.3	35.0	3.7
12H	59.3	35.3	5.4
U700 00H	72.3	27.6	0.1
12H	70.5	28.2	1.3
U500 00H	79.3	20.6	0.1
12H	78.1	20.7	1.1

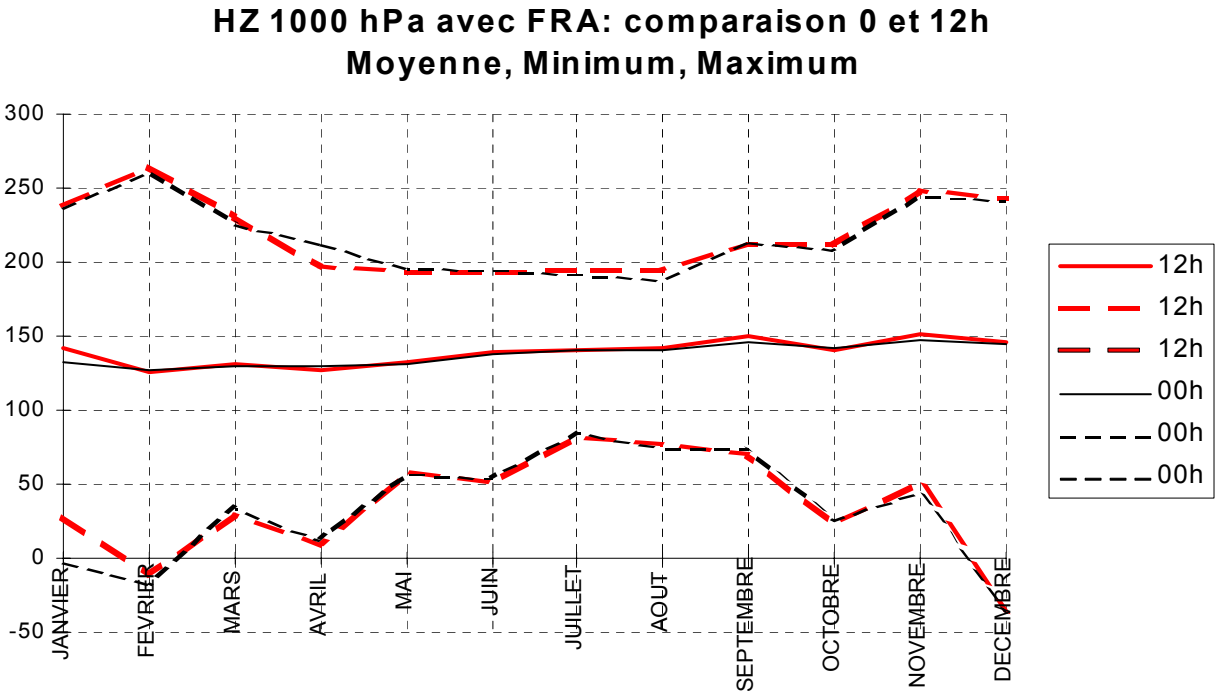
ANNEXE V-3:

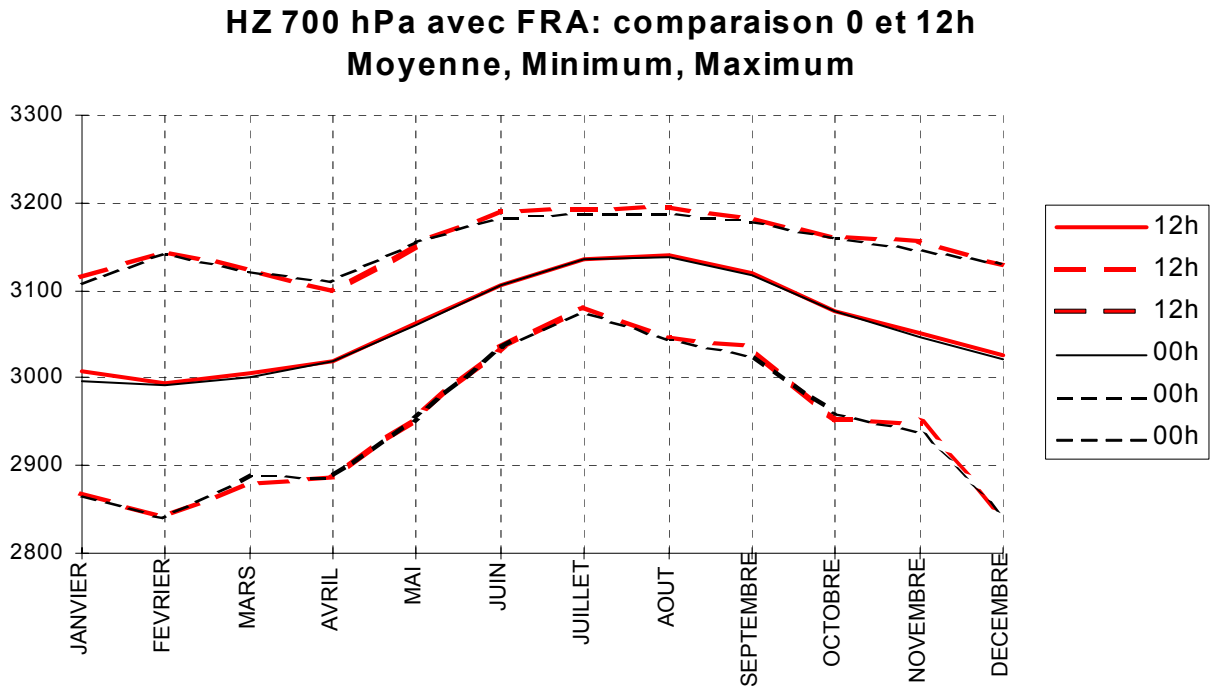
Champs moyens d'automne

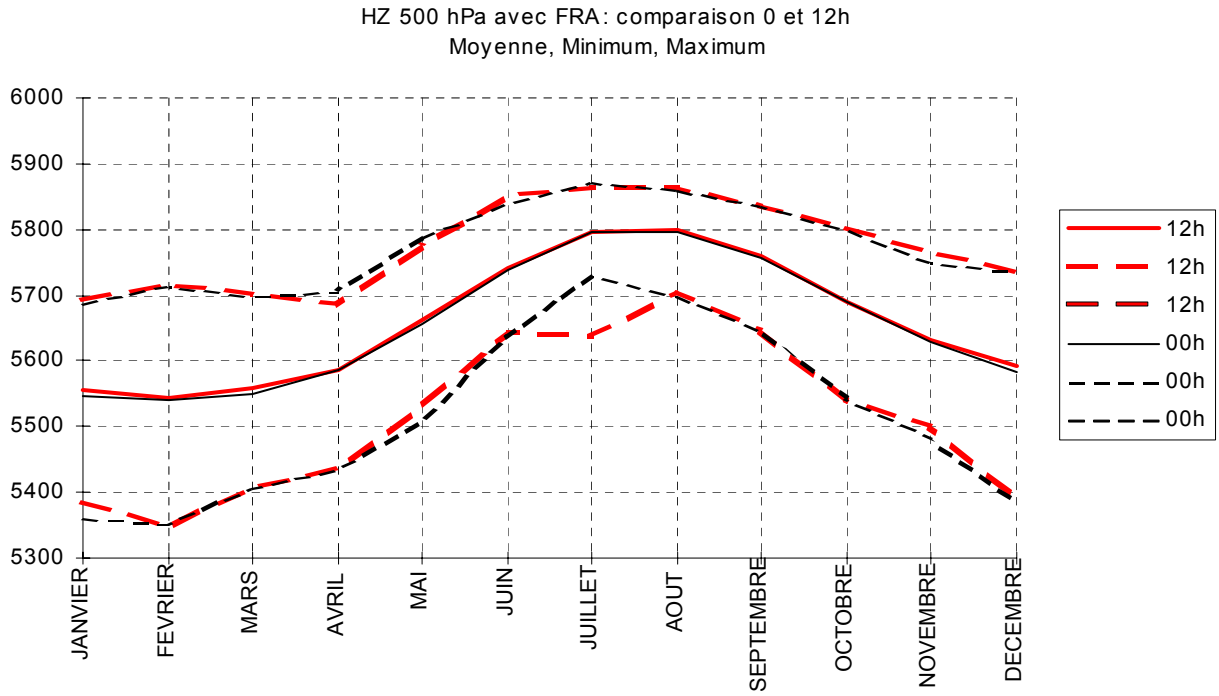
a) Autocorrélation temporelle

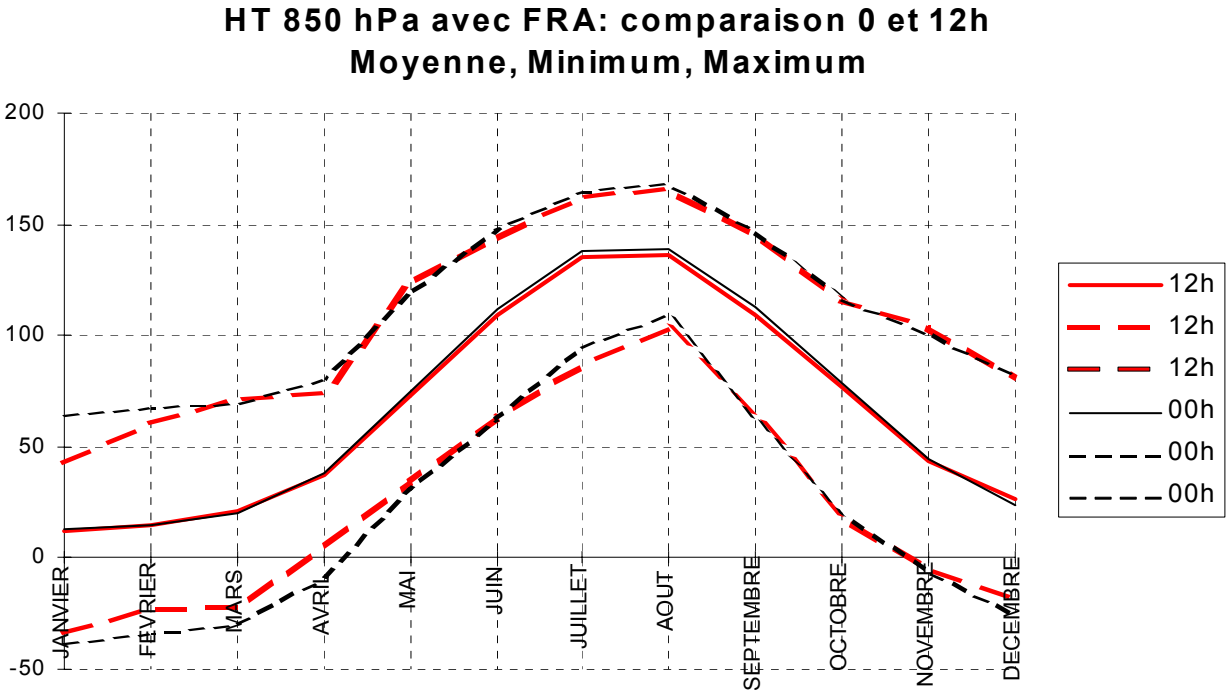


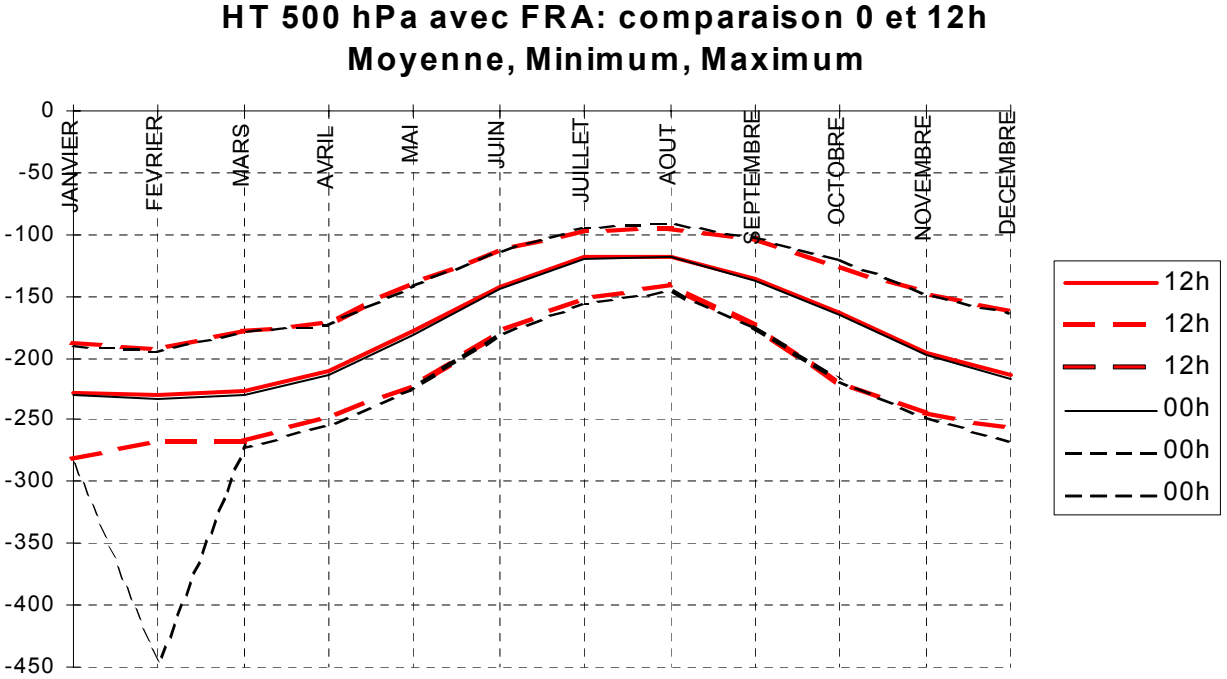
b) effet diurne (exemple sur archives françaises)

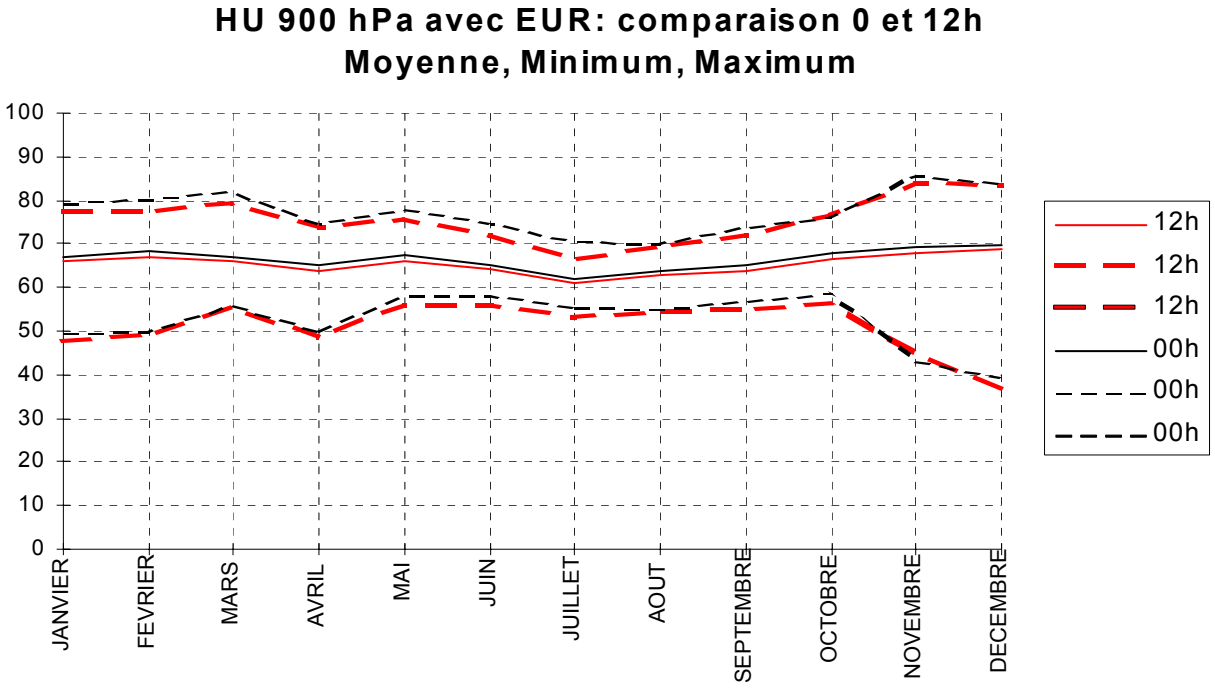


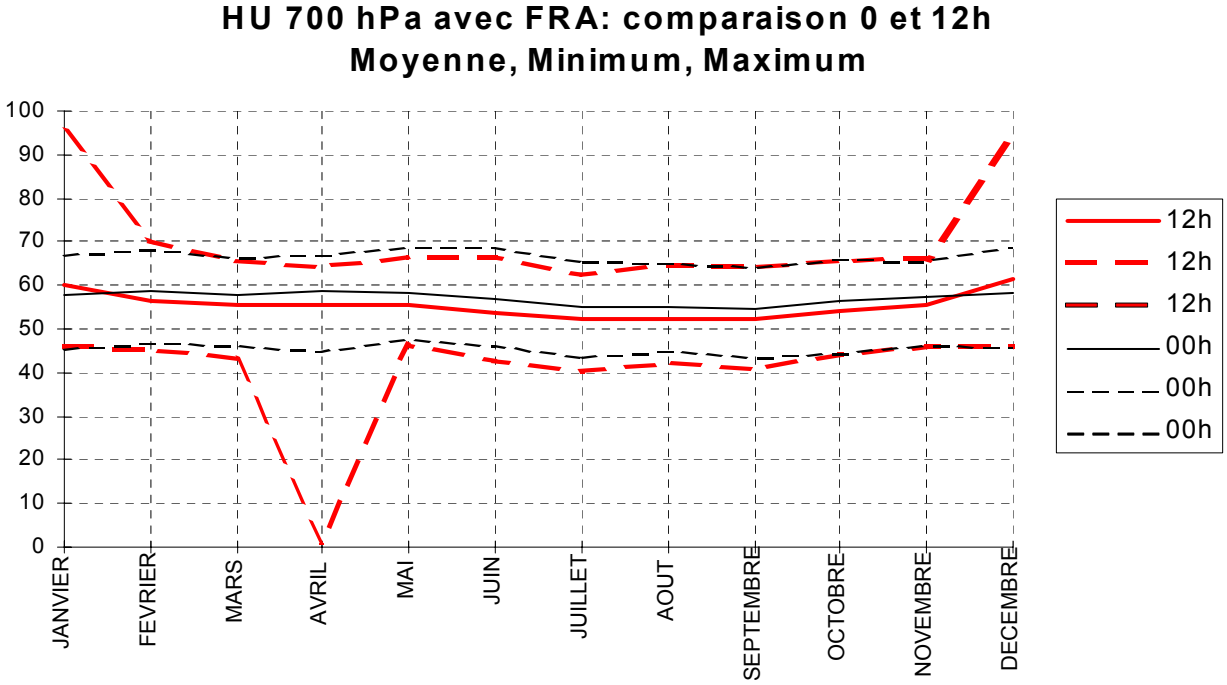

HZ 700 hPa avec FRA: comparaison 0 et 12h
Moyenne, Minimum, Maximum
3300
3200
3100
3000
2900
2800
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h

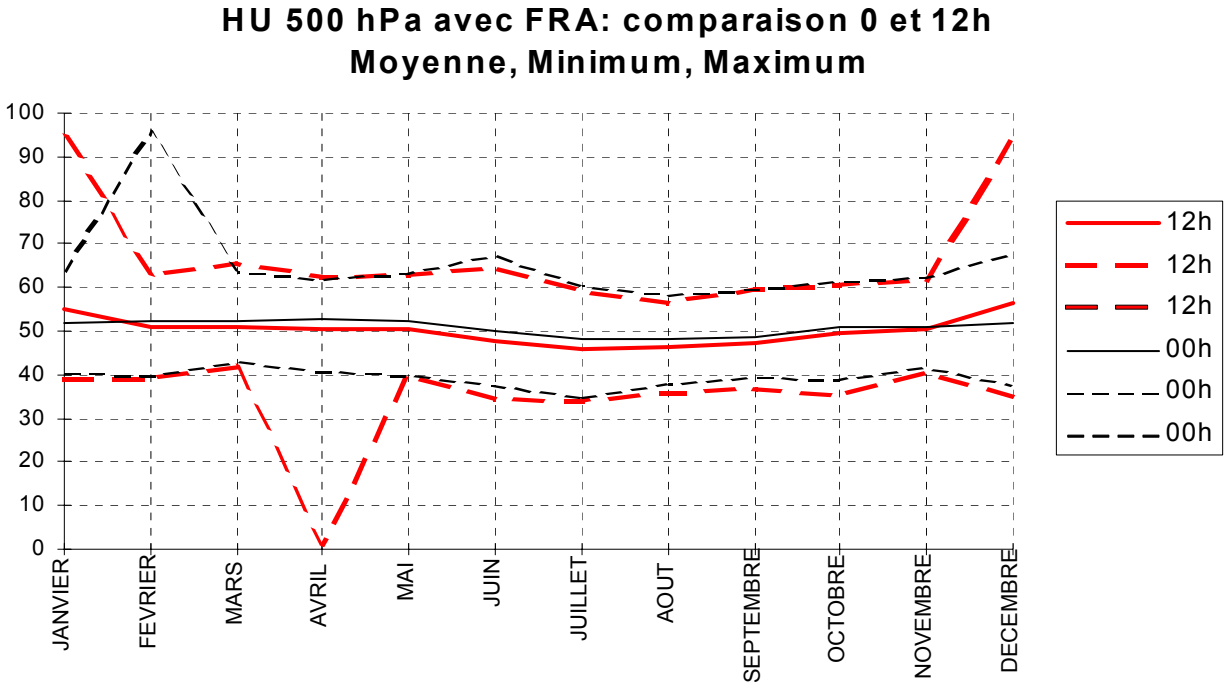

HZ 500 hPa avec FRA: comparaison 0 et 12h
Moyenne, Minimum, Maximum
6000
5900
5800
5700
5600
5500
5400
5300
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h


HT 850 hPa avec FRA: comparaison 0 et 12h
Moyenne, Minimum, Maximum
200
150
100
50
0
-50
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h


HT 500 hPa avec FRA: comparaison 0 et 12h
Moyenne, Minimum, Maximum
0
-50
-100
-150
-200
-250
-300
-350
-400
-450
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h


HU 900 hPa avec EUR: comparaison 0 et 12h
Moyenne, Minimum, Maximum
100
90
80
70
60
50
40
30
20
10
0
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h

