

Modélisation du Climat et Statistiques (2/3)

Statistiques Multivariées

Pascal Yiou

LSCE, Gif-sur-Yvette

Motivations

- Les données météorologiques ou océanographiques sont réparties en espace et en temps (champs)
 - Une carte par jour ou par mois
- L'information contenue dans un tel champ est très lourde à étudier
- Quelques « challenges »:
 - Peut-on réduire cette information à quelques paramètres?
 - Peut-on extraire des modes de variabilité temporelle à partir de structures de corrélations spatiales?

L'analyse en EOFs

- Détermination des "cartes" (EOFs) qui maximisent la variance du champ spatio-temporel $X(t,x)$, i.e. séparation temps-espace

$$X(t, x) = \sum_k a_k(t) E_k(x)$$

- Permet de réduire la dimensionalité des données et de séparer des mécanismes indépendants. En pratique, on ne garde que les K premières EOF représentant une partie de la variance totale.
- Utilisation courante en météorologie et océanographie.

Calcul des EOFs (1)

Pour un champ d'observations $X(t,x)$, on cherche un vecteur unitaire $E(x)$ qui **maximise** l'étendue de la projection de X sur E , i.e. **la variance de X expliquée par E** .

$$V_E = \frac{1}{N} \sum_t \sum_x |X(t,x)E(x)|^2$$

$$V_E = \frac{1}{N} \sum_t \left| \langle X(t,x), E(x) \rangle \right|^2 = \frac{1}{N} {}^t E X^t X E$$

$$V_E = {}^t E C E$$

$$C = \frac{1}{N} {}^t X X$$

Calcul des EOFs (2)

La première EOF est la solution de:

$$\max_{\|E\|=1} V_E = \max_{\|E\|=1} ({}^t E C E)$$

Algèbre... l'unique solution (au signe près) vérifie:

$$C E_1 = \lambda_1 E_1$$

λ_1 est la plus grande valeur propre de la matrice de covariance C , et E_1 est le vecteur propre associé: *le nuage de points $X(t,x)$ s'étend majoritairement dans la direction portée par E_1 .*

Les autres EOFs sont les vecteurs propres associés aux valeurs propres restantes.

Propriétés des EOFs (1)

Orthogonalité (non corrélation): par définition, les EOFs forment une base orthogonale. Leur corrélation (ou covariance) est nulle.

$$\langle E_k(x), E_{k'}(x) \rangle_x = {}^t E_k E_{k'} = \delta_{k,k'}$$

Coefficients temporels ou Composantes Principales (PC):

$$a_k(t) = \langle X(t, x), E_k(x) \rangle_x = X E_k$$

Propriétés des EOFs (2)

Covariance des PCs et variance « expliquée »:

$$\begin{aligned}\text{cov}(a_k(t), a_{k'}(t)) &= \langle a_k(t), a_{k'}(t) \rangle_t, \\ &= \langle XE_k, XE_{k'} \rangle_t, \\ &= {}^t E_k \underbrace{{}^t X X}_{C} E_{k'}, \\ &= \lambda_k ({}^t E_k E_{k'})\end{aligned}$$

En pratique: On conserve les k EOFs/PCs qui expliquent plus de 80% de la variance totale (par exemple).

Remarques

- On peut calculer les EOFs/PCs sur la matrice de covariance ou de corrélation. Le résultat n'est pas le même:
 - La matrice de corrélation permet d'identifier des structures de même variance
 - La matrice de covariance favorise l'étude de structures de grande variance.
- Il est parfois nécessaire de pondérer la matrice de corrélation
 - Aires des grilles spatiales (cos de latitude)

EOFs en résumé

- Les EOFs d'un champ d'observations sont les vecteurs propres de la matrice de covariance.
- Les PCs sont les coefficients temporels des EOFs.
- Les variances des PCs sont les valeurs propres de la matrice de covariance.
- Les observations peuvent être des séries hétérogènes (i.e., température et précipitation à un endroit donné) ou bien une répartition spatiale de données homogènes (e.g. température sur une grille).
- Une décomposition en EOF/PC est utile si la variance totale des observations est représentable sur *peu* d'EOFs.