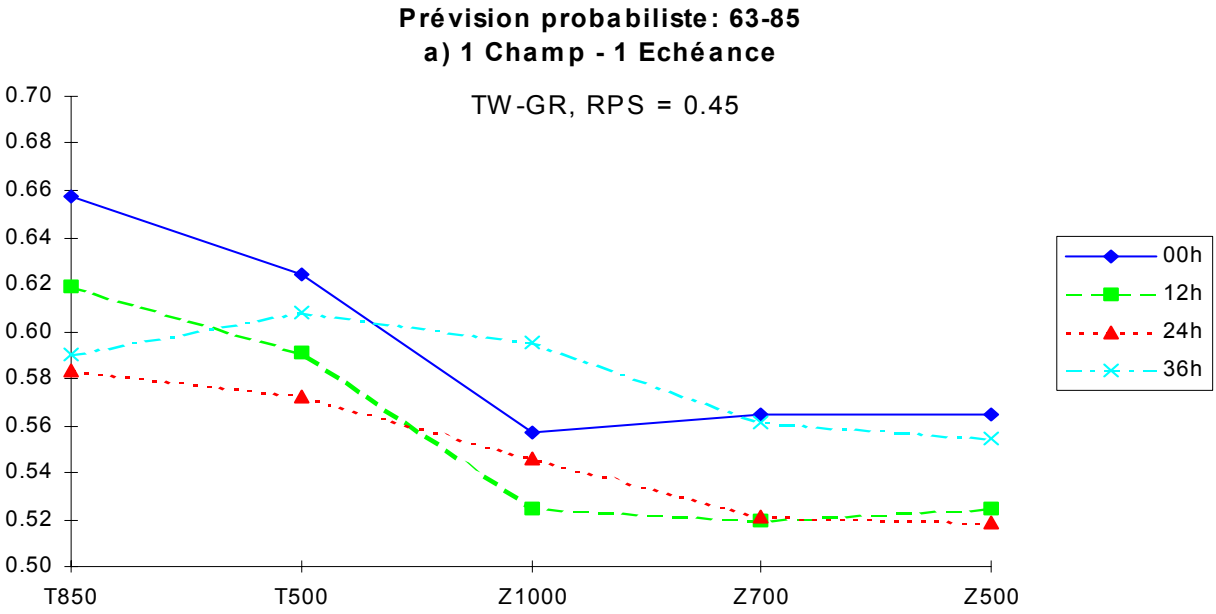

HU 700 hPa avec FRA: comparaison 0 et 12h
Moyenne, Minimum, Maximum
100
90
80
70
60
50
40
30
20
10
0
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h

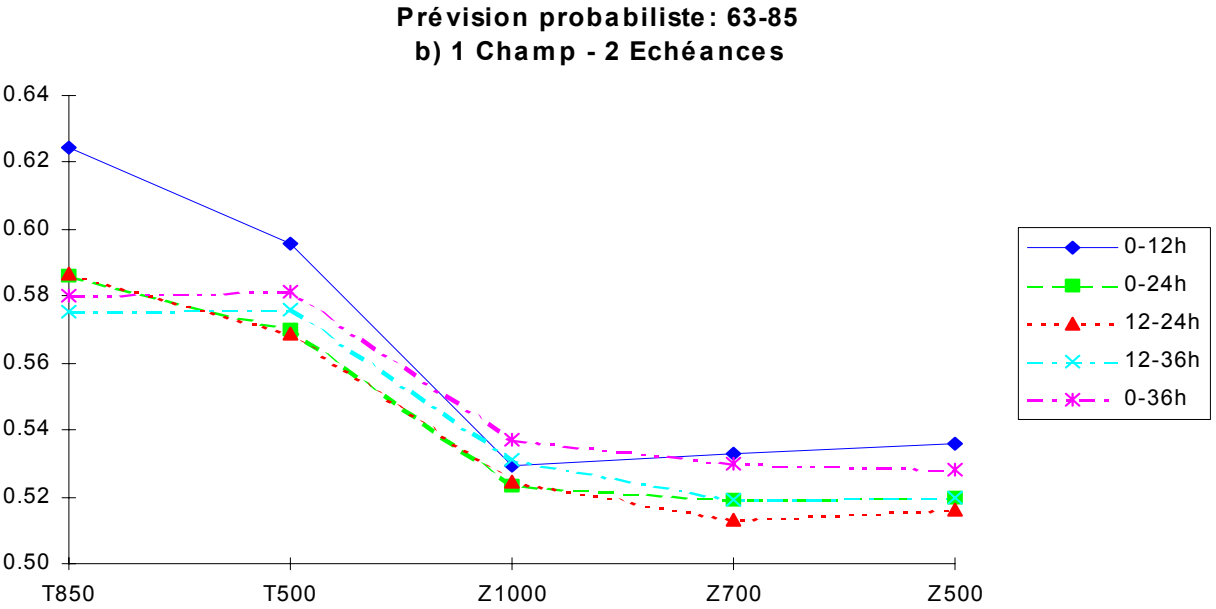

HU 500 hPa avec FRA: comparaison 0 et 12h
Moyenne, Minimum, Maximum
100
90
80
70
60
50
40
30
20
10
0
JANVIER
FEVRIER
MARS
AVRIL
MAI
JUIN
JUILLET
AOUT
SEPTEMBRE
OCTOBRE
NOVEMBRE
DECEMBRE
12h
12h
12h
00h
00h
00h

ANNEXE V-4:

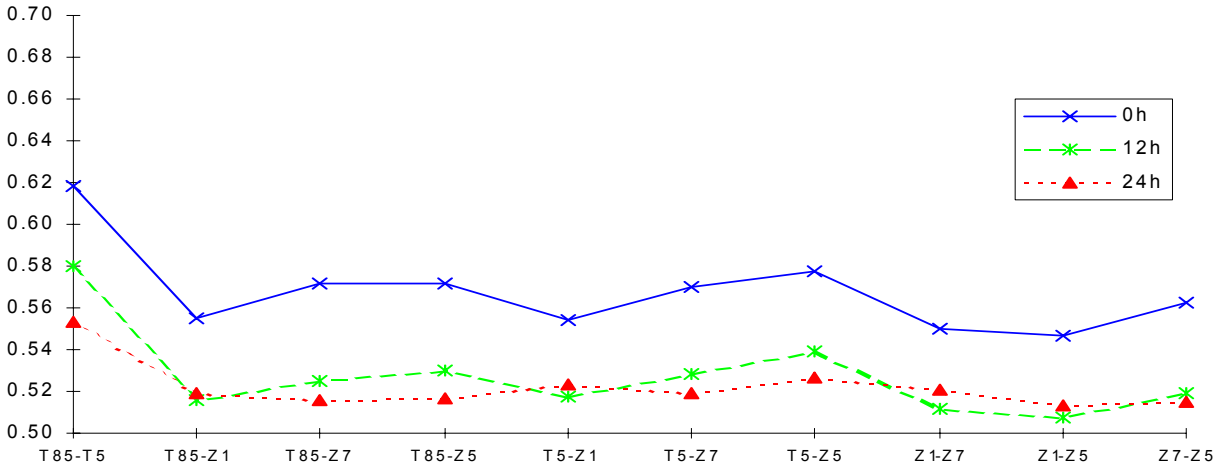
Utilisation des champs synoptiques complémentaires

Période 1963-1985

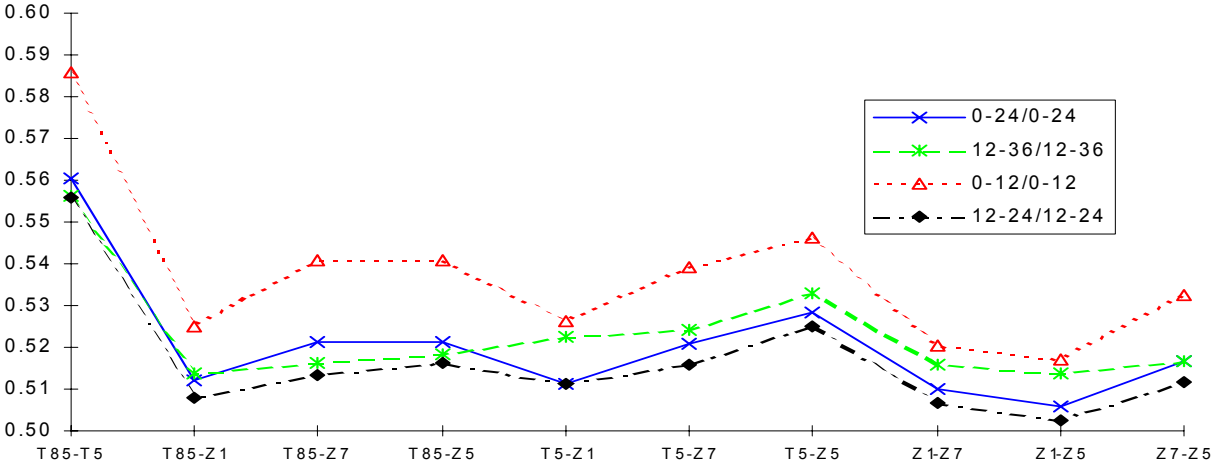




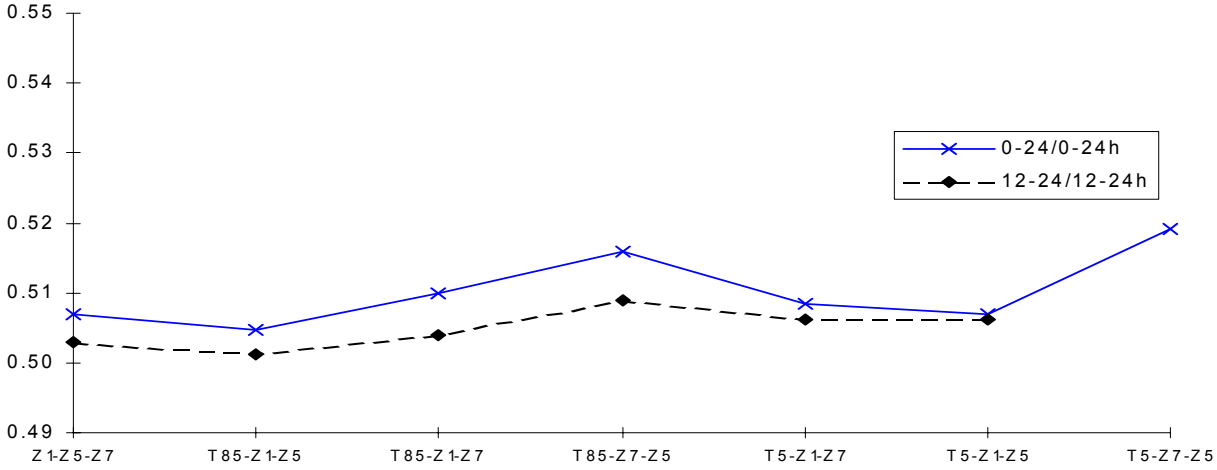
Prévision probabiliste: 63-85
c) 2 Champs - 1 Echéances



Prévision probabiliste: 63-85
d) 2 Champs - 2 Echéances



Prévision probabiliste: 63-85
e) 3 Champs - 2 Echéances



ANNEXE V-5 :

Les données de Nîmes

a) les données brutes

Le rapport de mélange à saturation (cf. Triplet & Roche, 1971):

Considérons une masse m_h d'air humide, occupant un volume v à la température T et sous la pression P, alors $m_h = m_a + m_v$ avec m_a masse d'air sec

et m_v masse de vapeur d'eau.

Le rapport de mélange, en g/g est défini de la façon suivant:

$$r = \frac{m_a}{m_v} \tag{1}$$

Il peut aussi s'écrire:

$$r = 0.622\frac{e}{P-e} \tag{2}$$

avec e, pression partielle de vapeur d'eau

et P-e=P_a, pression partielle de l'air sec

Maintenant, si on apporte de la vapeur d'eau à cette particule, son rapport de mélange r va augmenter. Cependant cette augmentation de r est limitée. En effet, quand la vapeur est saturante dans une particule, celle-ci ne peut plus en absorber davantage sans changer d'état.

On dit alors que l'air est saturé de vapeur d'eau. Son rapport de mélange est alors:

$$\boxed{r = 0.622\frac{e_w(T)}{P-e_w(T)} = r_w(T)} \tag{3}$$

avec $e_w(T)$, pression de vapeur saturante ou pression maximale de vapeur à la température T.

Cette valeur particulière de r s'appelle le *rapport de mélange à saturation*, $r_w(T)$.

Pour l'humidité relative (en %), la définition est la suivante, avec les notations précédentes:

$$\boxed{U = 100\frac{e}{e_w(T)}} \tag{4}$$

si l'air est sec, e = 0 donc U = 0%,

si l'air est saturé à la température T, la vapeur d'eau est saturante et e = $e_w(T)$ donc U = 100 %.

b) Indices d'instabilité

L'énergie convective potentielle disponible ou CAPE (pour Convective Available Potential Energy) est un paramètre quantitatif de l'instabilité convective. Il a éré défini par Moncrieff et Miller en 1976 comme l'intégrale des forces de flottabilité entre les niveaux de convection libre (Lfc) et d'équilibre thermique pour une adiabatique humide représentative de la couche convectivement instable (cf. figure 1).

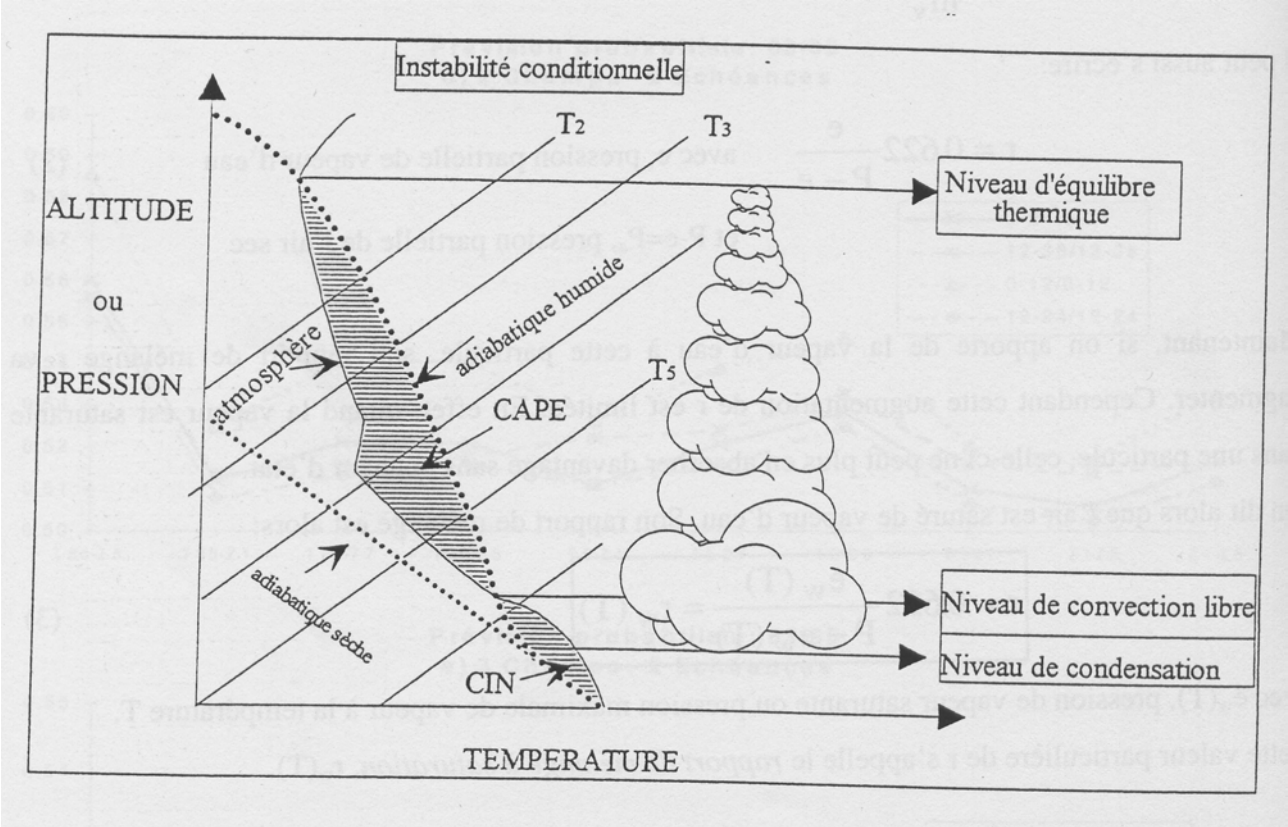


figure 1: Instabilité conditionnelle et CAPE (d'après Roux, 1991)

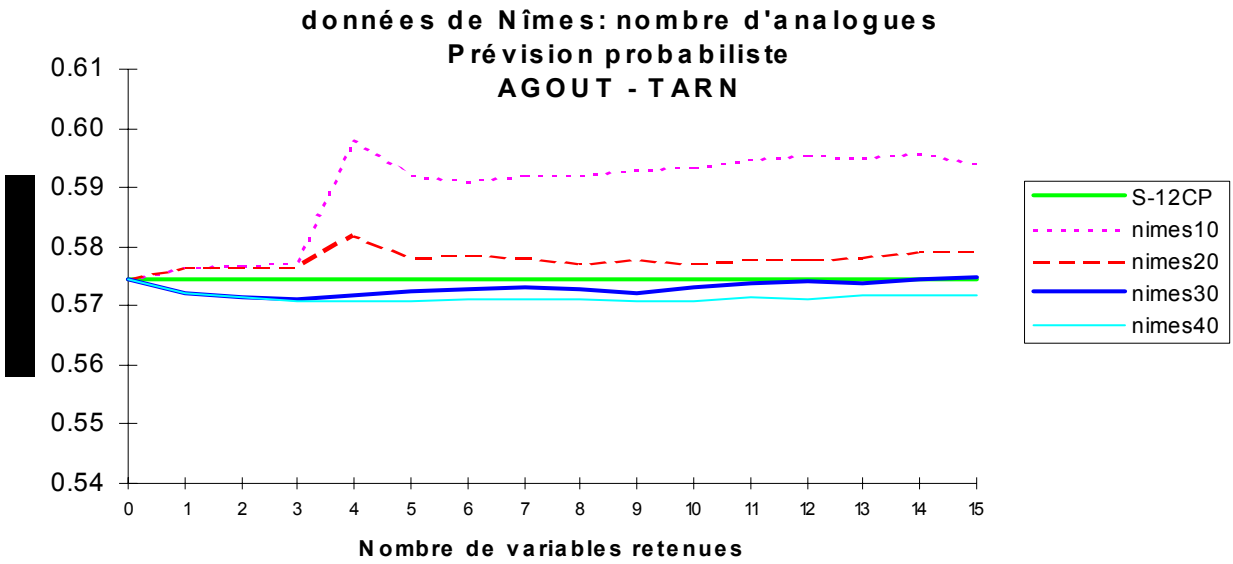
ANNEXE V-6:

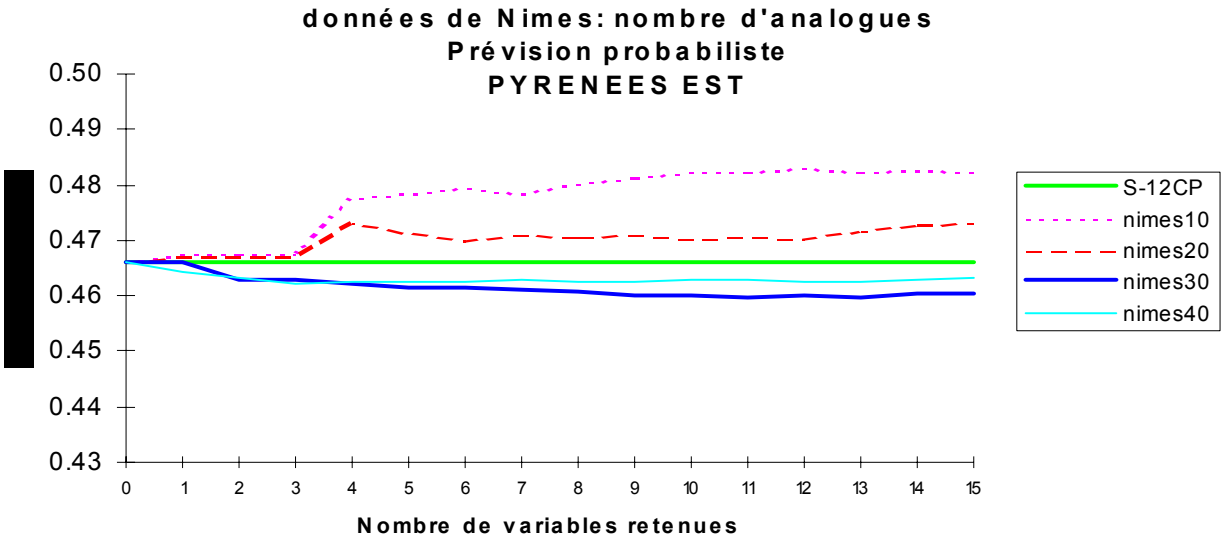
Comparaison des performances 1953-1993 et 1954-1983 pour la méthode S-12CP

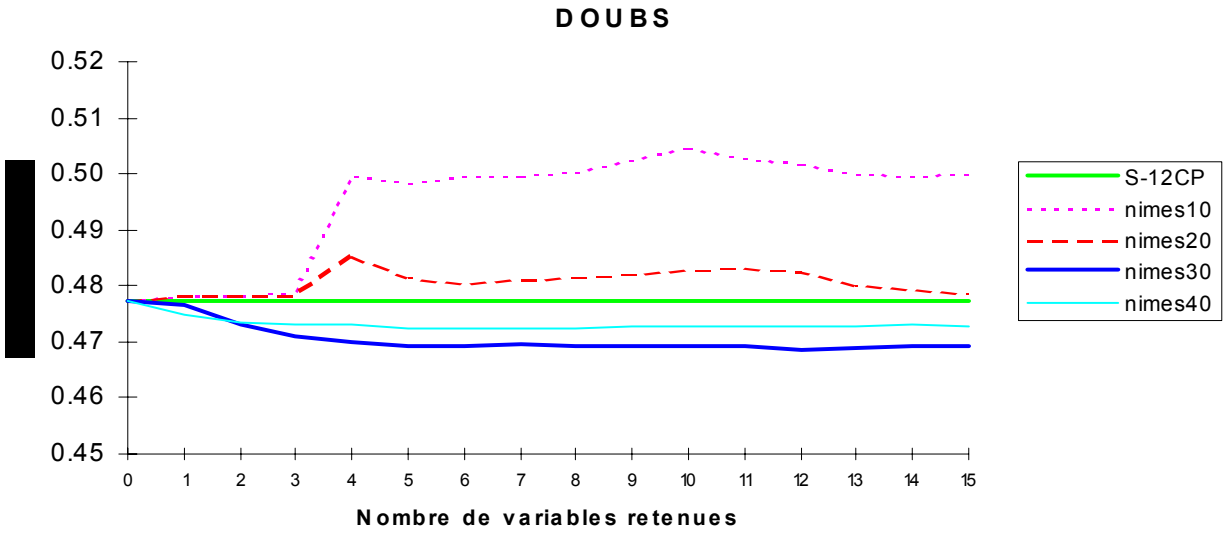
	prévision pluie / non pluie			prévision probabiliste en classes		
groupements	**53-93**	**54-83**	**écart**	**53-93**	**54-83**	**écart**
Creuse-Cher	80.22	80.00	-0.22	0.4503	0.4499	-0.0004
Vézère-Vienne-Thaurion	72.37	75.38	3.01	0.6365	0.5476	-0.0889
Dordogne	80.65	80.33	-0.32	0.4680	0.4646	-0.0034
Cère-Maronne	81.40	80.59	-0.81	0.5060	0.5133	0.0073
Truyère-Lot inférieur	79.25	79.27	0.02	0.4710	0.4663	-0.0047
Haut Tarn-Haut Lot	80.03	80.00	-0.03	0.4803	0.4866	0.0063
Agout-Tarn	75.05	75.49	0.44	0.5767	0.5746	-0.0021
Pyrénées Est	77.41	76.96	-0.45	0.4706	0.4662	-0.0044
Ariège-Vicdessos	75.93	76.74	0.81	0.5332	0.5247	-0.0085
Pique-Garonne-Salat	78.42	78.75	0.33	0.5357	0.5327	-0.0030
Gaves	74.30	76.85	2.55	0.5409	0.5235	-0.0174
Doubs	79.17	79.93	0.76	0.4816	0.4774	-0.0042
Ain-Valserine	77.49	79.78	2.29	0.5192	0.4988	-0.0204
Arve-Fier	76.31	79.38	3.07	0.5029	0.4834	-0.0195
Isère-Doron	81.05	80.29	-0.76	0.4272	0.4246	-0.0026
Isère moyenne	79.84	80.92	1.08	0.4941	0.4873	-0.0068
Romanche-Arc inférieur	80.76	80.48	-0.28	0.4574	0.4447	-0.0127
Drac	79.60	79.82	0.22	0.4691	0.4646	-0.0045
Buech-Drôme	79.87	80.37	0.50	0.4497	0.4455	-0.0042
Verdon	79.23	79.74	0.51	0.4510	0.4474	-0.0036
BVI Verdon	79.01	79.19	0.18	0.4626	0.4482	-0.0144
Var-Tinee-Roya	79.20	78.32	-0.88	0.4671	0.4616	-0.0055
Haute Durance	79.87	80.04	0.17	0.4327	0.4326	-0.0001
Durance moyenne	78.99	77.99	-1.00	0.4440	0.4317	-0.0123
Mont Cenis	78.34	77.40	-0.94	0.4904	0.4953	0.0049
Chassezac	77.89	77.62	-0.27	0.5377	0.5401	0.0024
Loire supérieure	76.98	76.01	-0.97	0.5227	0.5210	-0.0017
Doux-Eyrieux	78.53	79.30	0.77	0.4891	0.4807	-0.0084
Gard-Cèze	79.58	80.11	0.53	0.4879	0.4836	-0.0043
Loire moyenne	79.15	78.72	-0.43	0.4603	0.4601	-0.0002
Allier supérieur	78.42	78.75	0.33	0.4952	0.4971	0.0019
Sioule	81.56	81.94	0.38	0.4503	0.4526	0.0023
Cure	78.80	78.61	-0.19	0.5041	0.5070	0.0029
moyenne	**78.63**	**78.94**	**0.32**	**0.4899**	**0.4829**	**-0.0070**