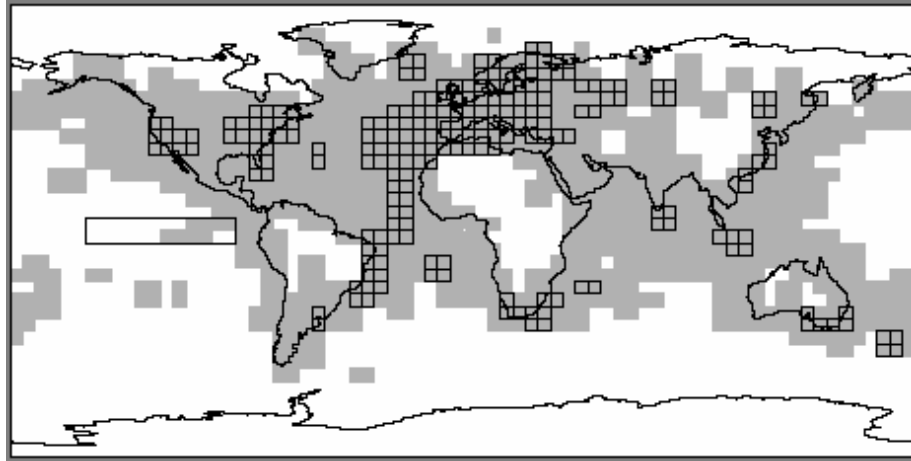
Applications & Examples

Reconstruction multi-proxy de températures

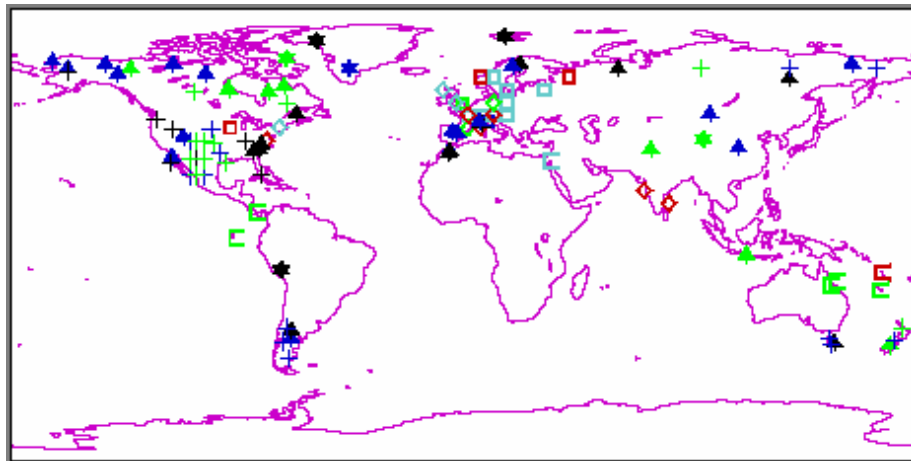
- Comment obtenir une reconstruction spatiale de température au cours du dernier millénaire?
- Base de données de températures depuis 1850
- Base de données dendro, isotopes, archives... sur le dernier millénaire

Températures observées et reconstruites

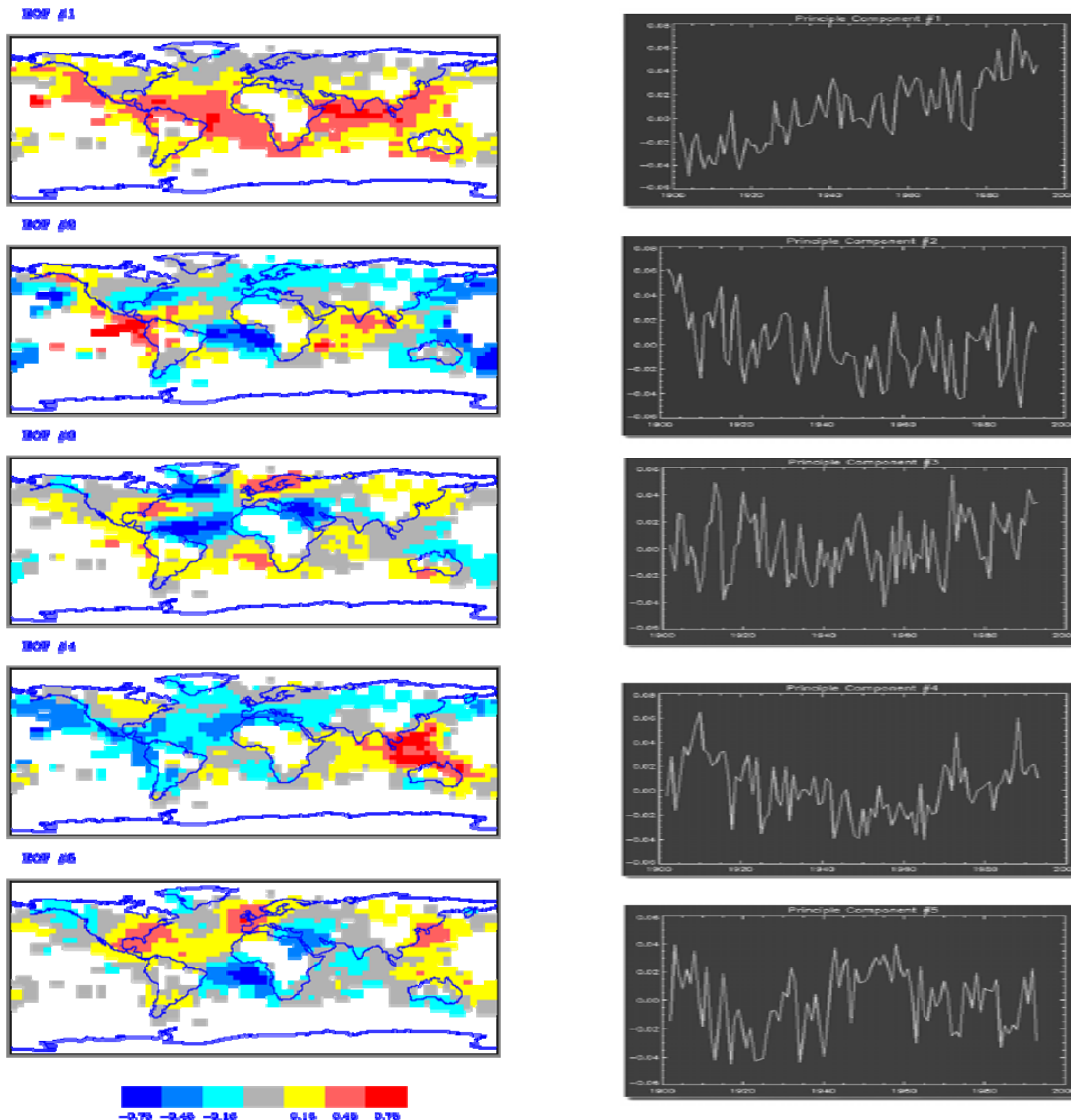
INSTRUMENTAL
TEMPERATURE
RECORD



GLOBAL
PROXY
CLIMATE
RECORDS



Premières EOFs de températures annuelles



Five leading patterns of global temperature variation during the 20th century.

Mann, M.E., Bradley, R.S., Hughes, M.K., Global-Scale Temperature Patterns and Climate Forcing Over the Past Six Centuries, *Nature*, 392, 779-787, 1998.

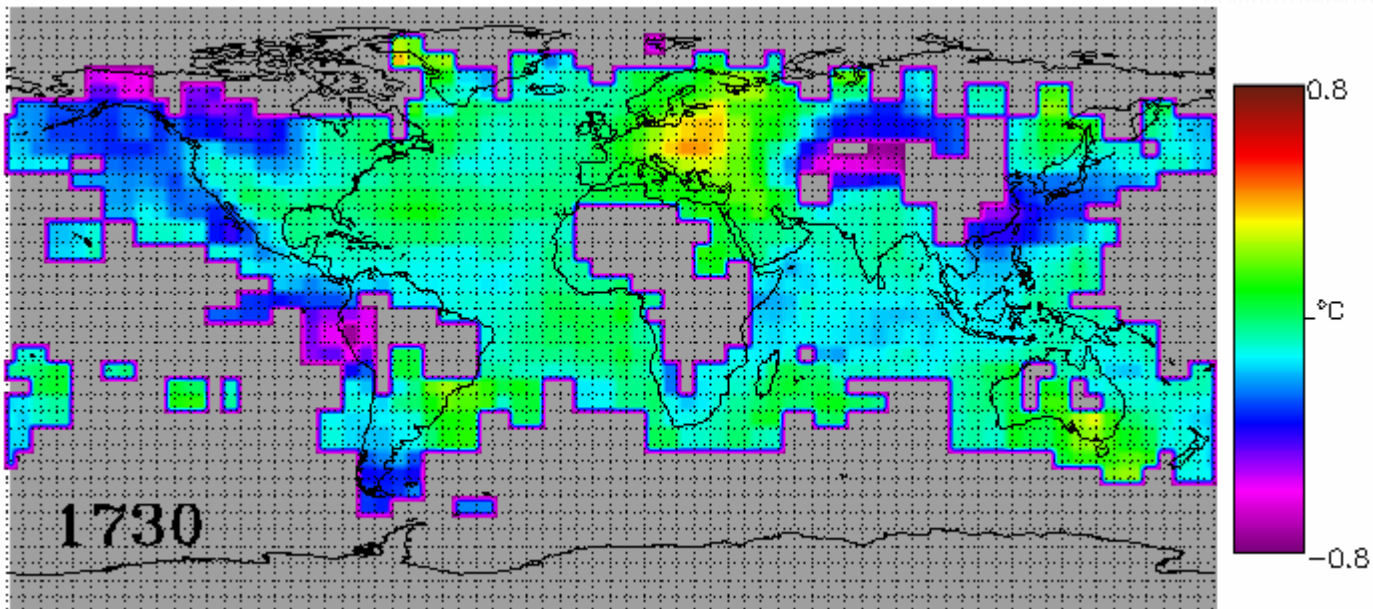
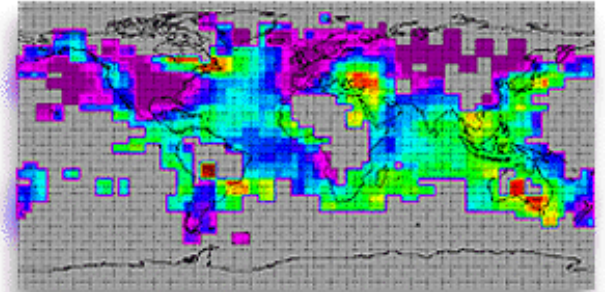
Stratégie de Mann et al. 1998

- Calculer les PCs/EOFs de la température observée depuis 1850 sur le globe, l'hémisphère nord, l'hémisphère sud
 - Les 5 premières EOF expliquent 83% de la variance.
- Faire une régression des proxys (cernes d'arbres, sédiments, coraux...) pour chaque PC de température sur une période d'apprentissage (20ème siècle).
- Vérifier la reconstruction sur 1850-1900.
- Détermination de la reconstruction sur le dernier millénaire, avec calcul d'intervalles de confiance liés aux données manquantes.