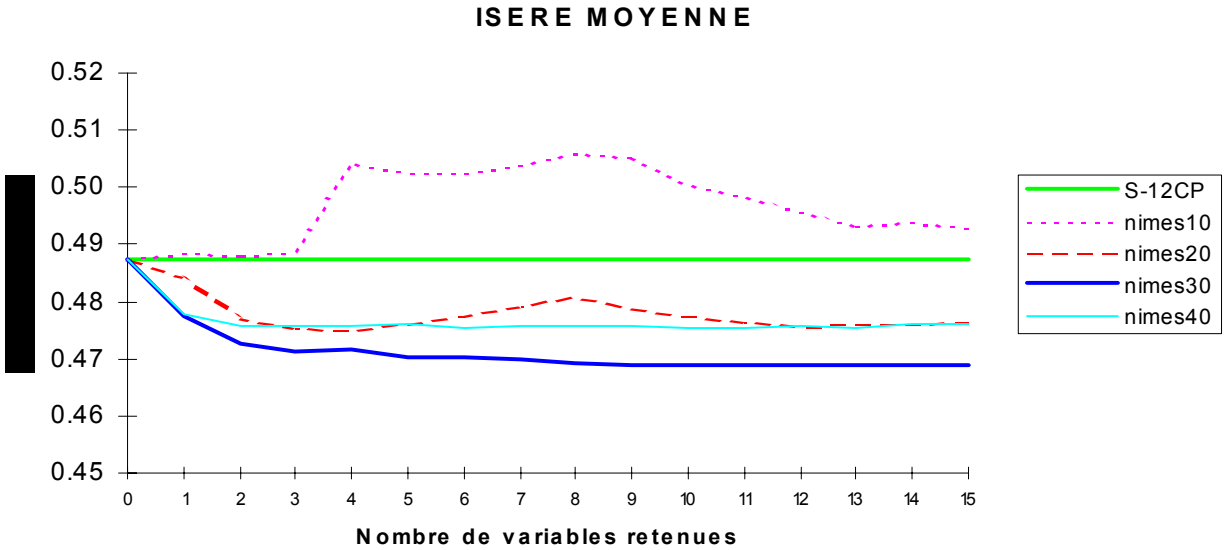
ANNEXE V-7:

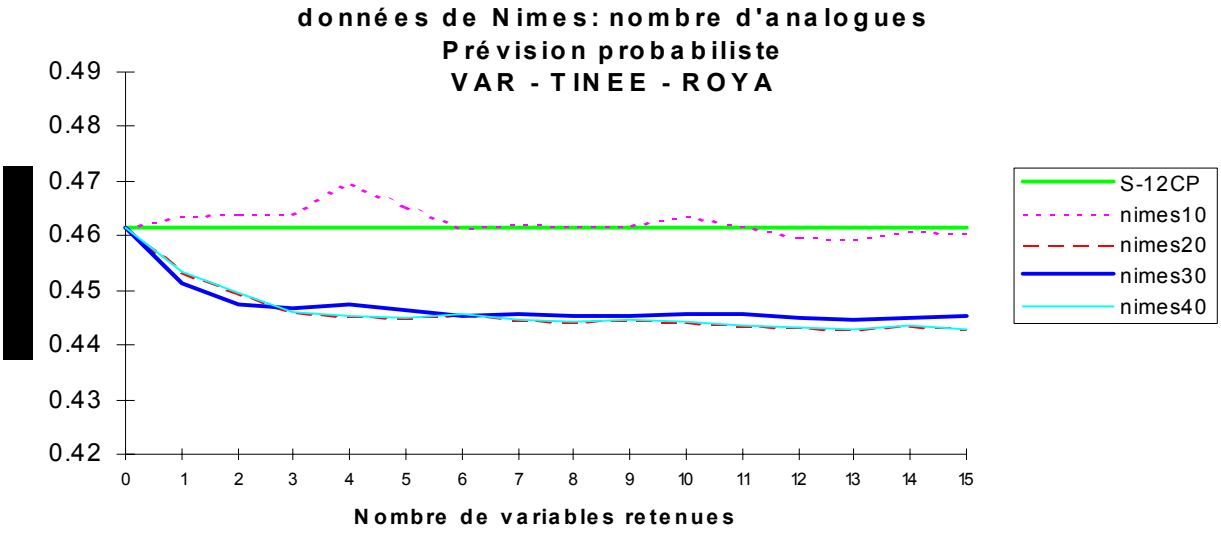
Données de Nîmes, nombre d'analagues

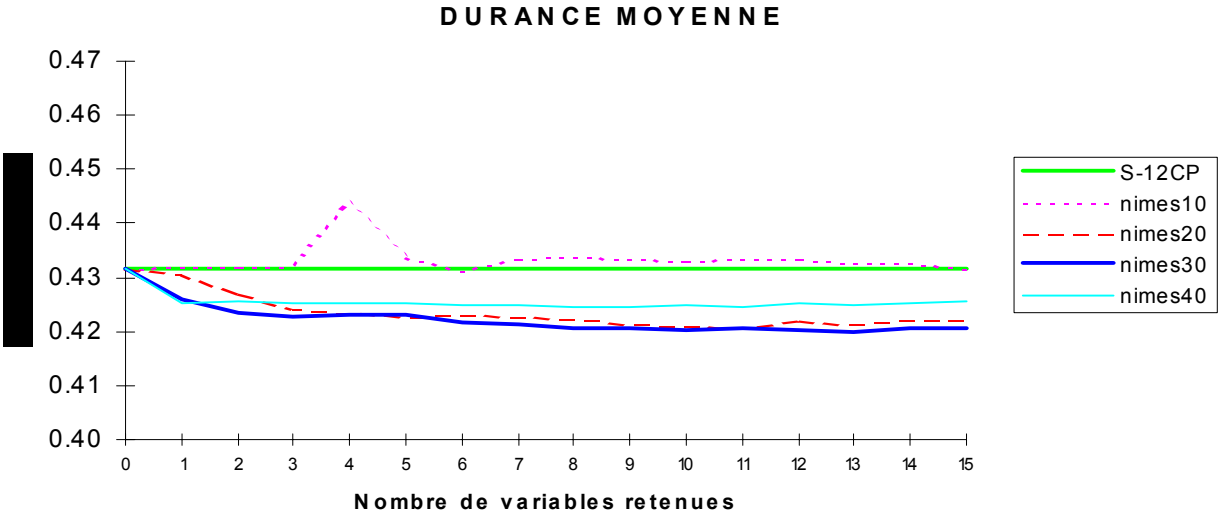


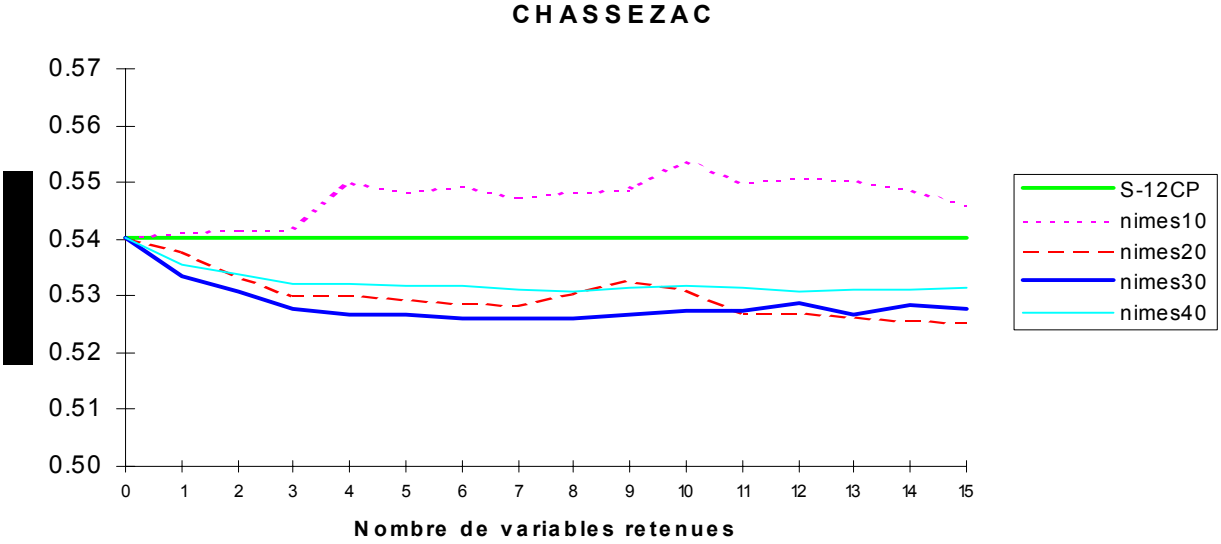

données de Nimes: nombre d'analogues
Prévision probabiliste
PYRENEES EST
0.50
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
nimes10
nimes20
nimes30
nimes40


DOUBS
0.52
0.51
0.50
0.49
0.48
0.47
0.46
0.45
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
nimes10
nimes20
nimes30
nimes40


ISERE MOYENNE
0.52
0.51
0.50
0.49
0.48
0.47
0.46
0.45
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
nimes10
nimes20
nimes30
nimes40

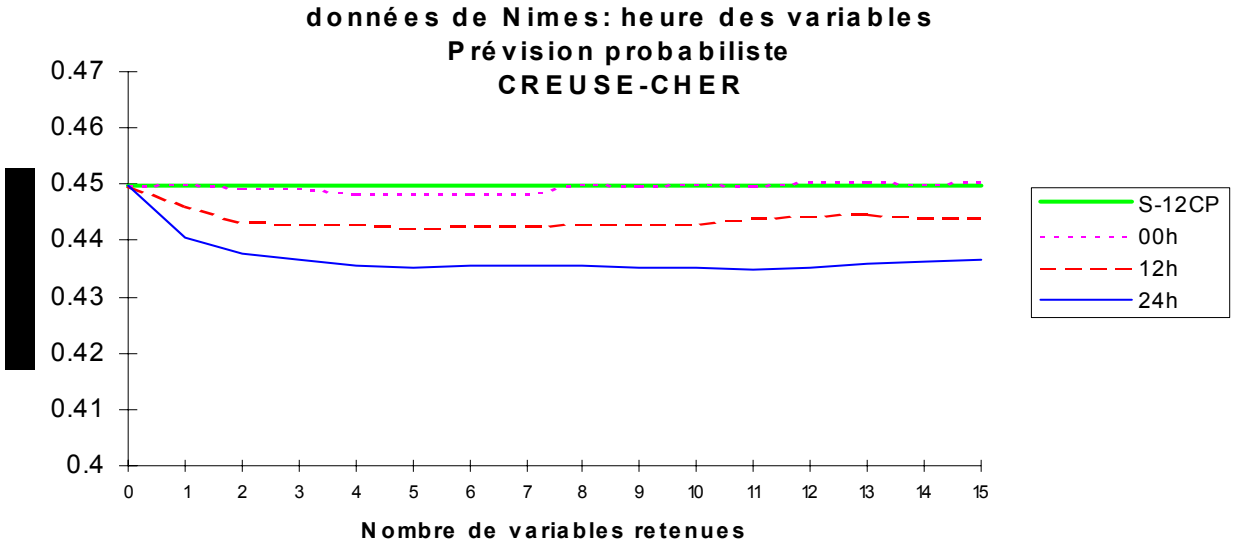

données de Nimes: nombre d'analogues
Prévision probabiliste
VAR - TINEE - ROYA
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
nimes10
nimes20
nimes30
nimes40

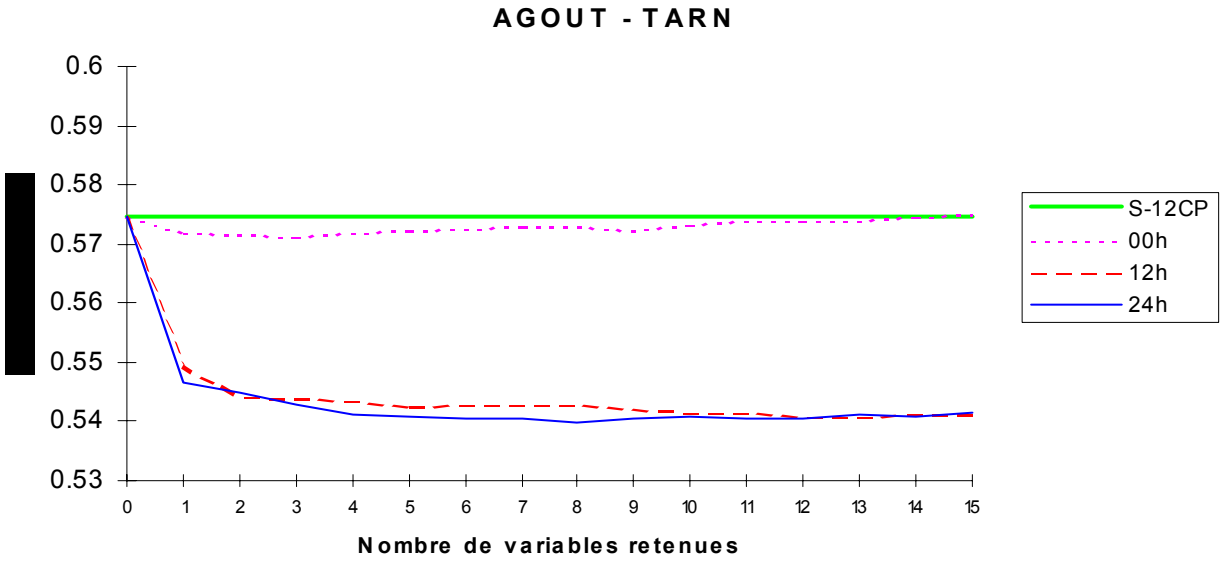

DURANCE MOYENNE
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.40
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
nimes10
nimes20
nimes30
nimes40

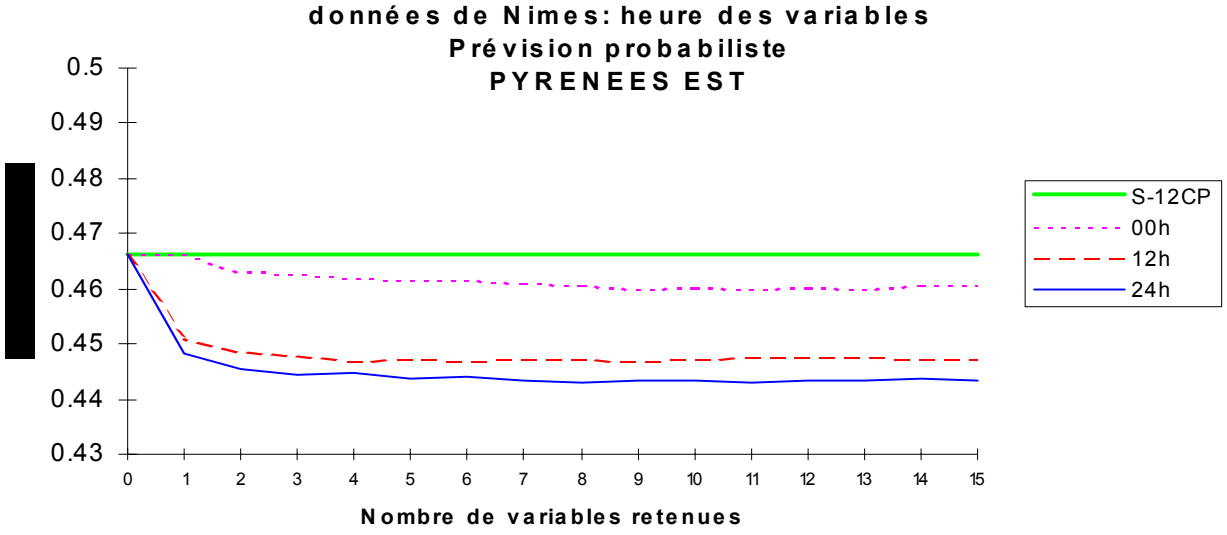

CHASSEZAC
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.50
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
nimes10
nimes20
nimes30
nimes40

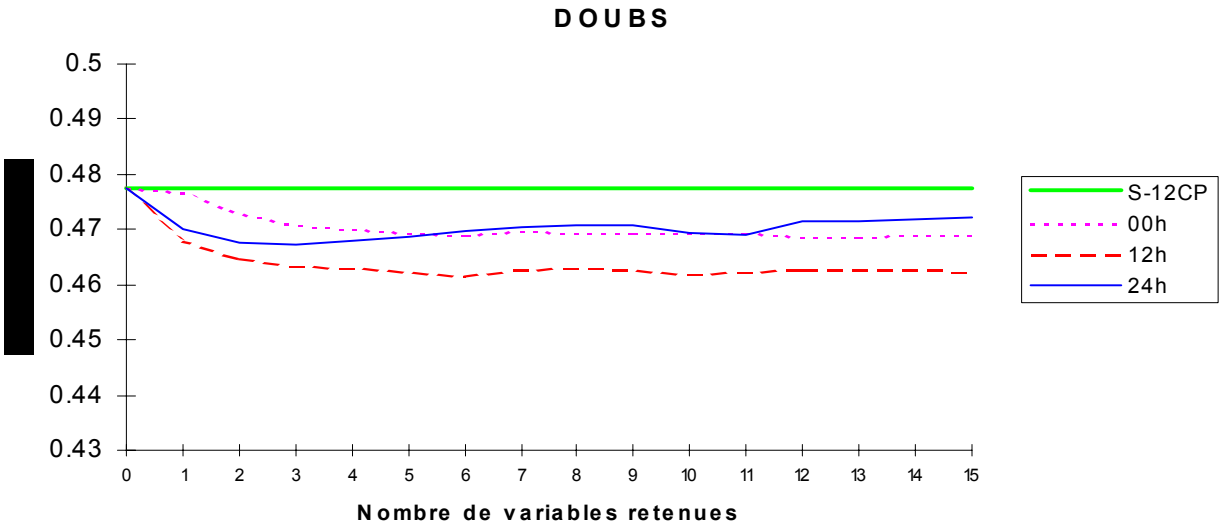
ANNEXE V-8:

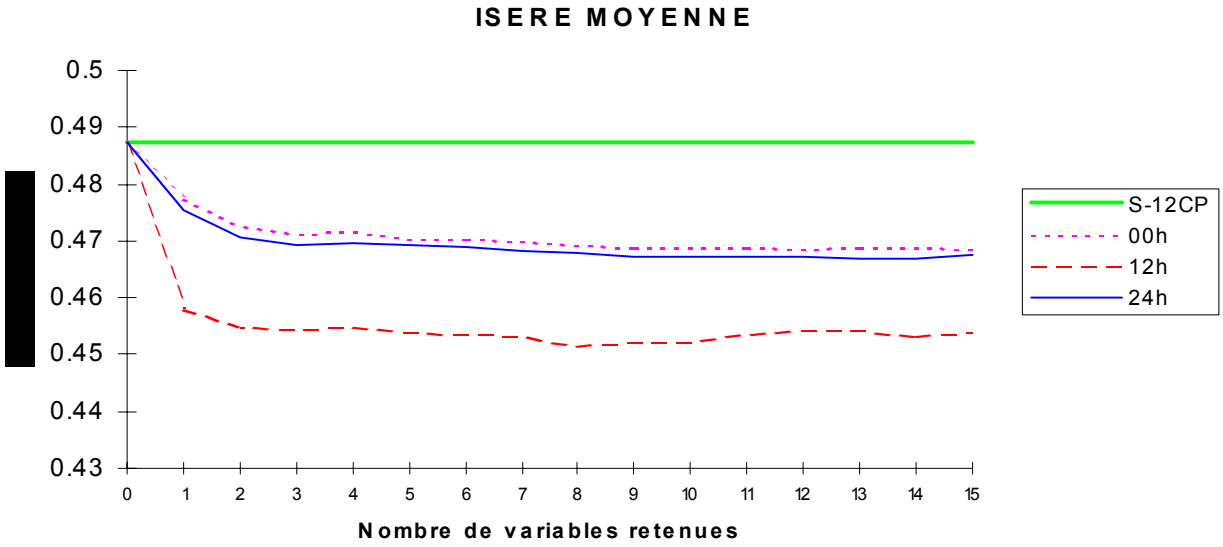
données de Nîmes: choix de l'échéance

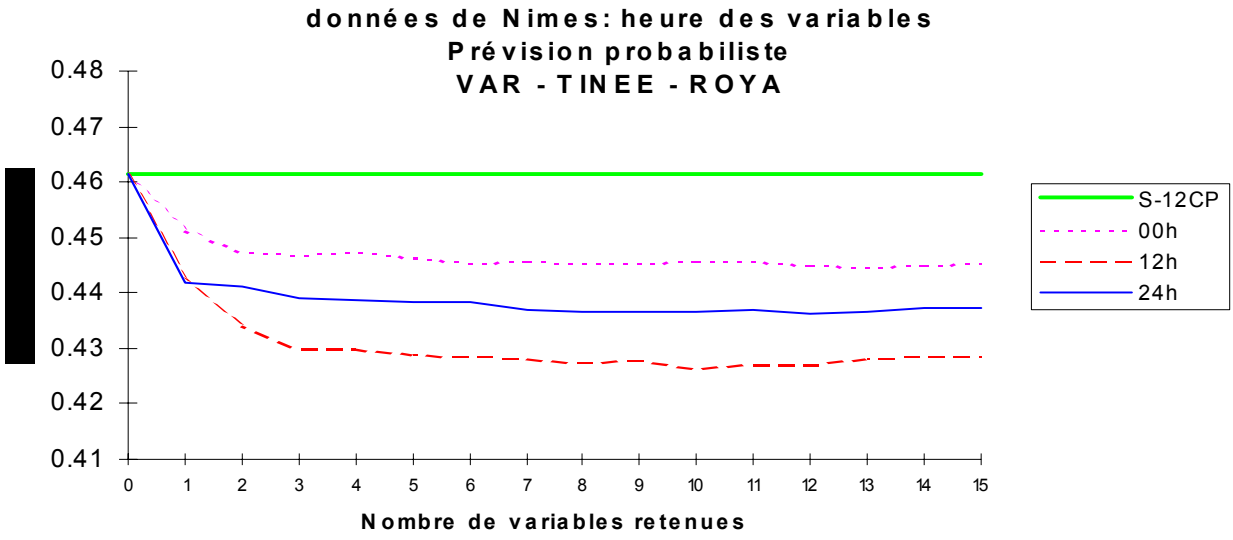


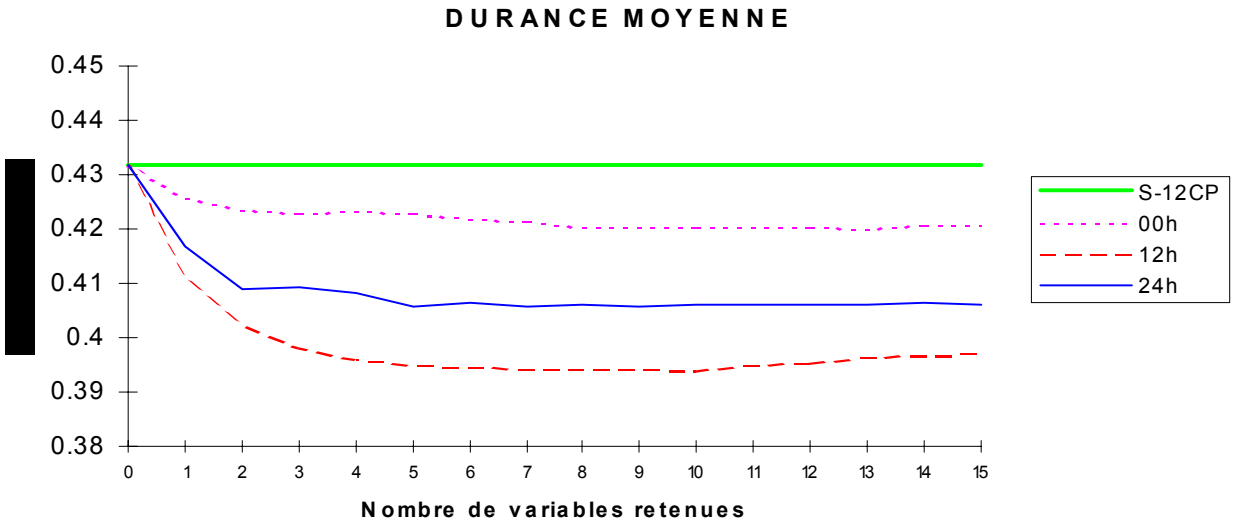


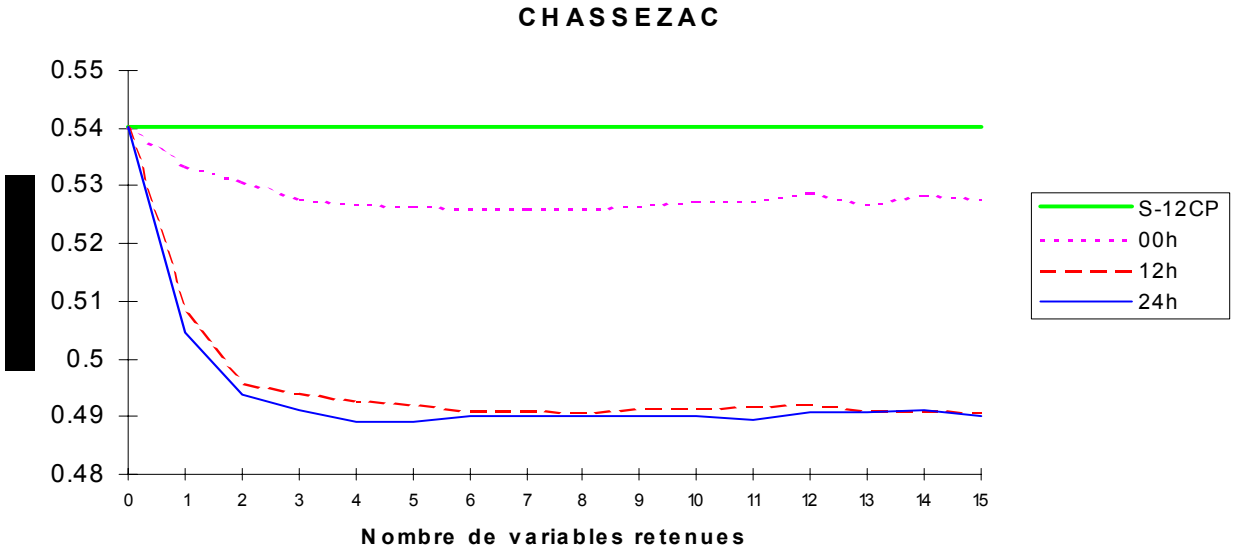

données de Nimes: heure des variables
Prévision probabiliste
PYRENEES EST
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
00h
12h
24h


DOUBS
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
00h
12h
24h


ISERE MOYENNE
0.5
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
00h
12h
24h


données de Nimes: heure des variables
Prévision probabiliste
VAR - TINEE - ROYA
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
00h
12h
24h


DURANCE MOYENNE
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0.38
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
00h
12h
24h


CHASSEZAC
0.55
0.54
0.53
0.52
0.51
0.5
0.49
0.48
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de variables retenues
S-12CP
00h
12h
24h

ANNEXE V-9:

Comparaison méthode S-12CP + données de Nîmes à 00 et 12h

	groupements	prévision pluie/non pluie			prévision en classes		
		S-12CP	**nimes 0h**	**nimes 12h**	**S-12CP**	**nimes 0h**	**nimes 12h**
1	Creuse-Cher	80.00	80.26	80.84	0.4499	0.4483	0.4425
2	Vézère-Vienne-Thaurion	75.38	75.35	76.19	0.5476	0.5473	0.5378
3	Dordogne	80.33	81.36	81.72	0.4646	0.4599	0.4474
4	Cère-Maronne	80.59	81.17	81.94	0.5133	0.5048	0.4869
5	Truyère-Lot inférieur	79.27	79.63	80.92	0.4663	0.4563	0.4389
6	Haut Tarn-Haut Lot	80.00	81.10	82.31	0.4866	0.4769	0.4445
7	Agout-Tarn	75.49	76.63	78.42	0.5746	0.5727	0.5428
8	Pyrénées Est	76.96	78.50	79.01	0.4662	0.4616	0.4468
9	Ariège-Vicdessos	76.74	76.85	78.97	0.5247	0.5202	0.5025
10	Pique-Garonne-Salat	78.75	79.93	79.93	0.5327	0.5273	0.5176
11	Gaves	76.85	77.77	78.35	0.5235	0.5231	0.5097
12	Doubs	79.93	80.81	81.43	0.4774	0.4691	0.4618
13	Ain-Valserine	79.78	81.03	80.73	0.4988	0.4886	0.4794
14	Arve-Fier	79.38	80.59	81.43	0.4834	0.4693	0.4597
15	Isère-Doron	80.29	81.43	81.28	0.4246	0.4128	0.4015
16	Isère moyenne	80.92	82.20	82.64	0.4873	0.4703	0.4536
17	Romanche-Arc inférieur	80.48	81.28	81.65	0.4447	0.4308	0.4091
18	Drac	79.82	81.39	81.98	0.4646	0.4494	0.4166
19	Buech-Drôme	80.37	81.14	82.49	0.4455	0.4305	0.4044
20	Verdon	79.74	80.88	82.09	0.4474	0.4339	0.4066
21	BVI Verdon	79.19	80.11	80.73	0.4482	0.4410	0.4186
22	Var-Tinee-Roya	78.32	80.62	80.88	0.4616	0.4454	0.4285
23	Haute Durance	80.04	81.43	82.71	0.4326	0.4183	0.3956
24	Durance moyenne	77.99	80.18	81.72	0.4317	0.4218	0.3945
25	Mont Cenis	77.40	78.42	79.93	0.4953	0.4799	0.4584
26	Chassezac	77.62	78.39	80.51	0.5401	0.5262	0.4910
27	Loire supérieure	76.01	76.89	78.02	0.5210	0.5121	0.4815
28	Doux-Eyrieux	79.30	80.40	81.94	0.4807	0.4714	0.4397
29	Gard-Cèze	80.11	82.12	83.70	0.4836	0.4723	0.4349
30	Loire moyenne	78.72	79.23	80.48	0.4601	0.4508	0.4321
31	Allier supérieur	78.75	79.56	81.03	0.4971	0.4890	0.4582
32	Sioule	81.94	82.12	82.38	0.4526	0.4500	0.4401
33	Cure	78.61	79.01	79.71	0.5070	0.5055	0.4990
	moyenne	**78.94**	**79.93**	**80.85**	**0.4829**	**0.4738**	**0.4540**

ANNEXE V-10:

Comparaison méthode TW-GR + données de Nîmes à 00 et 12 h

		prévision pluie/non pluie			prévision en classes		
groupements		**TW-GR**	**Nimes 0h**	**Nimes 12h**	**TW-GR**	**Nimes 0h**	**Nimes 12h**
1	Creuse-Cher	82.12	82.34	83.19	0.4234	0.4194	0.4153
2	Vézère-Vienne-Thaurion	78.72	79.38	79.82	0.5239	0.5249	0.5175
3	Dordogne	83.22	84.40	84.73	0.4362	0.4305	0.4215
4	Cère-Maronne	82.34	83.11	84.10	0.4719	0.4675	0.4536
5	Truyère-Lot inférieur	82.53	82.97	83.48	0.4251	0.4214	0.4059
6	Haut Tarn-Haut Lot	81.76	83.04	83.74	0.4408	0.4323	0.4107
7	Agout-Tarn	78.10	79.08	79.78	0.5208	0.5162	0.4963
8	Pyrénées Est	78.17	79.19	79.96	0.4373	0.4331	0.4235
9	Ariège-Vicdessos	79.19	80.00	79.74	0.4900	0.4886	0.4756
10	Pique-Garonne-Salat	79.38	80.26	79.67	0.5153	0.5121	0.5004
11	Gaves	77.51	78.83	79.12	0.5099	0.5095	0.4976
12	Doubs	83.08	83.70	84.03	0.4460	0.4398	0.4356
13	Ain-Valserine	83.81	84.76	84.87	0.4607	0.4538	0.4443
14	Arve-Fier	82.71	83.81	84.10	0.4385	0.4282	0.4184
15	Isère-Doron	80.51	82.49	82.86	0.3973	0.3882	0.3779
16	Isère moyenne	83.22	84.32	84.87	0.4449	0.4348	0.4204
17	Romanche-Arc inférieur	81.79	83.08	84.25	0.3903	0.3818	0.3656
18	Drac	81.17	83.15	83.77	0.4155	0.4034	0.3854
19	Buech-Drôme	80.59	82.71	83.70	0.4044	0.3940	0.3730
20	Verdon	80.26	82.09	83.08	0.4153	0.4061	0.3872
21	BVI Verdon	80.51	81.94	82.34	0.4263	0.4164	0.4023
22	Var-Tinee-Roya	81.10	82.20	82.23	0.4328	0.4193	0.4052
23	Haute Durance	81.06	82.38	83.15	0.3973	0.3882	0.3707
24	Durance moyenne	79.89	81.28	82.45	0.4081	0.3970	0.3804
25	Mont Cenis	80.04	80.95	81.10	0.4553	0.4477	0.4327
26	Chassezac	79.78	80.81	81.94	0.4943	0.4845	0.4596
27	Loire supérieure	79.38	80.22	81.36	0.4781	0.4668	0.4489
28	Doux-Eyrieux	79.89	81.14	82.67	0.4354	0.4281	0.4067
29	Gard-Cèze	80.88	82.20	83.55	0.4492	0.4384	0.4169
30	Loire moyenne	81.28	81.83	82.75	0.4252	0.4173	0.4061
31	Allier supérieur	80.15	81.06	82.67	0.4558	0.4472	0.4273
32	Sioule	82.60	83.52	83.44	0.4282	0.4217	0.4188
33	Cure	81.32	82.42	82.31	0.4631	0.4595	0.4574
moyenne		**80.85**	**81.96**	**82.57**	**0.4472**	**0.4399**	**0.4260**

ANNEXE V-11:

corrélation des données brutes

1000 mb

T	Td	H	r	ALT	T-Td	
1.00	0.72	-0.17	0.69	0.00	0.26	T
	1.00	0.55	0.97	-0.02	-0.48	Td
		1.00	0.55	-0.05	-0.98	H
			1.00	-0.04	-0.48	r
				1.00	0.03	ALT
					1.00	T-Td

950 mb

1.00	0.72	-0.03	0.69	0.26	0.11
	1.00	0.65	0.97	0.09	-0.61
		1.00	0.66	-0.15	-0.97
			1.00	0.07	-0.60
				1.00	0.16
					1.00

900 mb

1.00	0.62	-0.09	0.61	0.44	0.16
	1.00	0.70	0.97	0.13	-0.68
		1.00	0.69	-0.23	-0.96
			1.00	0.12	-0.64
				1.00	0.25
					1.00

850 mb

1.00	0.44	-0.25	0.47	0.59	-0.30
	1.00	0.72	0.95	0.11	-0.72
		1.00	0.69	-0.34	-0.96
			1.00	0.12	-0.65
				1.00	0.34
					1.00

800 mb

1.00	0.33	-0.31	0.36	0.69	0.35
	1.00	0.75	0.94	0.11	-0.78
		1.00	0.73	-0.35	-0.95
			1.00	0.13	-0.70
				1.00	0.35
					1.00

700 mb

1.00	0.28	-0.27	0.31	0.80	0.31
	1.00	0.80	0.92	0.17	-0.82
		1.00	0.78	-0.29	-0.95
			1.00	0.18	-0.74
				1.00	0.31
					1.00

600 mb

T	Td	H	r	ALT	T-Td
1.00	0.33	-0.22	0.31	0.85	0.29
	1.00	0.80	0.84	0.22	-0.81
		1.00	0.73	-0.27	-0.95
			1.00	0.19	-0.68
				1.00	0.31
					1.00

500 mb

1.00	0.45	-0.16	0.36	0.87	0.25
	1.00	0.77	0.73	0.32	-0.75
		1.00	0.61	-0.24	-0.95
			1.00	0.24	-0.56
				1.00	0.30
					1.00

400 mb

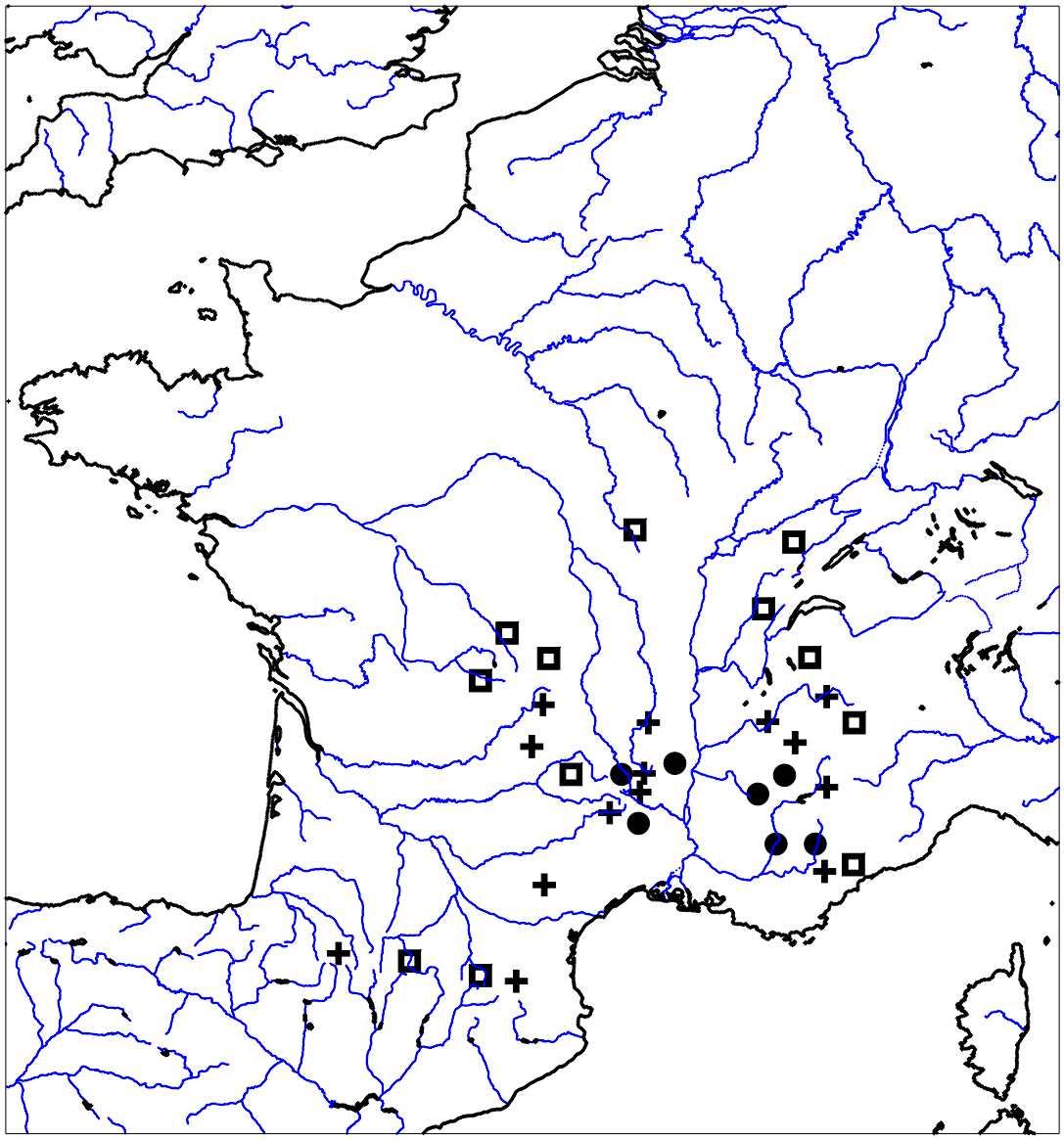
1.00	0.48	-0.10	0.15	0.88	0.18
	1.00	0.75	0.41	0.36	-0.73
		1.00	0.33	-0.17	-0.95
			1.00	0.10	-0.30
				1.00	0.24
					1.00

300 mb

1.00	0.07	-0.04	-0.30	0.80	0.07
	1.00	0.84	0.05	-0.03	-0.94
		1.00	0.05	-0.06	-0.87
			1.00	-0.20	-0.04
				1.00	0.11
					1.00

ANNEXE V-12:

carte de répartition des gains pour la méthode TW-GR + Nîmes 12 h (prévision pluie/non pluie)



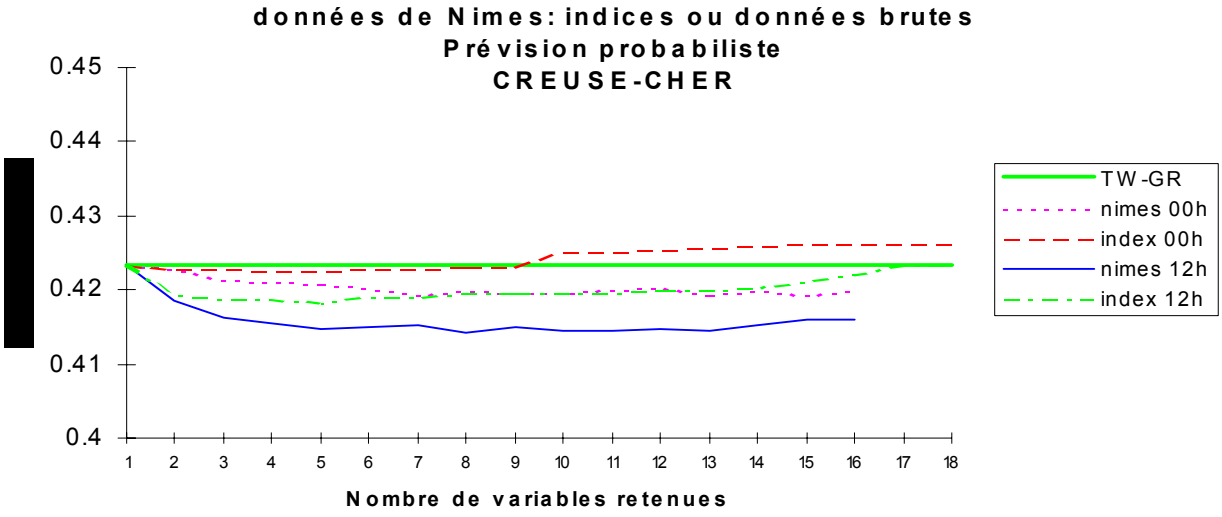
● gain de IR > 2.5

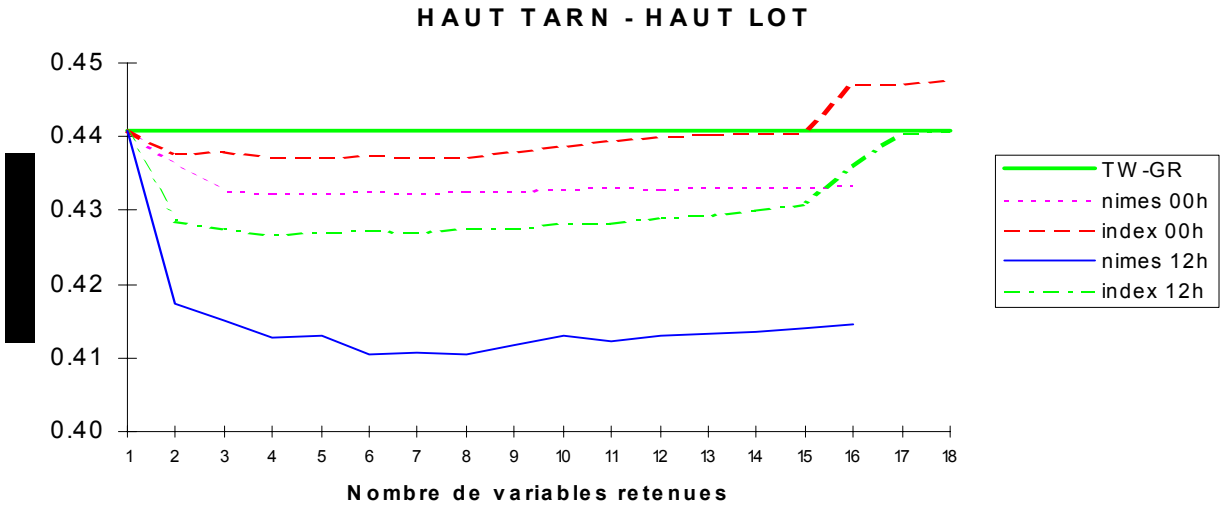
\+ 1.5 < gain IR < 2.5

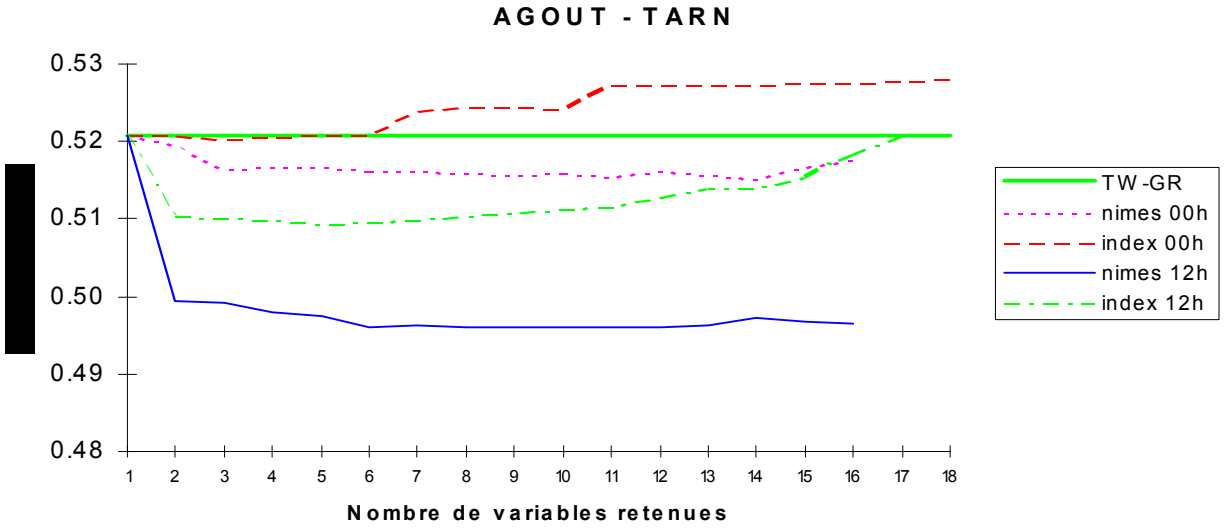
□ gain de IR < 1.5

ANNEXE V-13:

Utilisation des indices d'instabilité







ANNEXE V-14:

corrélation données de Nîmes / pluie moyenne des 33 groupements

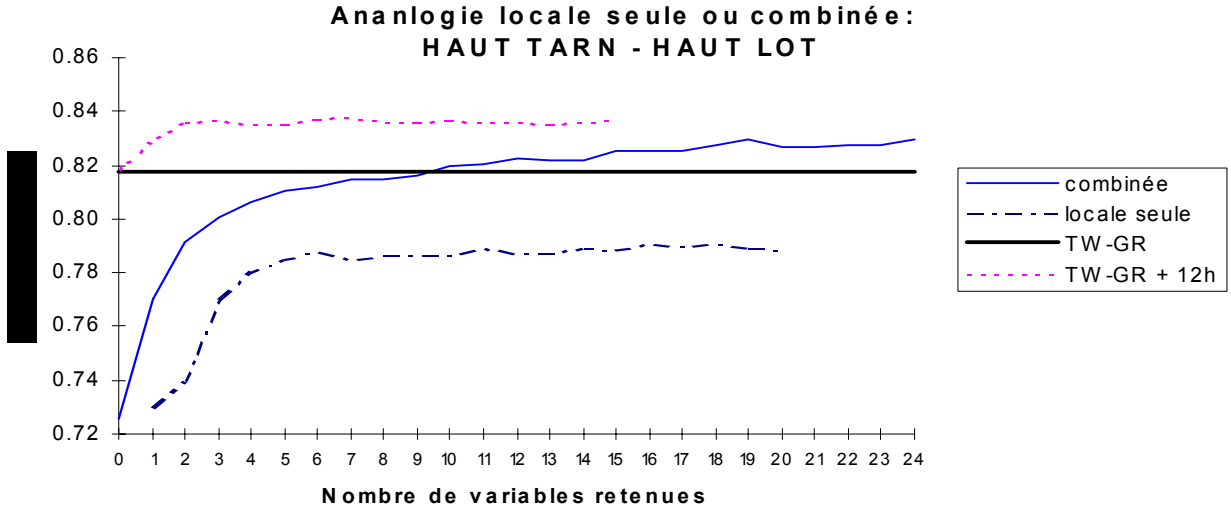
	T	Td	**H**	r	**ALT**	T-Td
1000 mb	-0.02	0.16	**0.27**	0.18	**-0.28**	0.13
950 mb	0.00	0.18	**0.27**	0.19	**-0.34**	0.05
900 mb	0.00	0.19	**0.26**	0.21	**-0.32**	0.04
850 mb	-0.03	0.21	**0.27**	0.23	**-0.30**	0.04
800 mb	-0.06	0.23	**0.30**	0.25	**-0.28**	0.05
700 mb	-0.08	0.23	**0.30**	0.24	**-0.24**	0.05
600 mb	-0.07	0.22	**0.29**	0.20	**-0.22**	0.05
500 mb	-0.05	0.20	**0.27**	0.15	**-0.19**	0.05
400 mb	-0.02	0.17	**0.24**	0.06	**-0.16**	0.05
300 mb	-0.02	0.18	**0.22**	0.11	**-0.14**	0.03

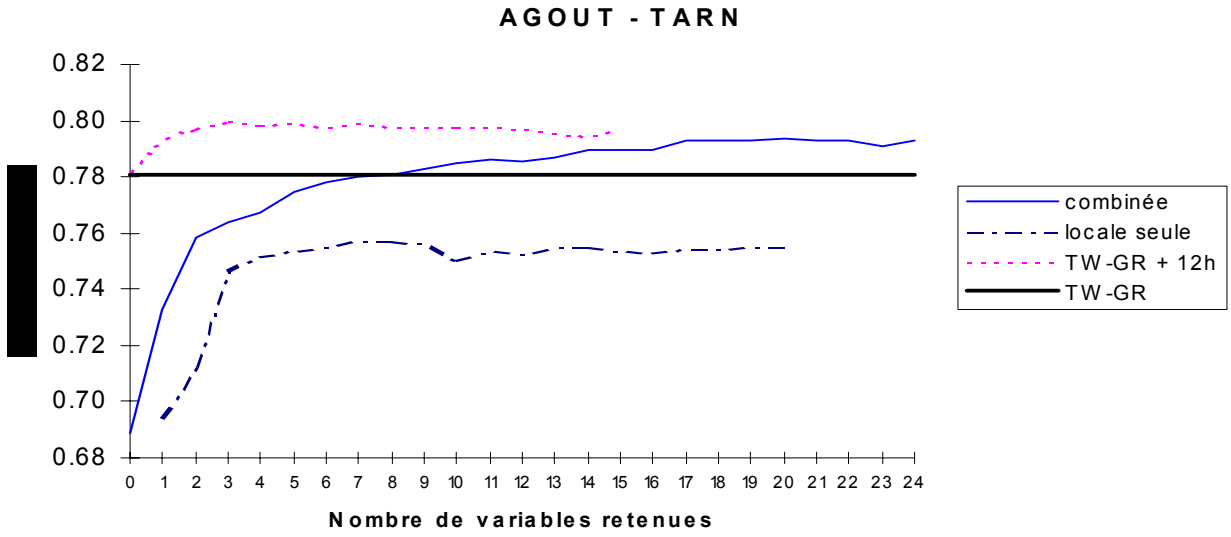
Corrélation pluie / indice d'instabilité de Nîmes à 12h

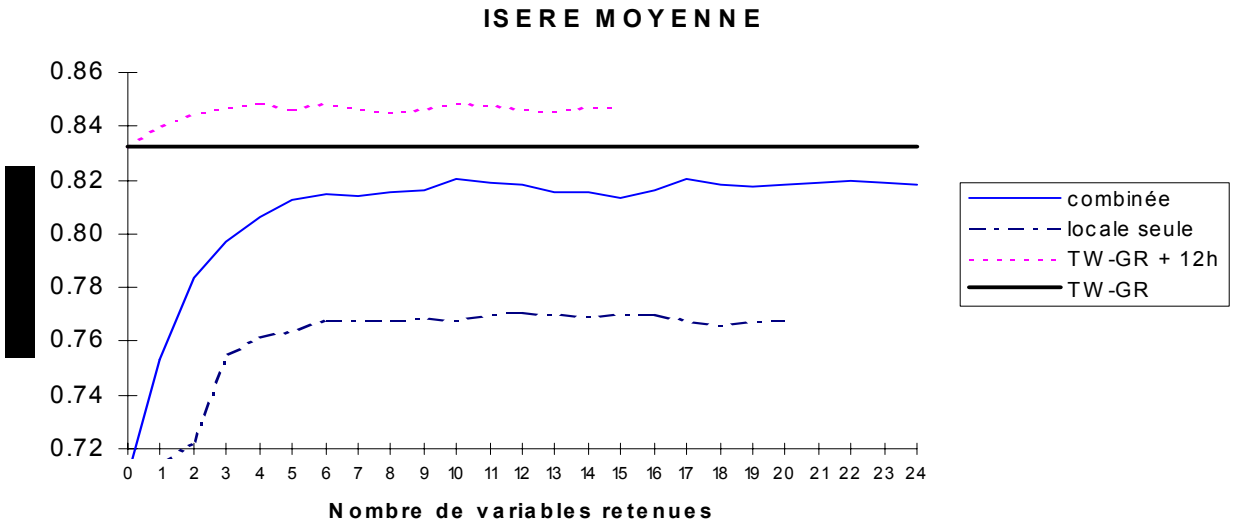
SI	0.12
LI	0.12
KI	0.14
TT	0.13
TTHI	0.15
SWEAT	0.13
PRES 0°	0.13
ALT 0°	-0.13
MAP1	0.13
MAP2	0.13
MAP3	0.13
MAP4	0.13
MAP5	0.13
CAPE	0.18
RI	0.07
VX	0.13

ANNEXE V-15:

L'analogie locale seule ou combinée avec l'analogie synoptique







ANNEXE V-16:

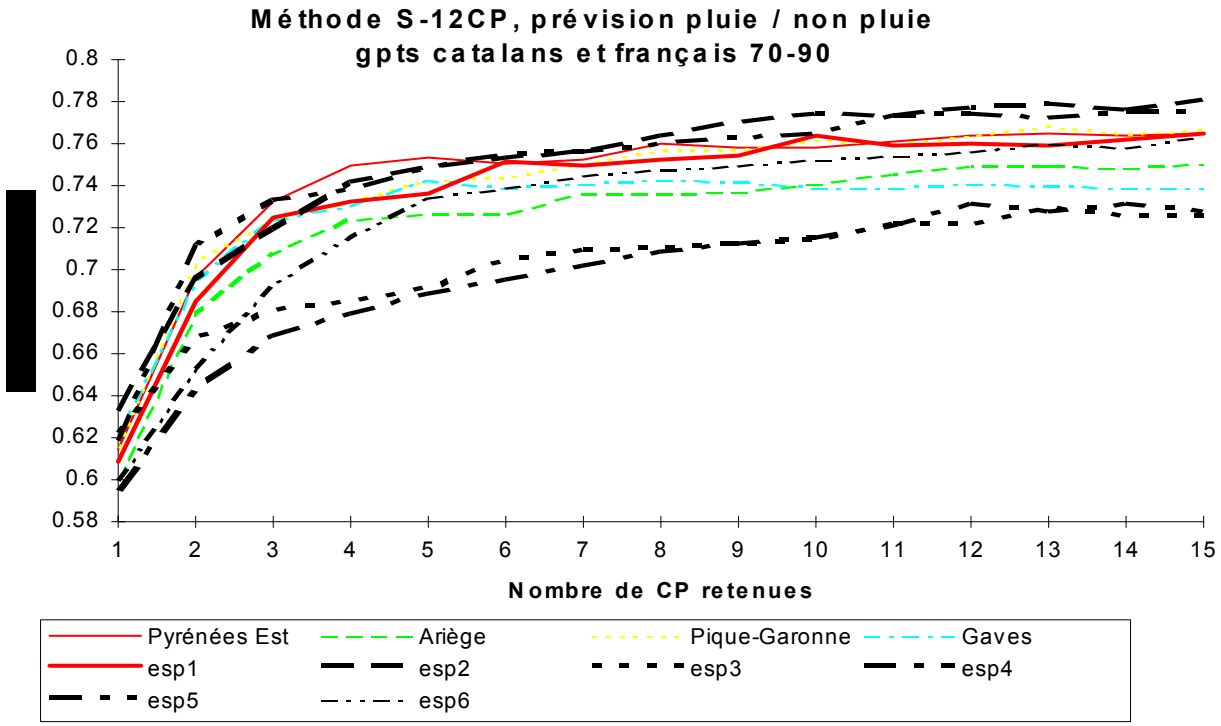
l'analogie locale avant l'analogie synoptique (pluie/non pluie)

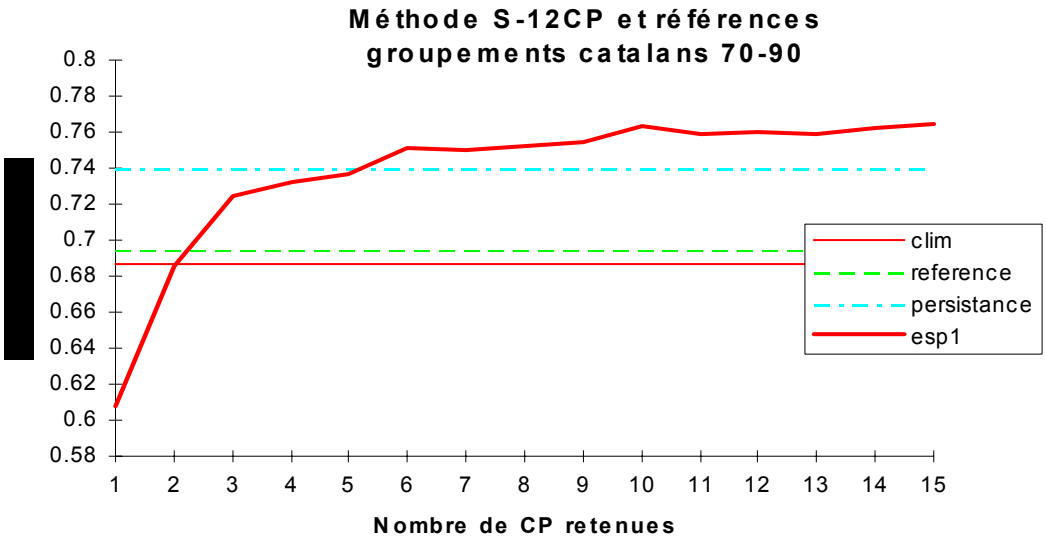
	analogie locale seule	analogie locale avant l'analogie synoptique
5. Truyère-Lot inférieur	75.5	74.9
6. Haut Tarn-Haut Lot	76.9	76.6
7. Agout-Tarn	74.6	74.5
14. Arve-Fier	72.5	71.4
15. Isère Doron	71.4	70.9
16. Isère moyenne	75.5	75.3
17. Romanche-Arc inférieur	74.8	75.4
18. Drac	77.3	76.6
19. Buech-Drôme	75.9	75.7
20. Verdon	77.5	77.3
21. BVI Verdon	75.6	75.0
22. Var-Tinee-Roya	74.0	72.6
23. Haute Durance	77.3	77.5
24. Durance moyenne	76.5	76.7
25. Mont Cenis	75.8	76.3
26. Chassezac	76.2	76.6
27. Loire supérieure	76.4	76.6
28. Doux-Eyrieux	76.0	76.0
29. Gard-Cèze	77.9	77.8
30. Loire moyenne	75.9	74.2
31. Allier supérieur	77.8	77.1
moyenne	**75.8**	**75.5**

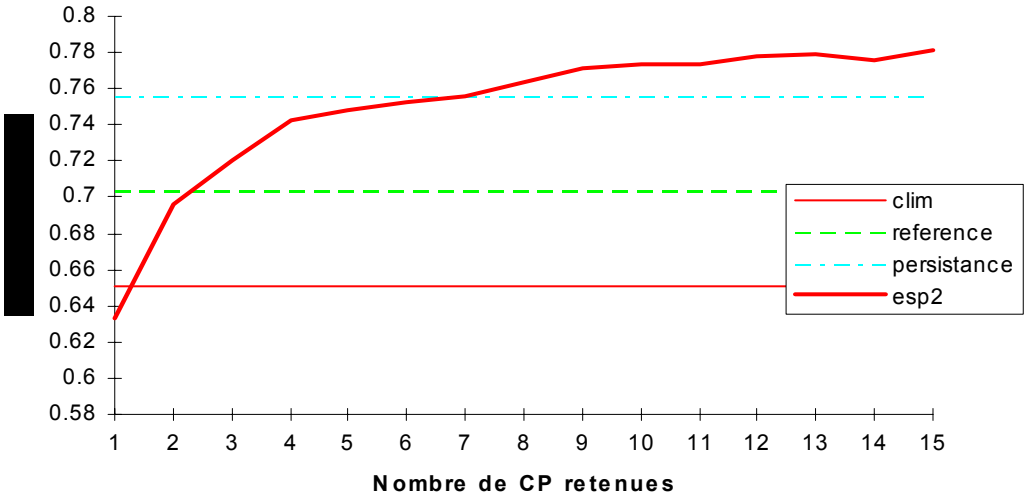
ANNEXES DU CHAPITRE VI

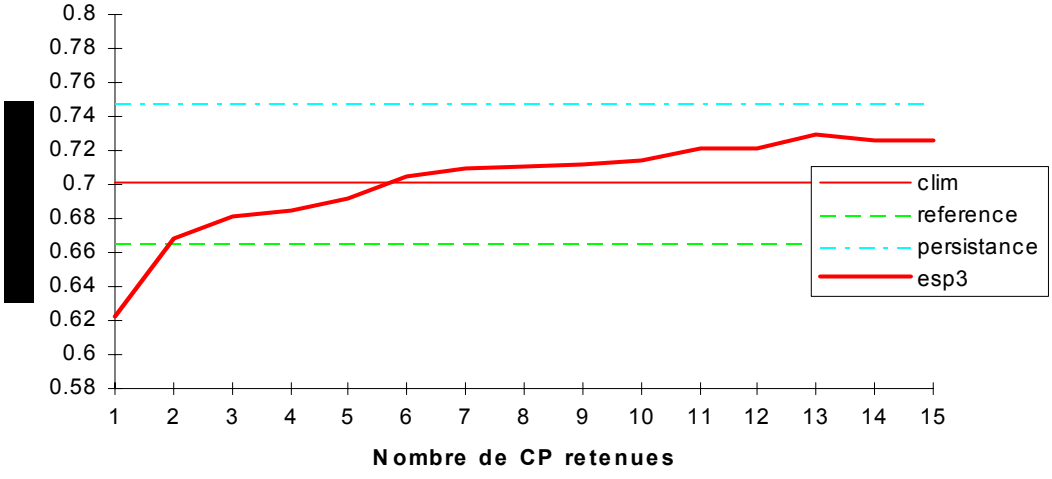
ANNEXE VI-1:

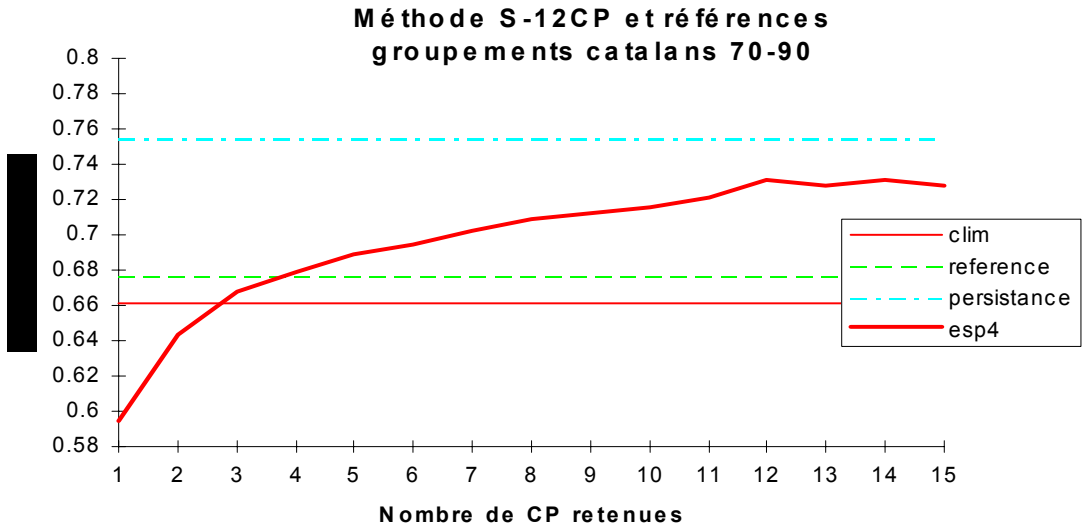
Espagne-Catalogne

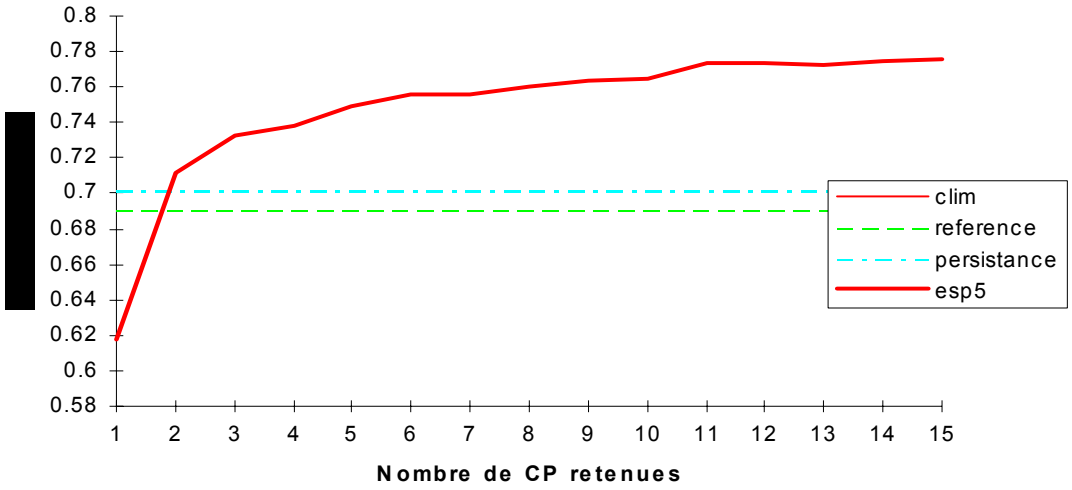


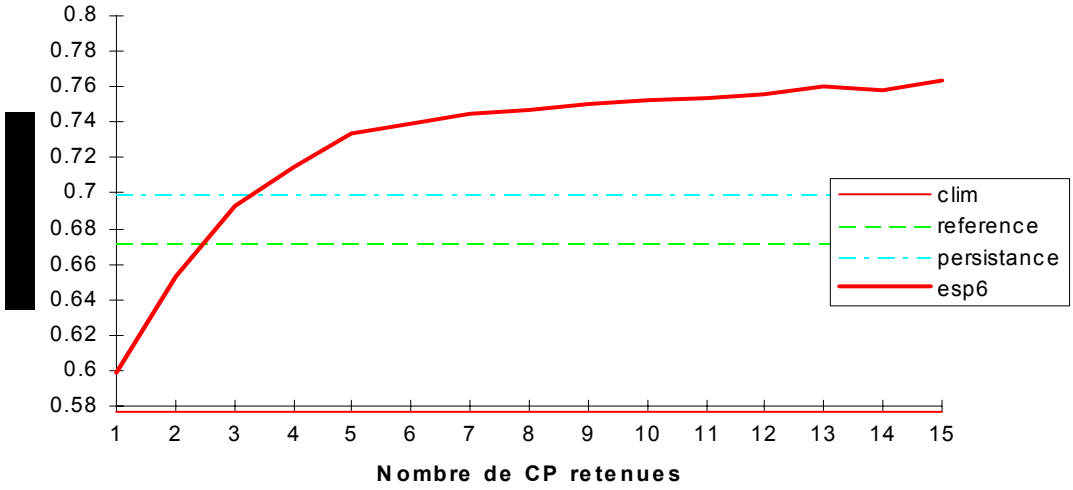

Méthode S-12CP et références
groupements catalans 70-90
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
esp1