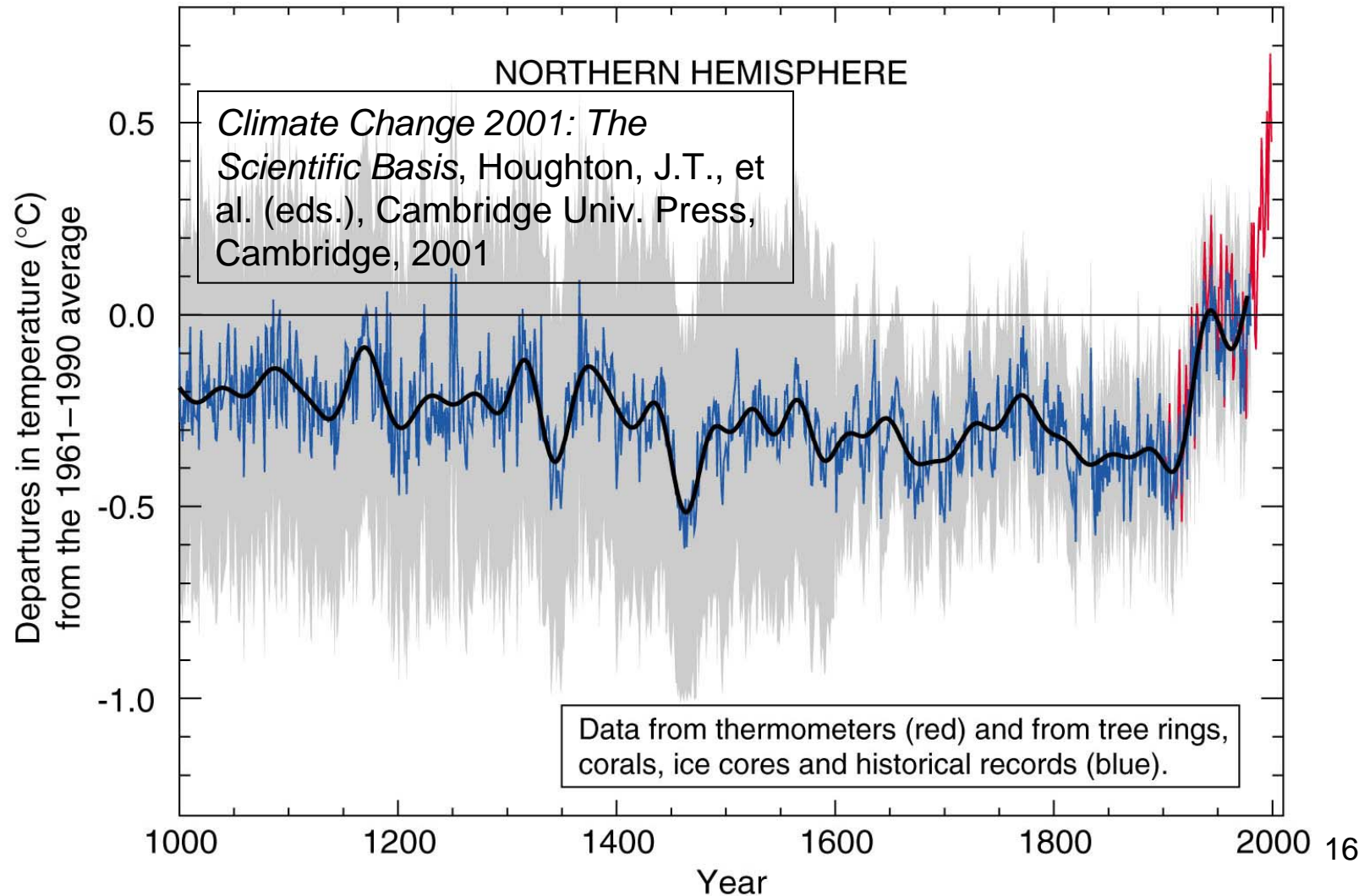
Exemples

Après l'éruption du Tambora →

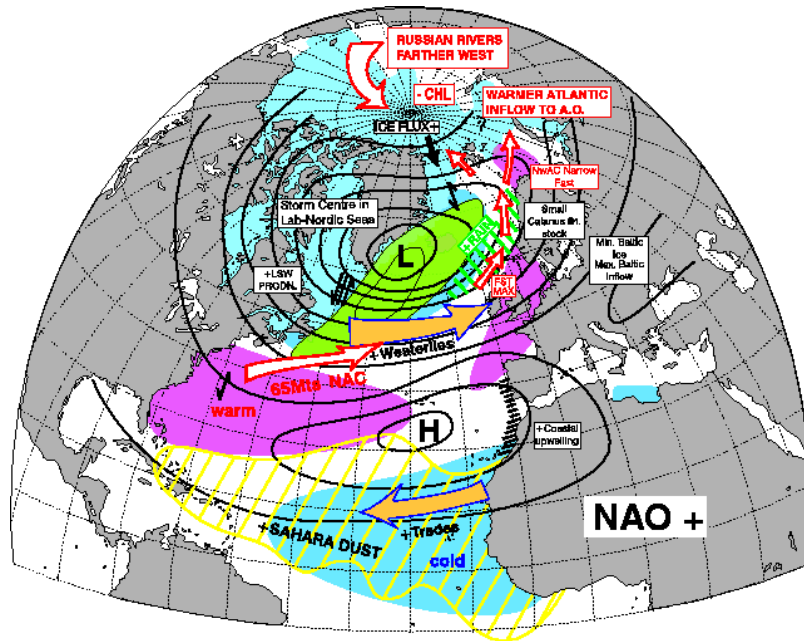
1816
("A Year Without A Summer")



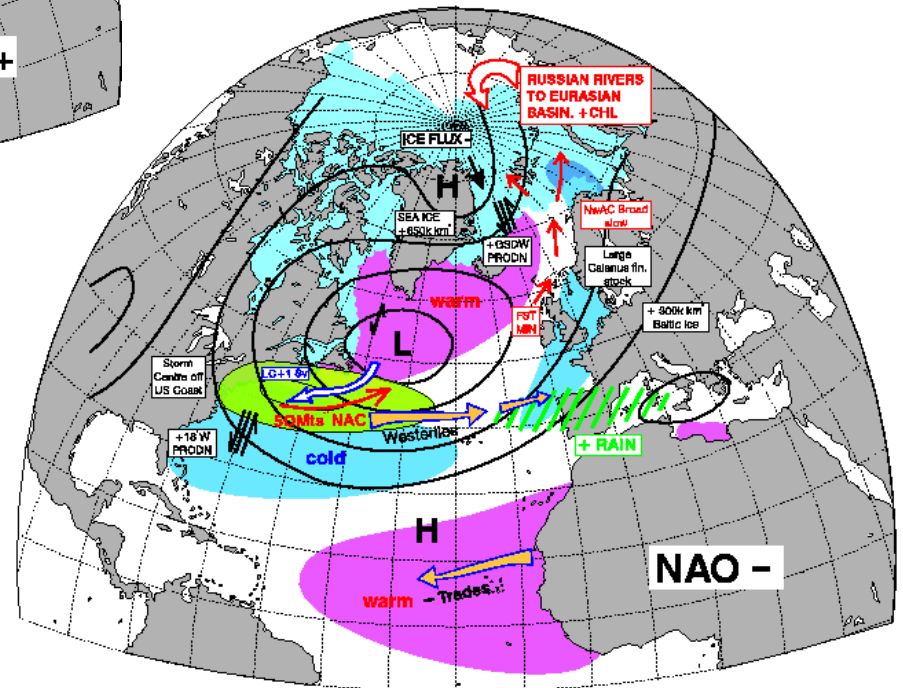
Températures de l'hémisphère nord



L'Oscillation Nord Atlantique

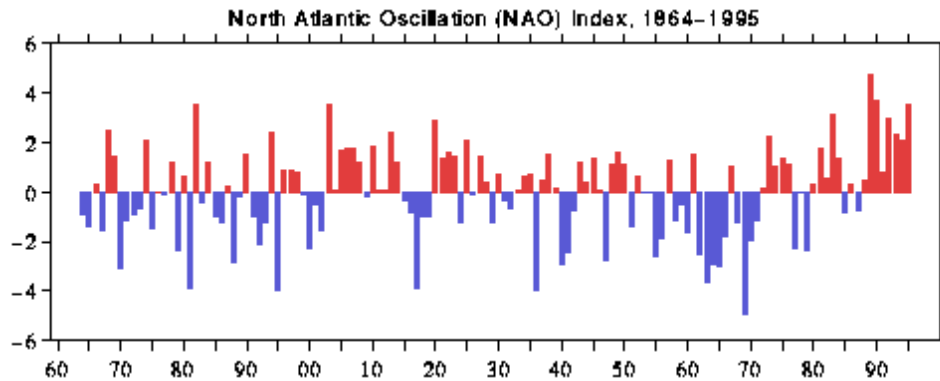


Les deux phases de la NAO
(D. Stephenson)



Indice NAO

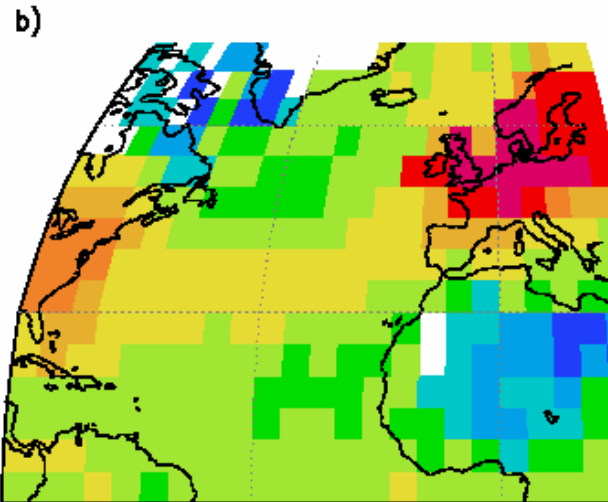
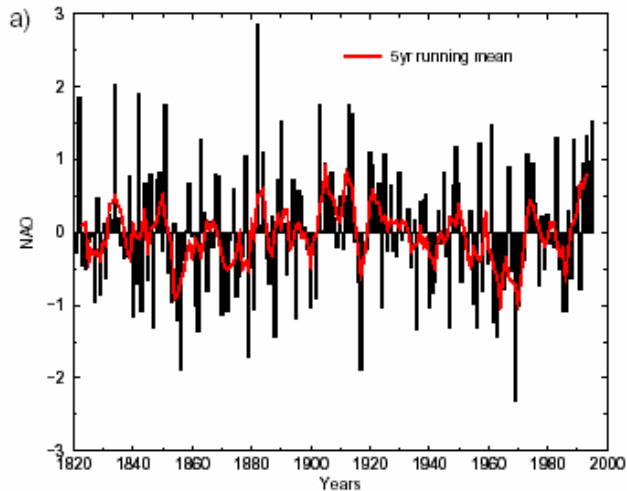
Différence de pression normalisée entre les Açores et l'Islande



Vent géostrophique:
$$V_g = \frac{1}{\rho f} \vec{k} \times \text{grad}P$$

L'indice NAO exprime l'intensité du vent zonal (vers l'est) dans les extra-tropiques. Il peut être défini comme une EOF de pression de surface.