0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
esp2


0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
esp3

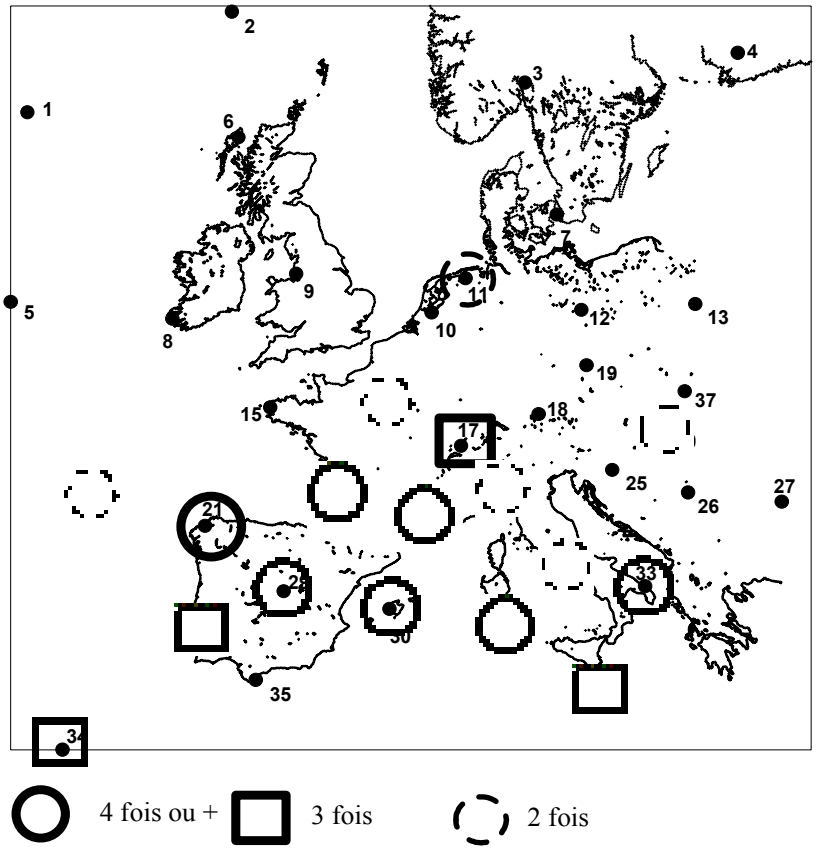

Méthode S-12CP et références
groupements catalans 70-90
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
esp4


0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
esp5

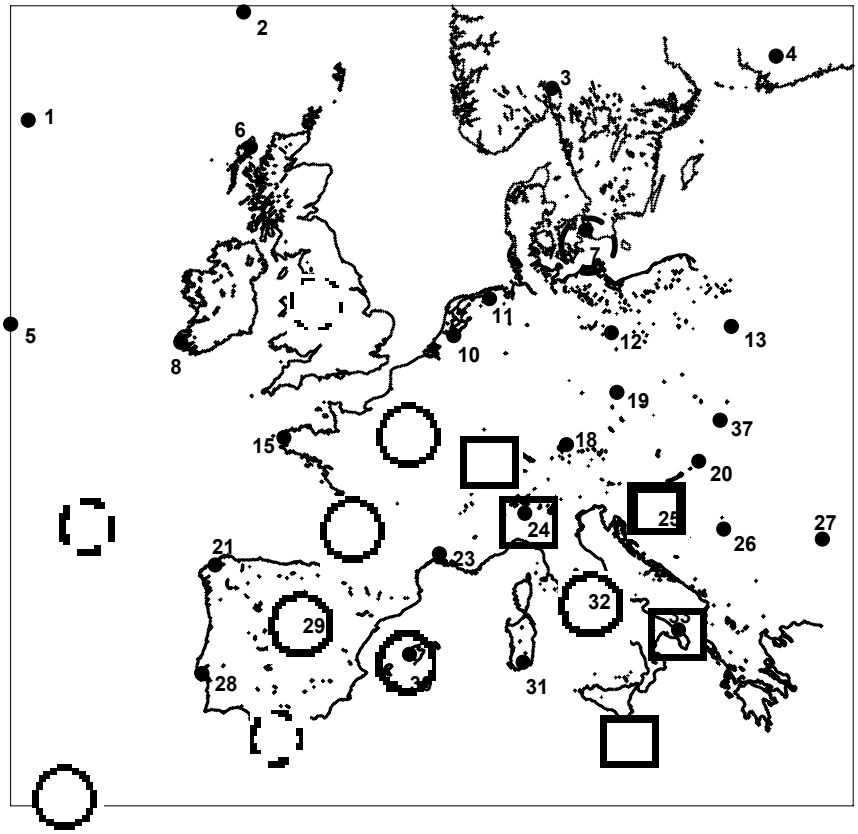

0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
0.58
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
esp6

ANNEXE VI-2:

Catalogne - Les RS les plus sélectionnées



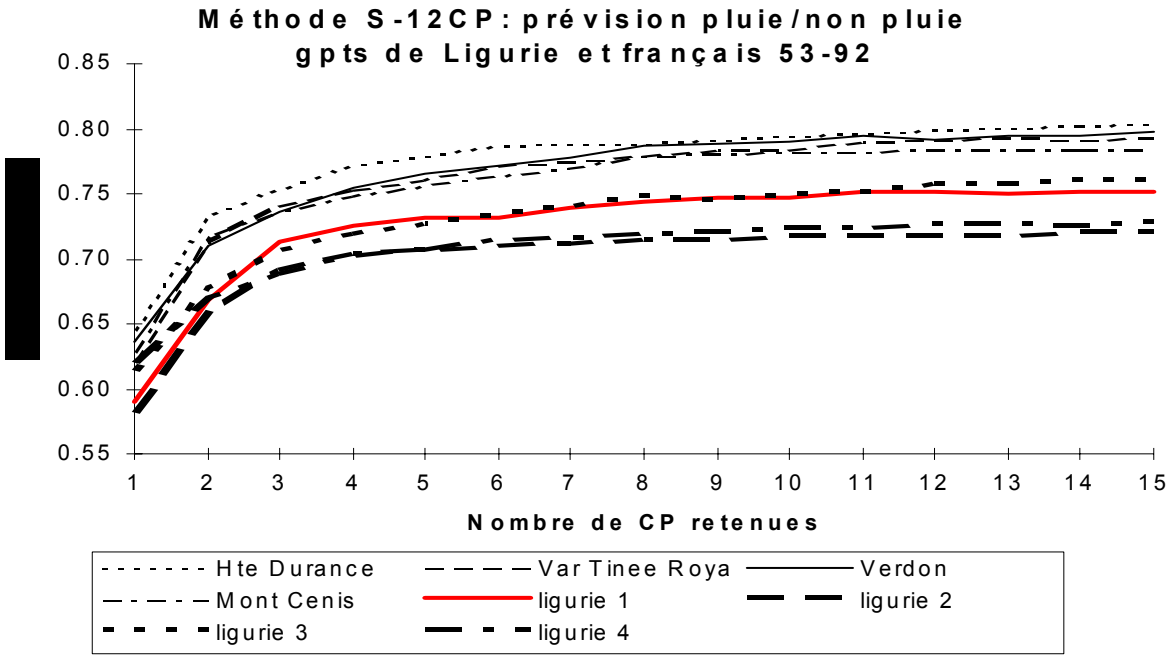
prévision en pluie/non pluie

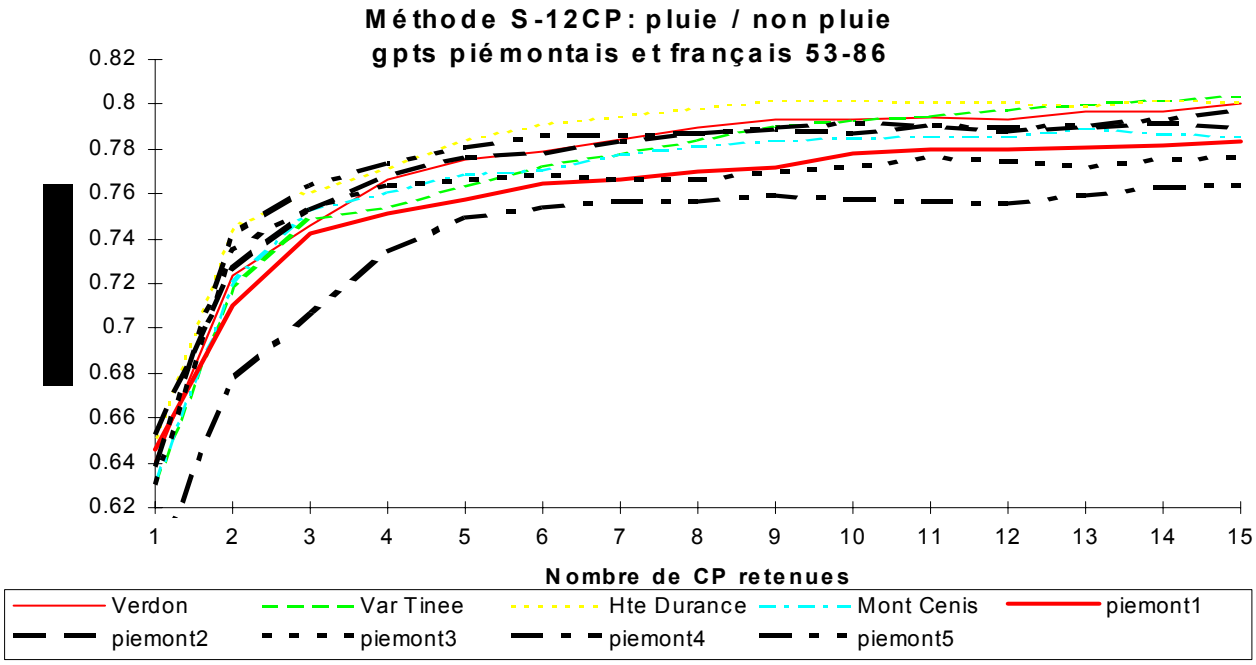


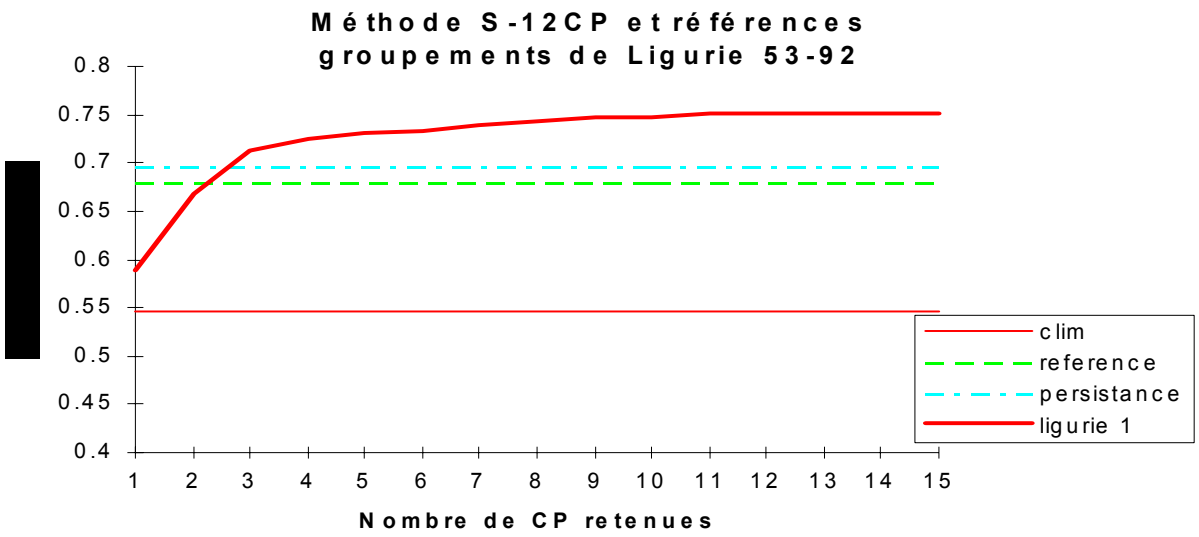
prévision en classes

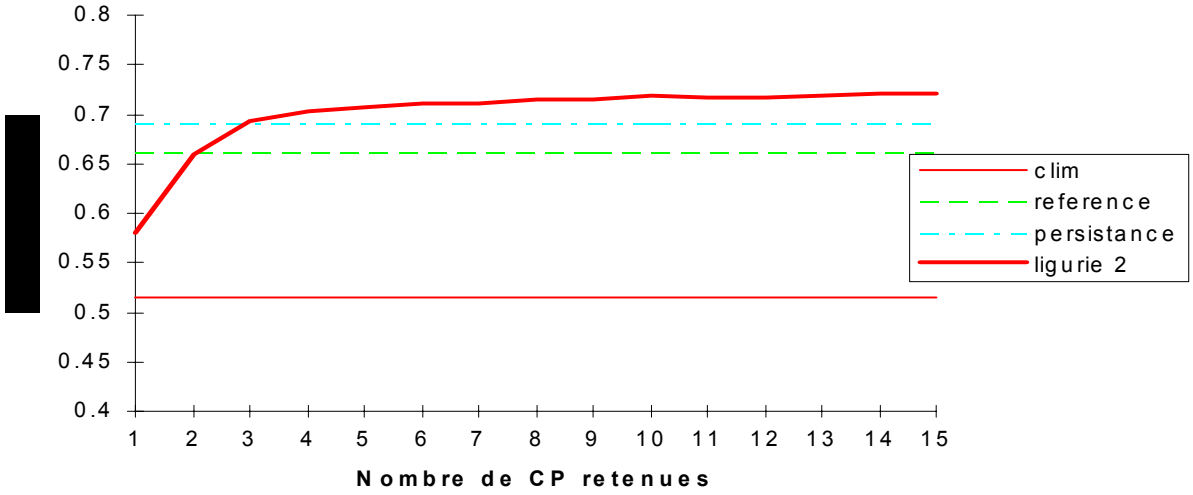
ANNEXE VI-3 :

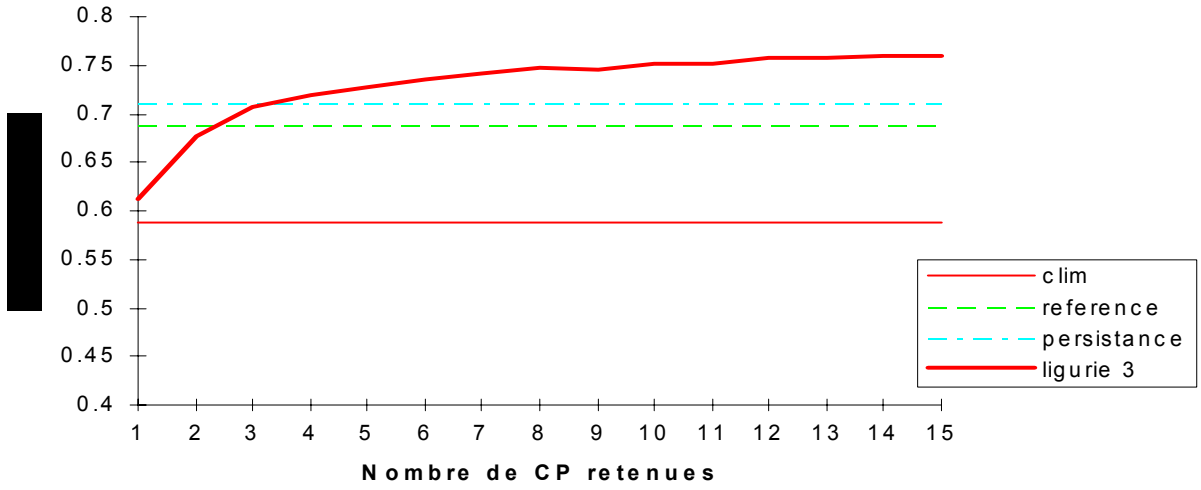
Italie-Ligurie et Piémont

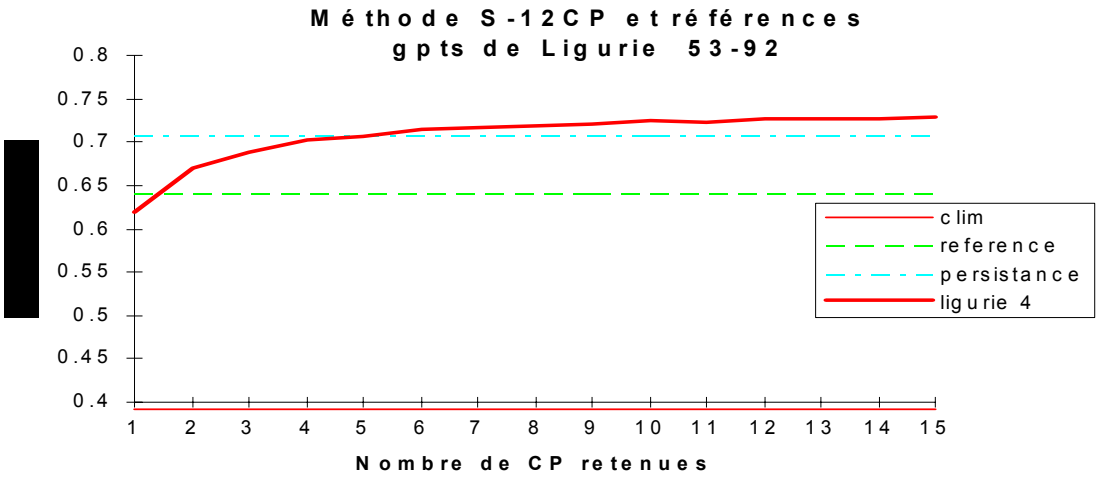


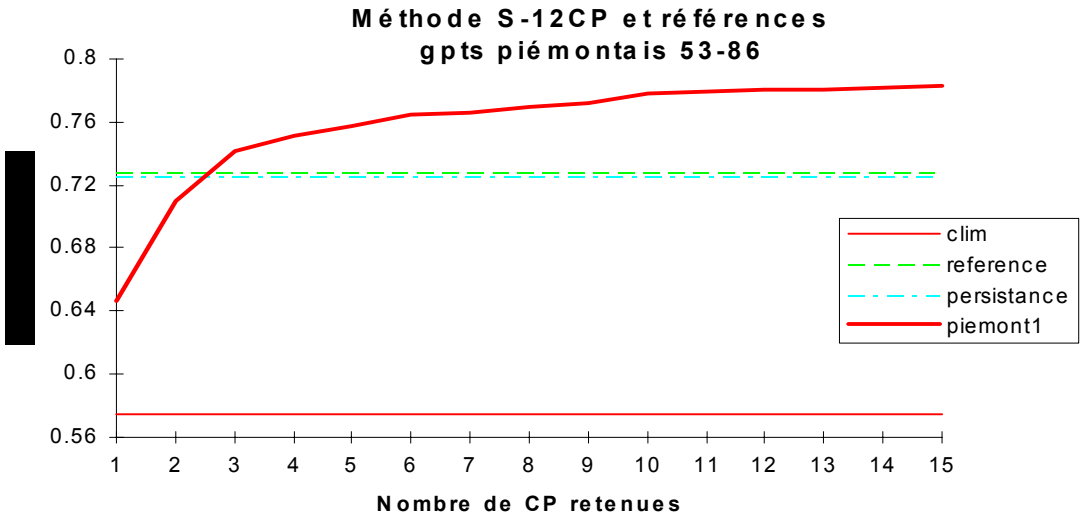


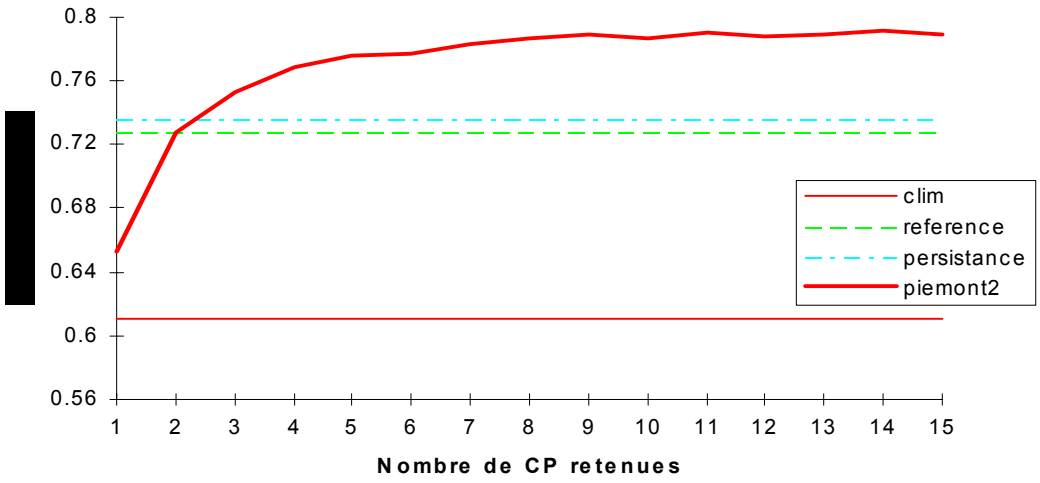

Méthode S-12CP et références
groupements de Ligurie 53-92
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
ligurie 1