Le Climat de l'Atlantique Nord

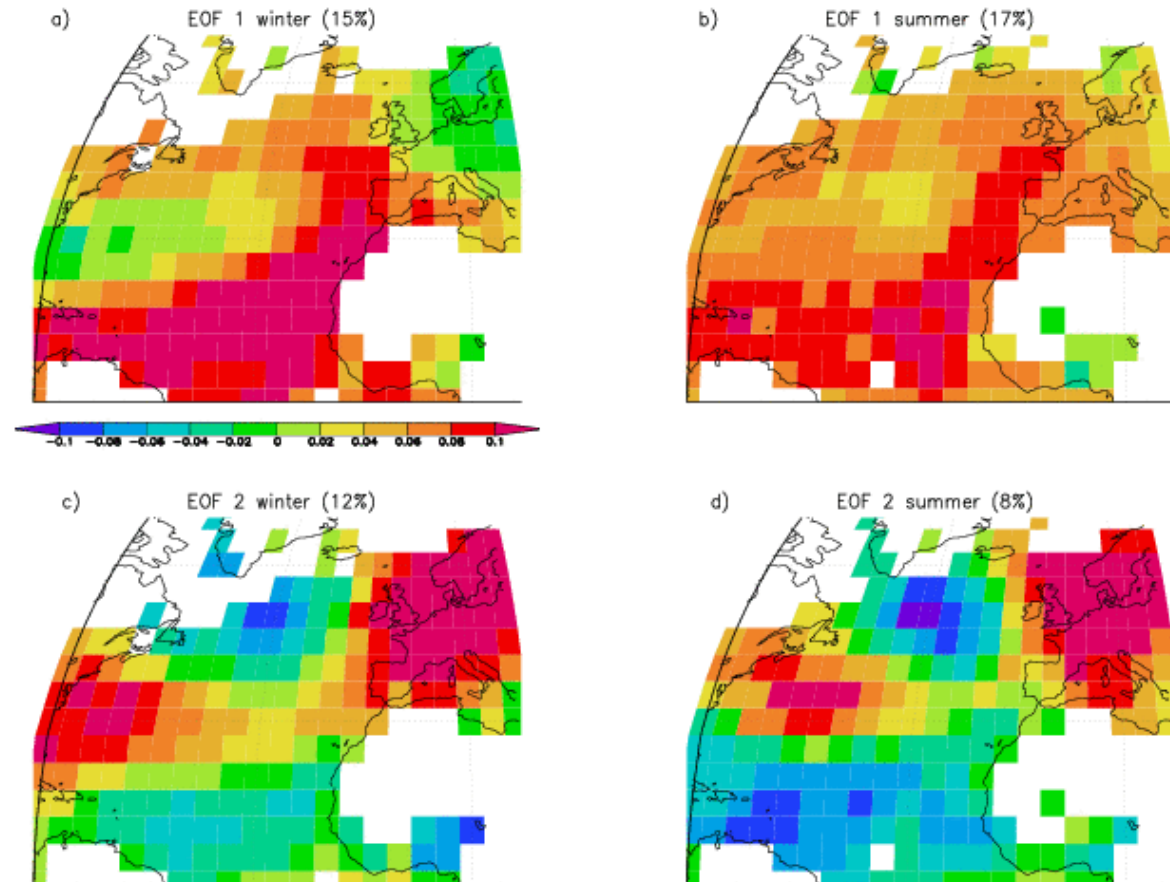


Corrélation entre NAO et température de surface:

- Influence de la circulation atmosphérique sur les températures de surface
- Structure quadripolaire

Le Climat de l'Atlantique Nord

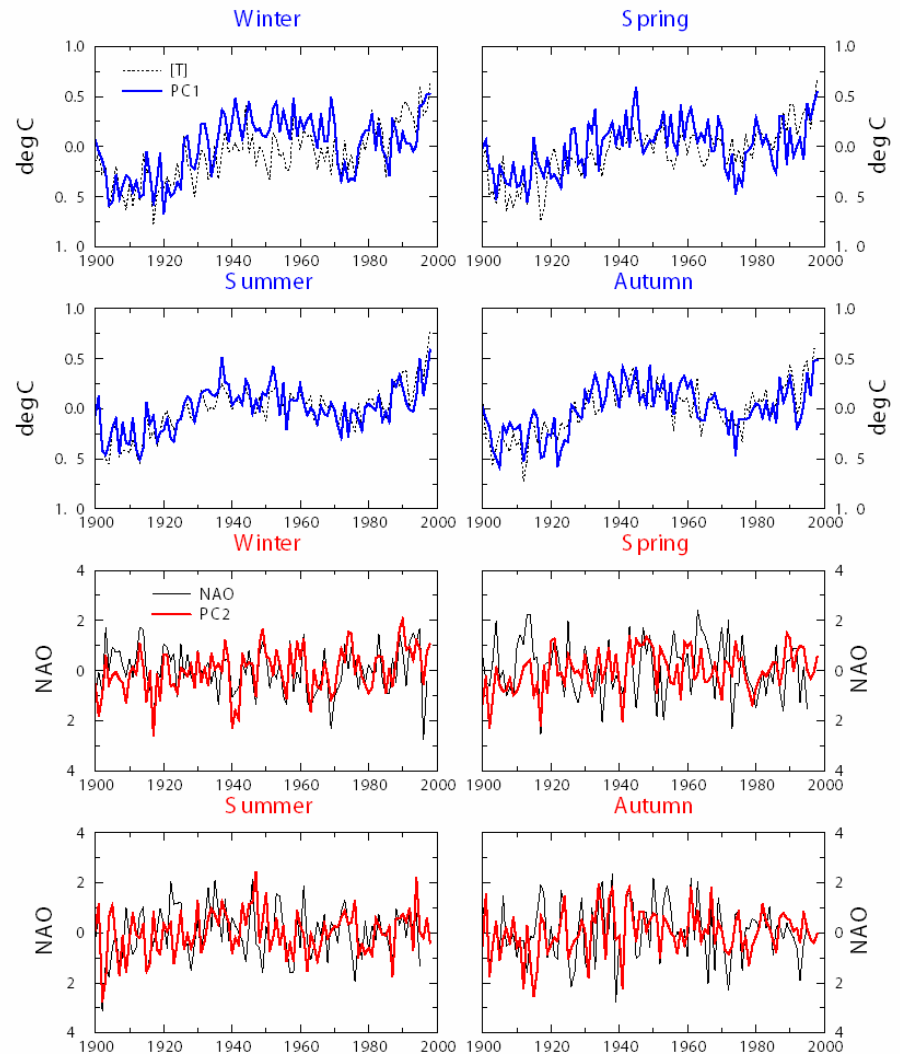
EOFs des températures de surface observées au 20ème siècle



- a) EOF^T 1 winter,
- b) EOF^T 1 summer,
- c) EOF^T 2 winter,
- d) EOF^T 2 summer.

Climat de l'Atlantique Nord

Première PC de la température de surface: réchauffement global



Deuxième PC de la température de surface: indice NAO

Utilisation des EOFs

- Réduire la dimension des observations à 2 ou 3, tout en conservant un pourcentage significatif de variance totale.
 - Étape indispensable avant certains types d'analyses (M-SSA, classification de champs multivariés...).
- Signification physique éventuelle des « modes » de variance.
 - Exemple: *l'Oscillation Arctique* définie comme la première EOF du z500 dans l'hémisphère nord.
- Calcul de variables composites entre plusieurs types d'observations
 - Exemple: séries de plusieurs isotopes dans une même carotte sédimentaire

Comparer deux champs

- Analyse en corrélation canonique (CCA)
 - Pour deux champs $X(t,x)$ et $Y(t,y)$, on cherche deux combinaisons linéaires de ces champs dont la *corrélation* soit maximale
- Analyse en décomposition de valeurs singulières (SVD)
 - Pour deux champs $X(t,x)$ et $Y(t,y)$, on cherche deux combinaisons linéaires de ces champs dont la *covariance* soit maximale

Analyse en Corrélation Canonique (CCA)

X est un champ en m_X dimensions et Y est un champ en m_Y dimensions. On cherche deux vecteurs f_X et f_Y de dimensions respectives m_X et m_Y (i.e. des cartes), tels que les coefficients temporels $\beta^X = \langle X, f_X \rangle$ et $\beta^Y = \langle Y, f_Y \rangle$ ont une corrélation maximale.

$$\begin{aligned}\rho &= \frac{\text{Cov}(\beta^X, \beta^Y)}{\sqrt{\text{Var}(\langle X, f_X \rangle) \text{Var}(\langle Y, f_Y \rangle)}}, \\ &= \frac{f_X^T \text{Cov}(X, Y) f_Y}{\sqrt{\text{Var}(\langle X, f_X \rangle) \text{Var}(\langle Y, f_Y \rangle)}}\end{aligned}$$

$$\text{Var}(\langle X, f_X \rangle) = f_X^T \Sigma_{XX} f_X = 1$$

$$\text{Var}(\langle Y, f_Y \rangle) = f_Y^T \Sigma_{YY} f_Y = 1$$

CCA

- Problème de maximisation sous contrainte (norme unitaire des vecteurs)
- Solution sous forme de valeurs et vecteurs propres.
- La corrélation optimale est la racine carrée de la plus grande valeur propre de:

$$M = \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^T$$

Quelques propriétés

- La maximisation de la corrélation permet de trouver des « patterns » f_X et f_Y de variances équivalentes
- On peut aussi s'intéresser à une maximisation de la covariance (résultats différents)
- Les coefficients temporels (β_i^X) sont décorrélés, les « patterns » peuvent être corrélés (différence avec les EOFs et PCs).

Combinaison avec les EOFs

- Quand la dimension est très grande, on a intérêt à réduire l'information des deux champs sur quelques EOFs/PCs, et de faire une CCA sur les quelques PCs.

EOFs et CCA

Décomposition en EOFs de X et Y:

$$X \approx \sum_{i=1}^{k_X} \alpha_i^{X+} e_X^{i+},$$

$$Y \approx \sum_{i=1}^{k_Y} \alpha_i^{Y+} e_Y^{i+}.$$

La CCA est appliquée sur les vecteurs:

On retrouve les « patterns » dans l'espace de départ:

Normalisation par l'écart type:

$$\alpha_i^+ = (\lambda_i)^{-1/2} \alpha_i,$$

$$e^{i+} = (\lambda_i)^{1/2} e_i,$$

$$X' = (\alpha_1^{X+}, \dots, \alpha_{k_X}^{X+}), \text{ et}$$

$$Y' = (\alpha_1^{Y+}, \dots, \alpha_{k_Y}^{Y+})$$

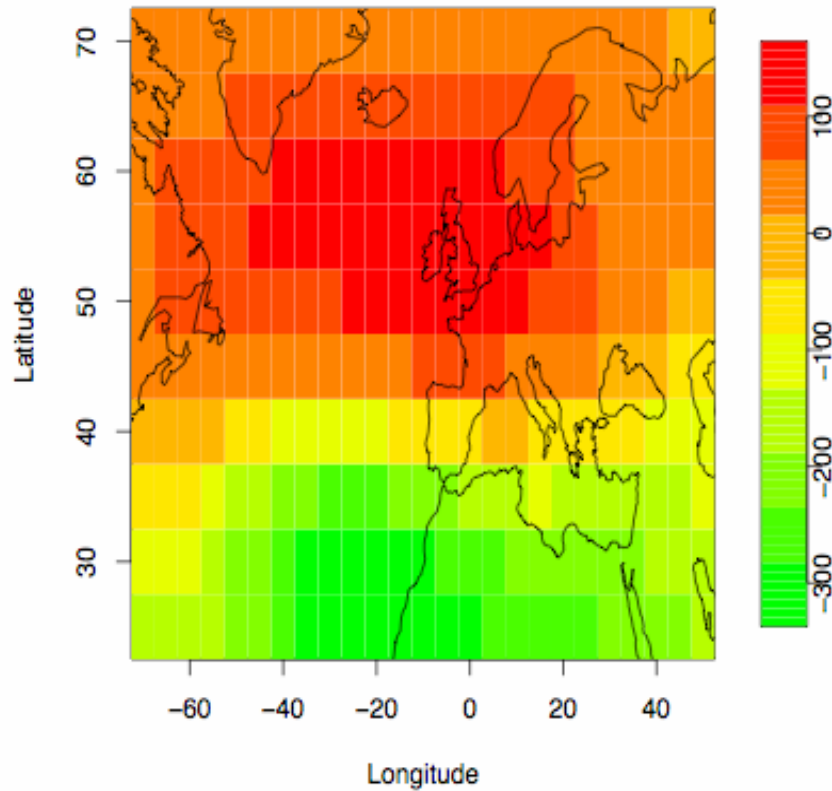
$$f_X^i = \sum_{j=1}^{k_X} (\lambda_j^X)^{1/2} (f_{X'}^i) e_X^j$$