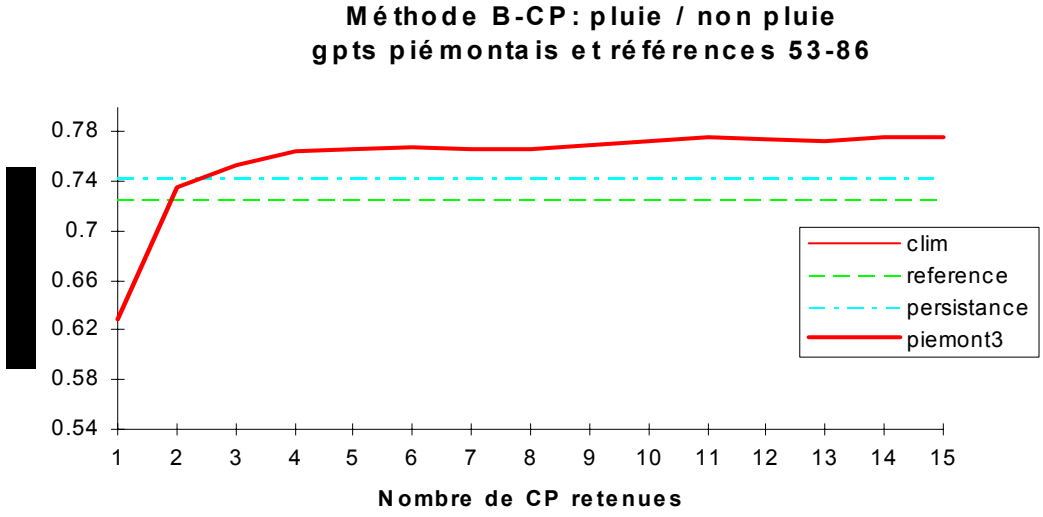

0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
ligurie 2

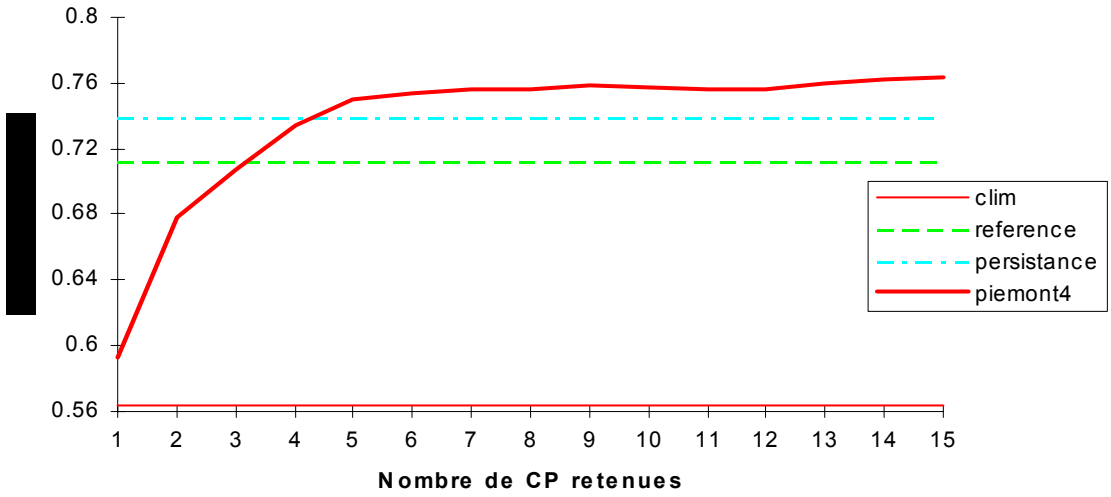

0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
ligurie 3

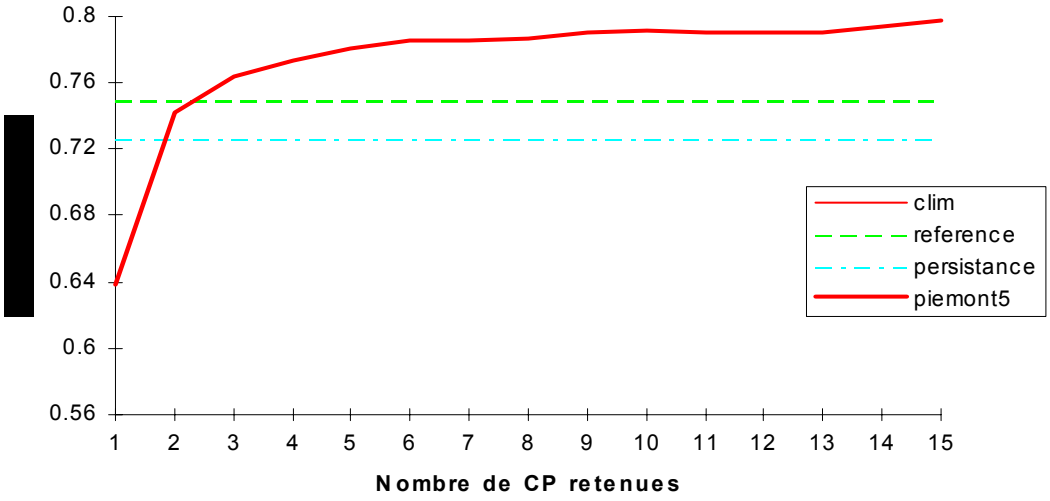

Méthode S-12CP et références
gpts de Ligurie 53-92
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
ligurie 4


Méthode S-12CP et références
gpts piémontais 53-86
0.8
0.76
0.72
0.68
0.64
0.6
0.56
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
piemont1


0.8
0.76
0.72
0.68
0.64
0.6
0.56
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
piemont2

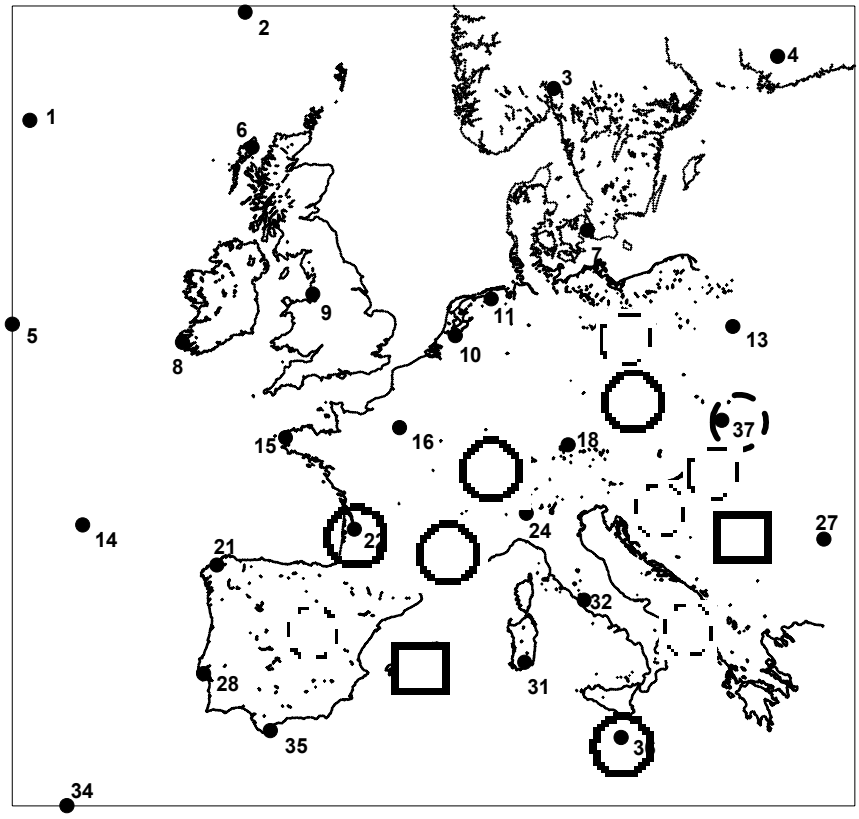

Méthode B-CP: pluie / non pluie
gpts piémontais et références 53-86
0.78
0.74
0.7
0.66
0.62
0.58
0.54
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
piemont3


0.8
0.76
0.72
0.68
0.64
0.6
0.56
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
piemont4


0.8
0.76
0.72
0.68
0.64
0.6
0.56
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de CP retenues
clim
reference
persistance
piemont5

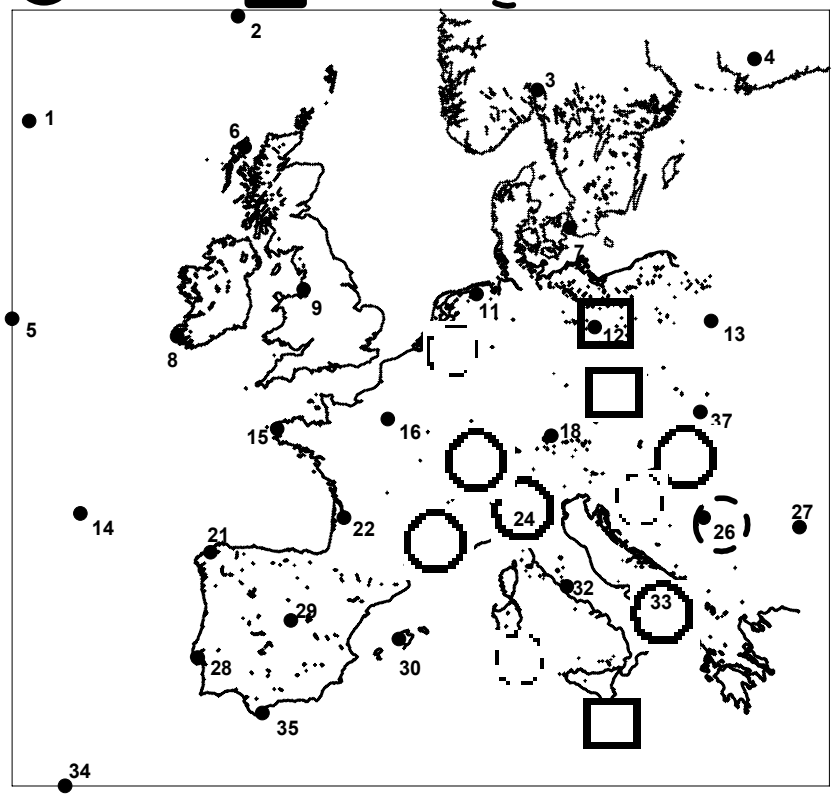
ANNEXE VI-4:

Ligurie - Les RS les plus sélectionnées



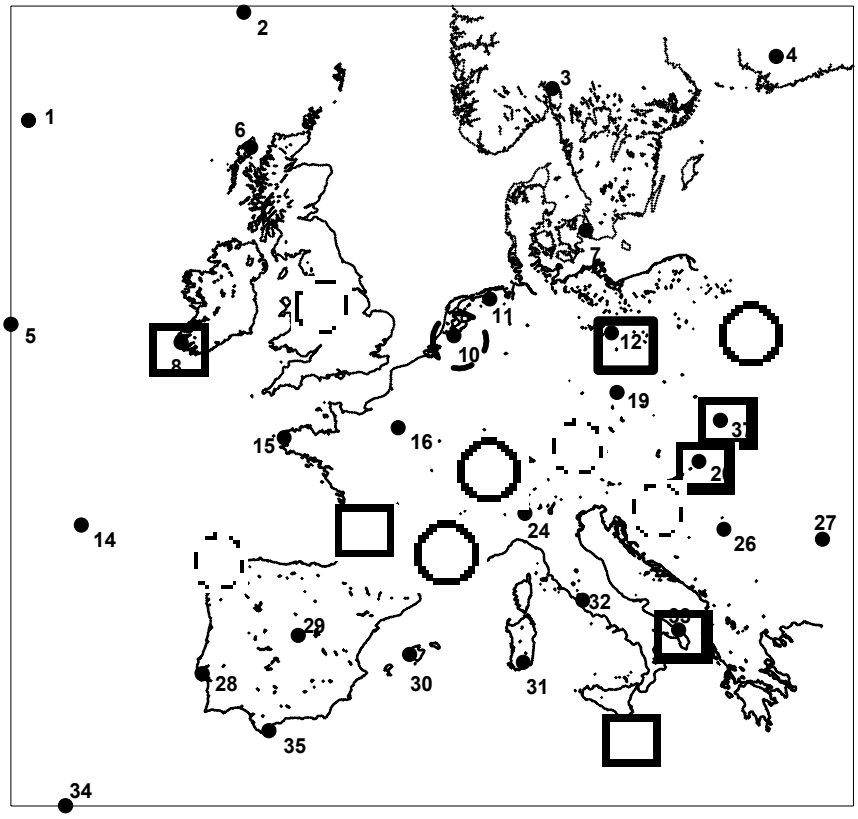
prévision en pluie/non pluie

4 fois ou + 3 fois 2 fois



prévision en classes

Piémont - Les RS les plus sélectionnées



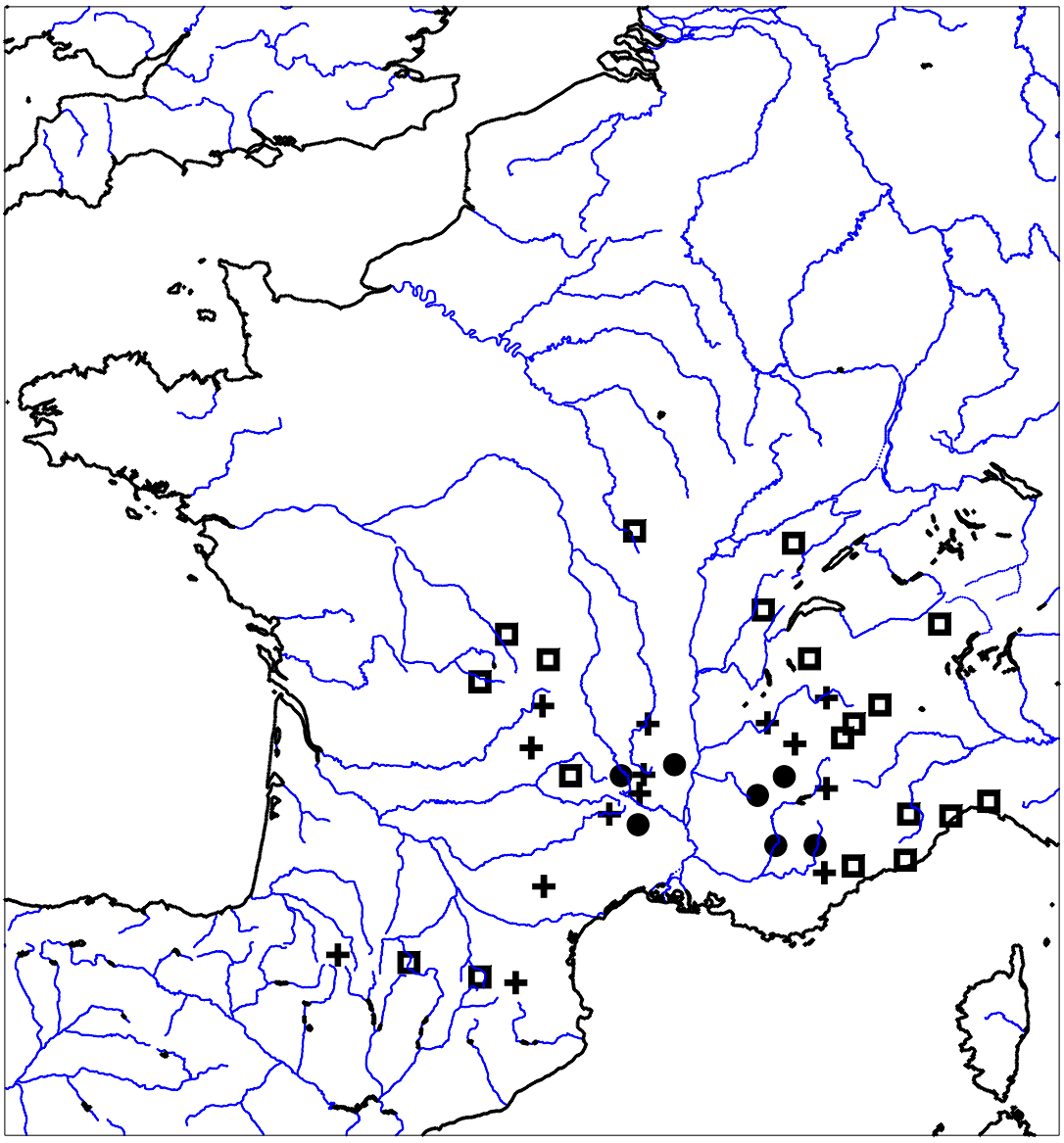
prévision en pluie/non pluie

4 fois ou + 3 fois 2 fois

prévision en classes

ANNEXE VI-5:

Carte de répartition des gains pour les bassins français et italiens prévision pluie/non pluie



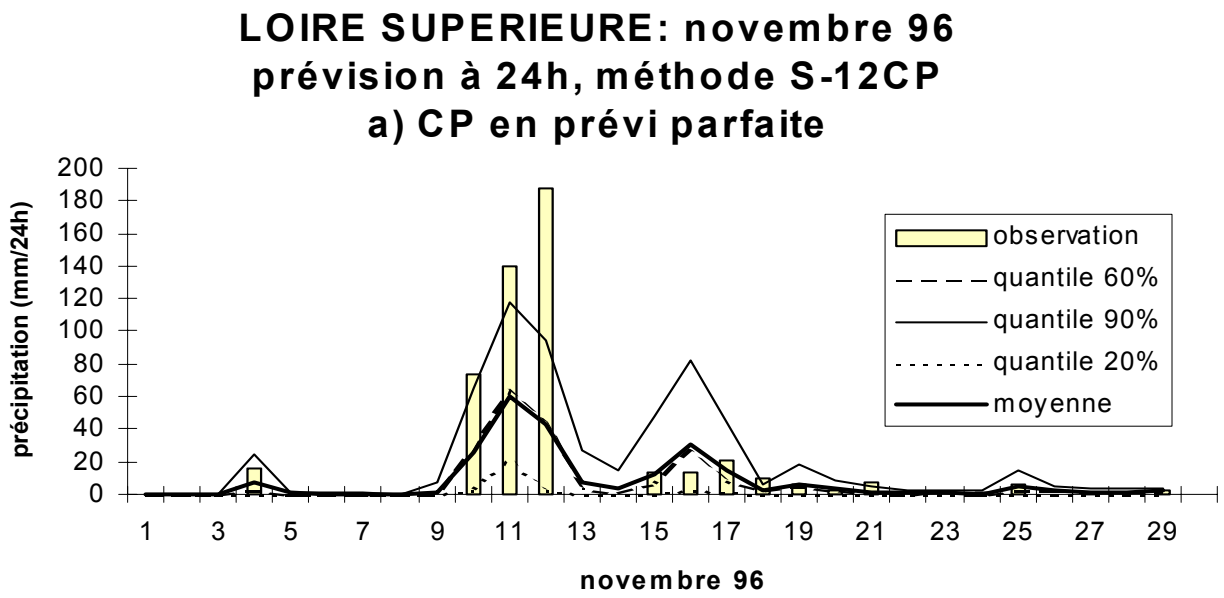
■ gain de IR > 2.5

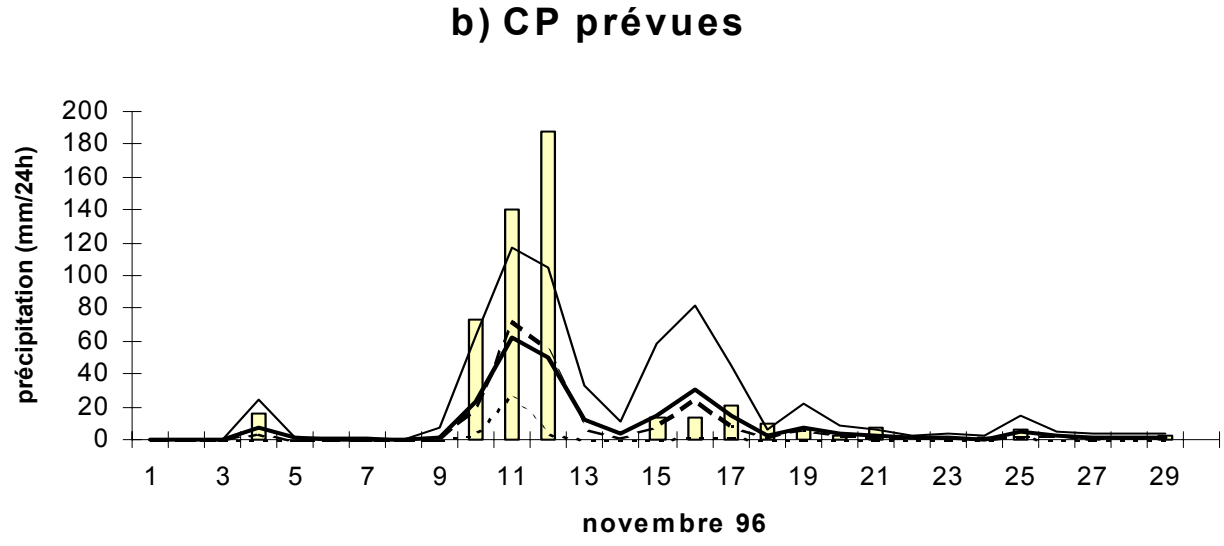
– 1.5 < gain IR < 2.5

□ gain de IR < 1.5

ANNEXE VI-6:

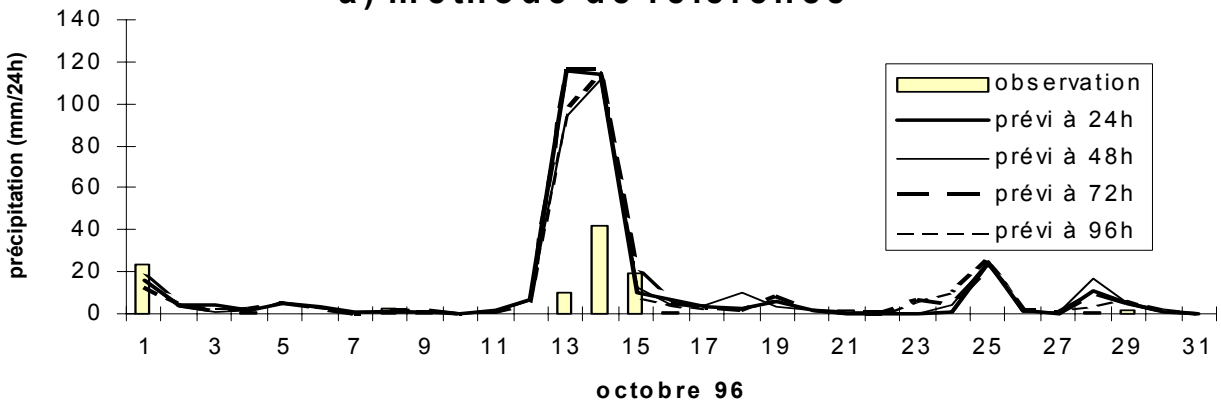
Loire supérieure, novembre 1996, méthode S-12CP



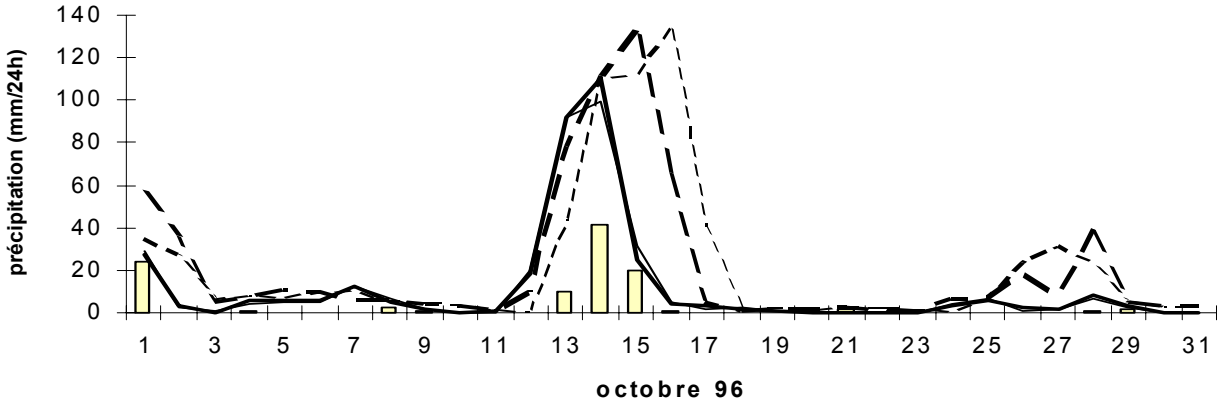


ANNEXE VI-7:

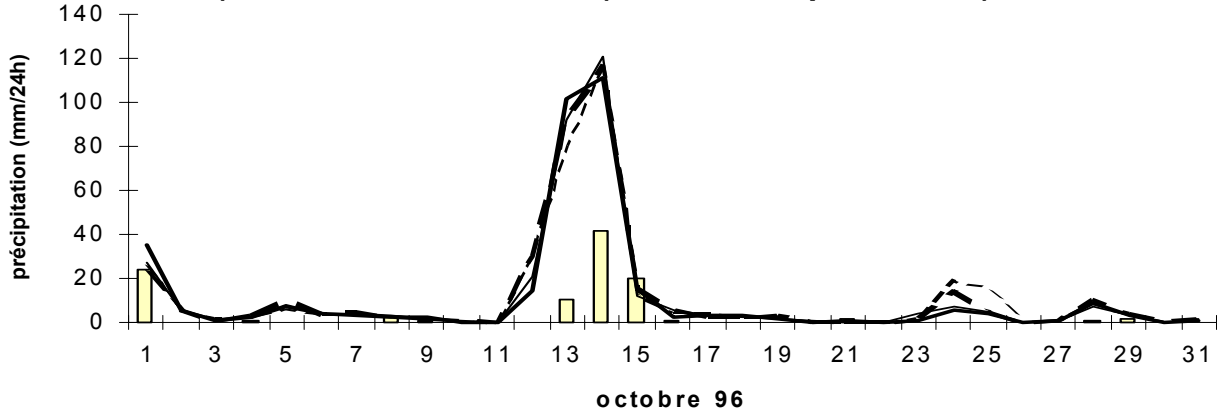
LOIRE SUPERIEURE: octobre 96
prévision à 24, 48, 72, 96 h (quantile 90%)
a) méthode de référence



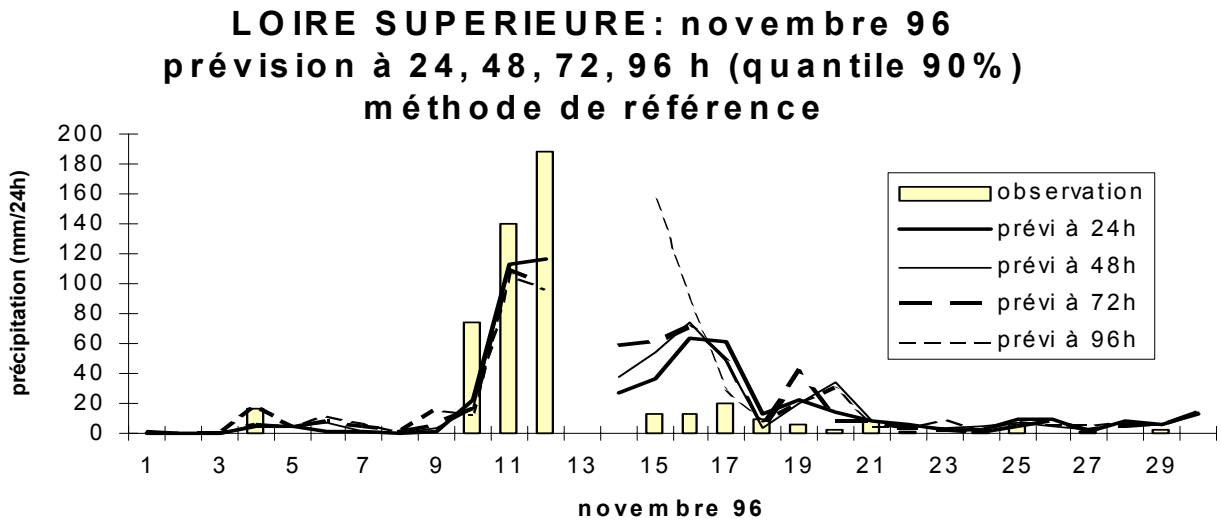
b) méthode S-12CP (CP prévues)

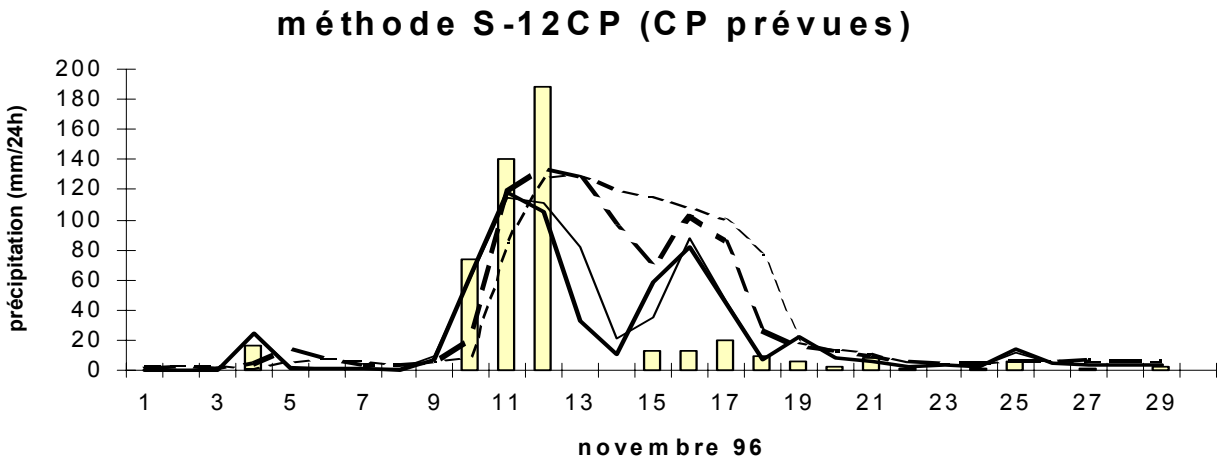


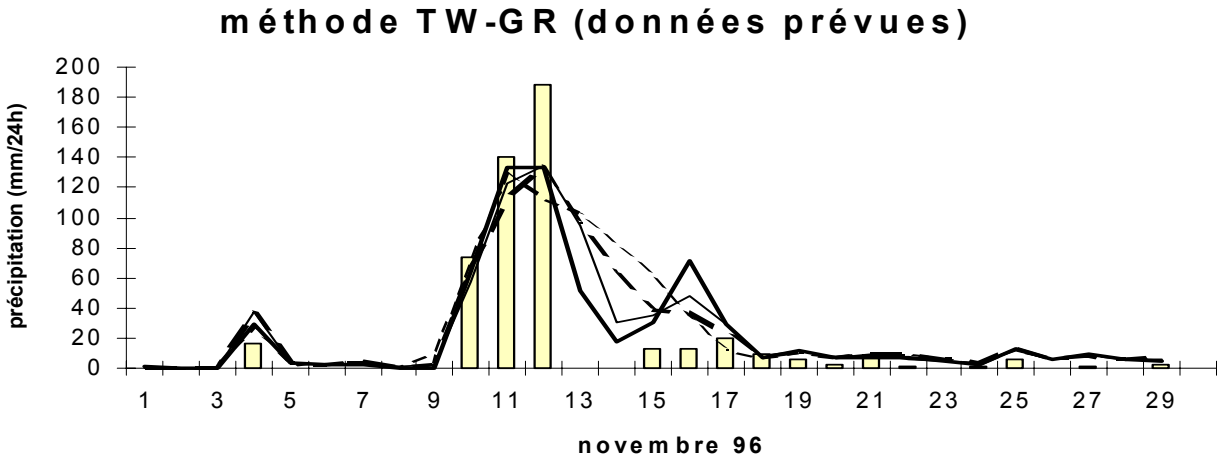
c) méthode TW-GR (données prévues)



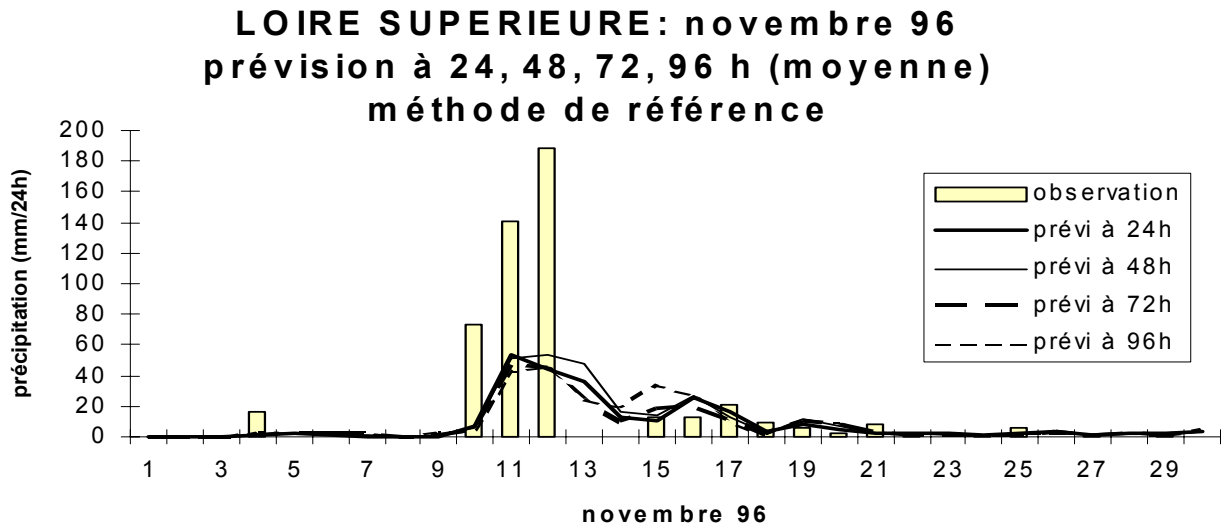
ANNEXE VI-8:

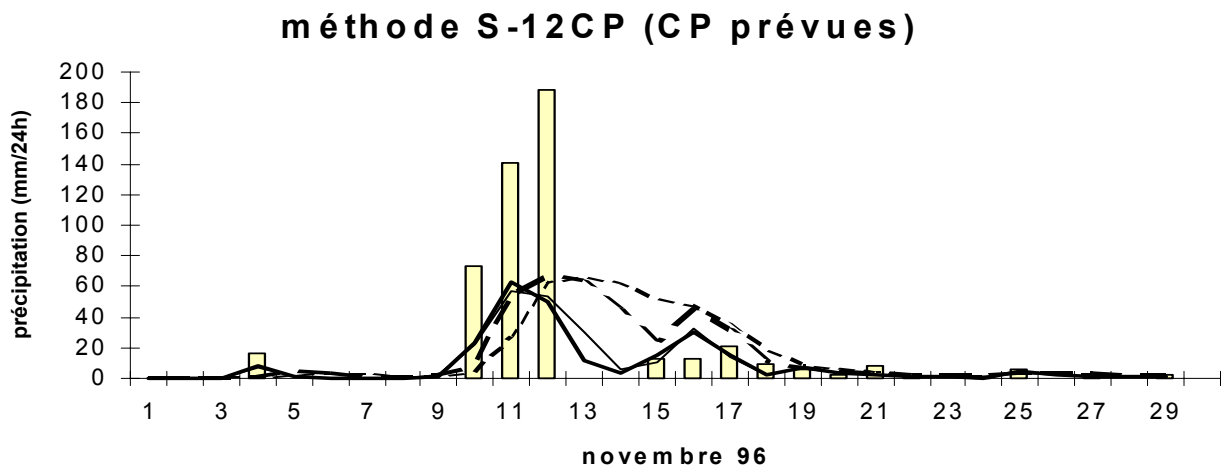

LOIRE SUPERIEURE: novembre 96
prévision à 24, 48, 72, 96 h (quantile 90%)
méthode de référence
précipitation (mm/24h)
observation
prévi à 24h
prévi à 48h
prévi à 72h
prévi à 96h
novembre 96

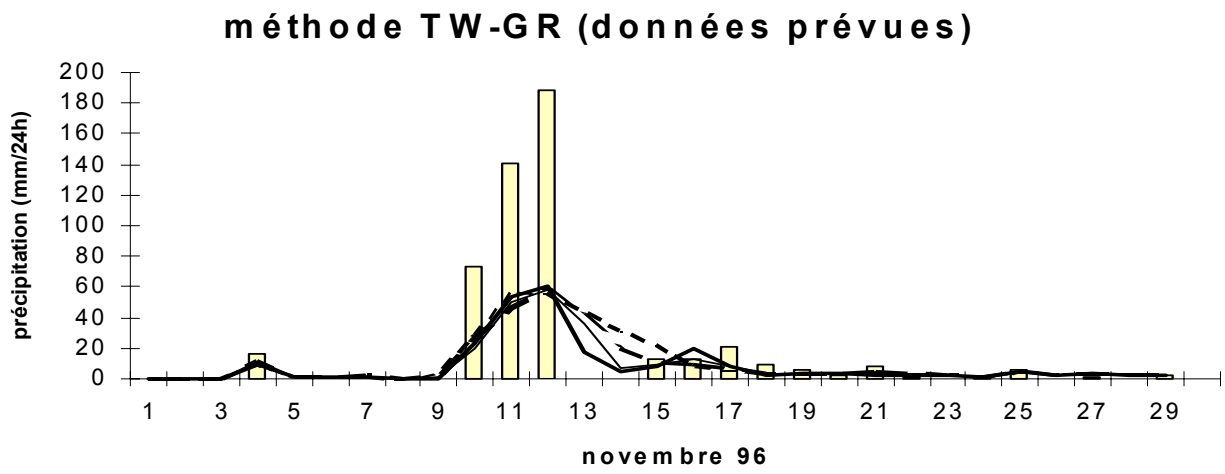

méthode S-12CP (CP prévues)
précipitation (mm/24h)
novembre 96


méthode TW-GR (données prévues)
précipitation (mm/24h)
novembre 96

ANNEXE VI-9:

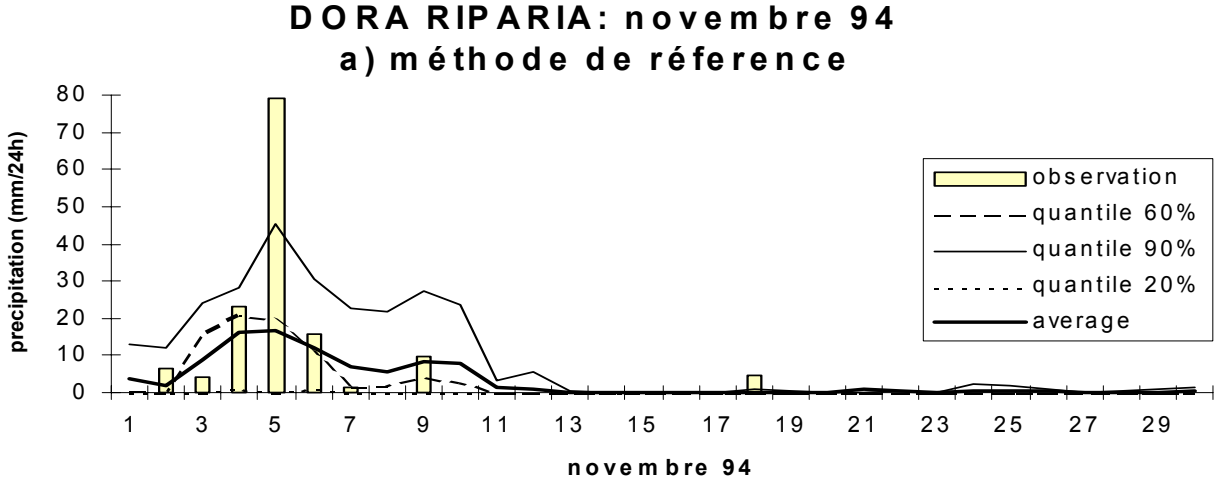

LOIRE SUPERIEURE: novembre 96
prévision à 24, 48, 72, 96 h (moyenne)
méthode de référence
précipitation (mm/24h)
observation
prévi à 24h
prévi à 48h
prévi à 72h
prévi à 96h
novembre 96

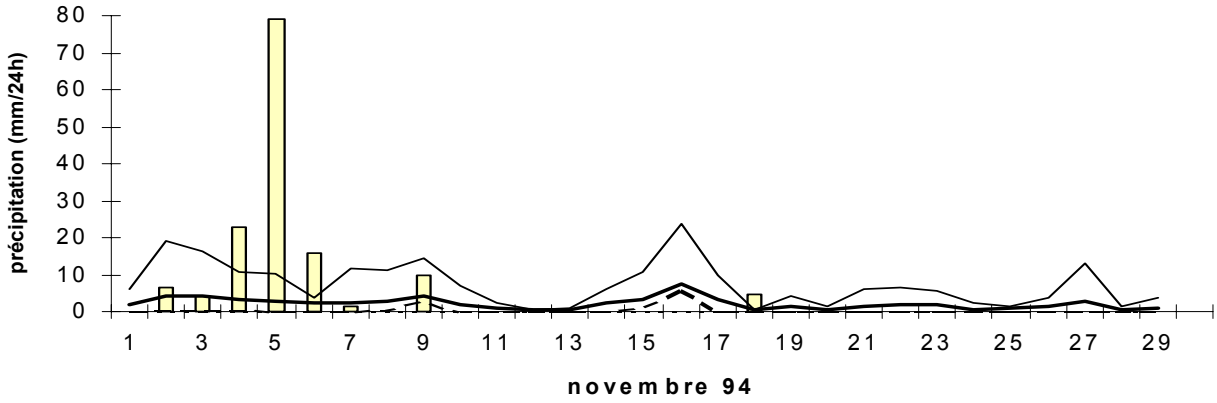

méthode S-12CP (CP prévues)
précipitation (mm/24h)
novembre 96

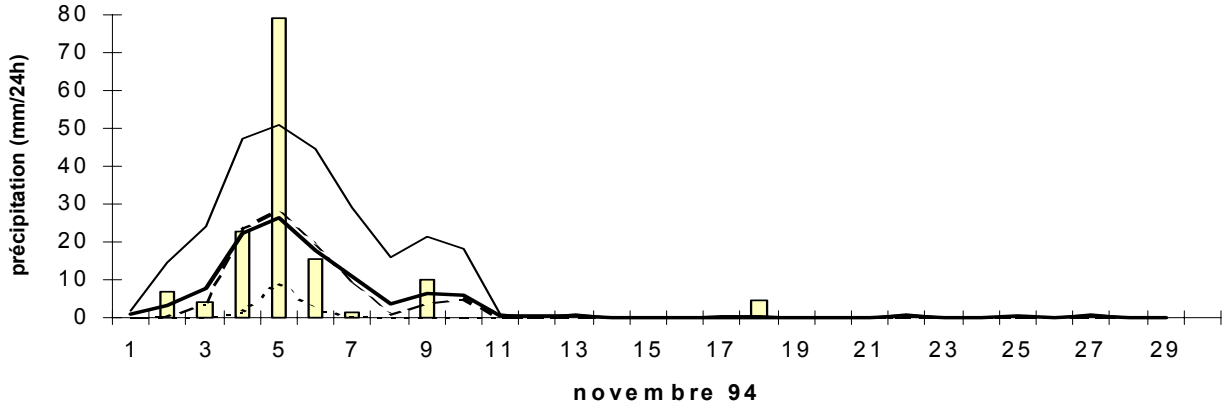

méthode TW-GR (données prévues)
précipitation (mm/24h)
novembre 96

ANNEXE VI-10:

Dora Riparia, novembre 1994, prévision à 24 h par les 3 méthodes







THESE DE DOCTORAT

Titre de l'ouvrage :

Prévision quantitative des précipitations journalières par une méthode statistico-dynamique de recherche d'analogues

Application à des bassins du pourtour méditerranéen

Nom de l'auteur :

Sophie GUILBAUD

Etablissement :

Institut National Polytechynique de Grenoble

RESUME

La méthode utilisée est une méthode statistique, reliant les pluies à la circulation atmosphérique issue d'un modèle météorologique. Elle est basée sur une technique de recherche d'analogue: des situations météorologiques similaires à la situation du jour sont extraites d'un fichier historique, contenant les champs de géopotentiels 1000 et 700 mb à 0h, condensés par Analyse en Composantes Principales. Puis, la prévision des précipitations pour 24h est effectuée à partir des précipitations observées lors des journées analogues.

Cette méthode a été élaborée par le Service Ressources en Eau d'EDF dans les années 70 sur des bassins montagneux français, pour assurer la sécurité de ses installations en cas de crue.

Cette étude a permis d'améliorer les performances de la méthode en reconsidérant les prédicteurs et la sélection des analogues qui était faite avec une distance euclidienne sur les 6 premières composantes principales (CP) du champ 700 mb.

Tout d'abord, on a vu qu'il fallait utiliser les CP des 2 champs (700 et 1000 mb) à 0 et 24h, et pas seulement les 6 premières. Ainsi une nouvelle distance, élaborée bassin par bassin, a apporté une nette amélioration.

Puis, un autre critère, le score de Teweles-Wobus, spécialement construit pour des champs en points de grille, a été utilisé pour sélectionner les analogues, ce qui a donné des résultats supérieurs à ceux obtenus avec la distance euclidienne.

Ensuite, sur divers prédicteurs synoptiques testés, seule l'utilisation de l'humidité en basses couches s'est révélée prometteuse. Quant à l'utilisation d'information locale, en deuxième niveau de sélection, elle est intéressante pour des bassins situés dans un rayon d'influence d'environ 200 km.

Enfin, une validation des meilleurs essais sur les 3 derniers automnes a confirmé les résultats obtenus en calage. Pour terminer, le modèle amélioré a été testé sur des bassins espagnols et italiens à risque.

MOTS CLES

Prévision météorologique, Précipitation journalière, Situations analogues, Critère de proximité, Reconnaissance des formes, Episodes convectifs méditerranéens

Sophie GUILBAUD **PREVISION QUANTTATIVE DES PRECIPITATIONS JOURNALIERES PAR UNE METHODE STATISTICO-DYNAMIQUE DE RECHERCHE D'ANALOGUES** **1997**

Application à des bassins du pourtour méditerrannéen