Exemple: SLP et SST

- SLP et SST de 1856 à 2003 en moyennes mensuelles
- Analyse sur DJF
- Conservation des 10 premières PCs pour les deux champs
 - 96% de la variance de SLP
 - 83% de la variance de SST

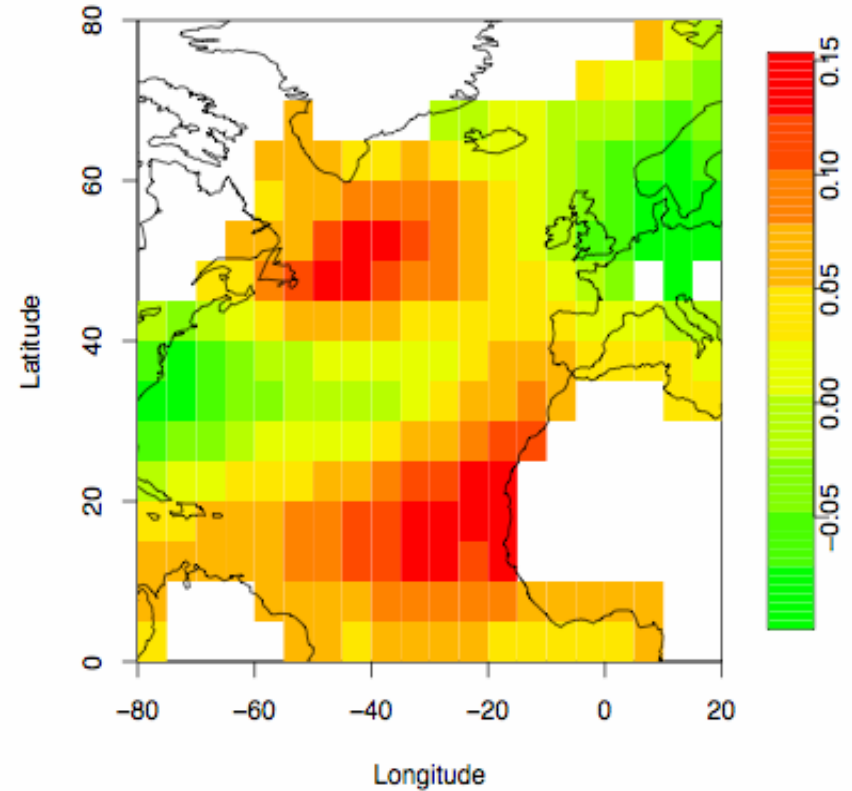
Exemple: SLP et SST

SLP EOF1



35% de la variance

SST EOF2



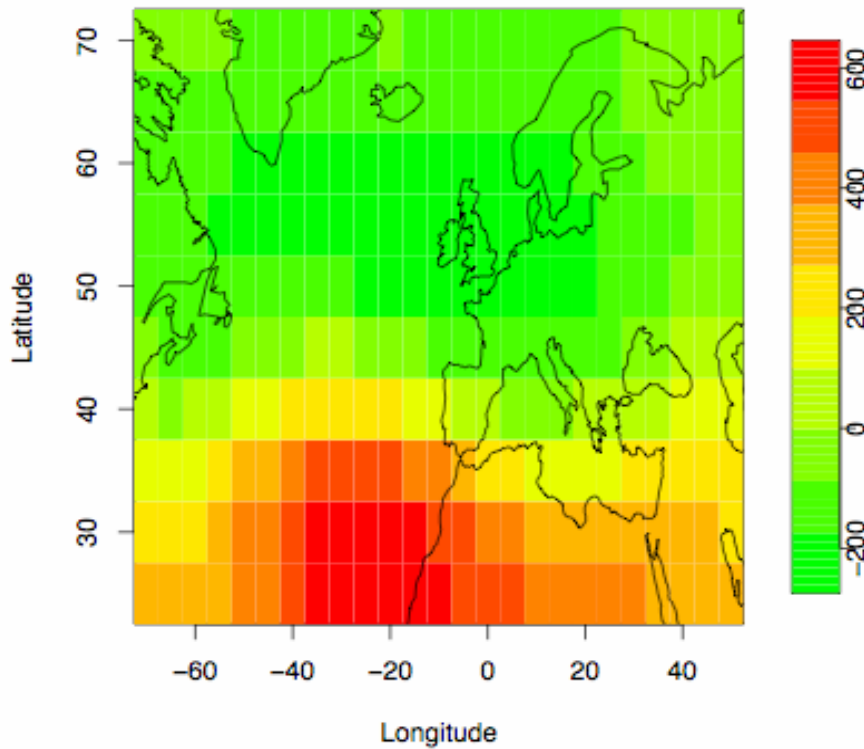
17% de la variance

$R=-0.38$

Patterns CCA de SLP et SST

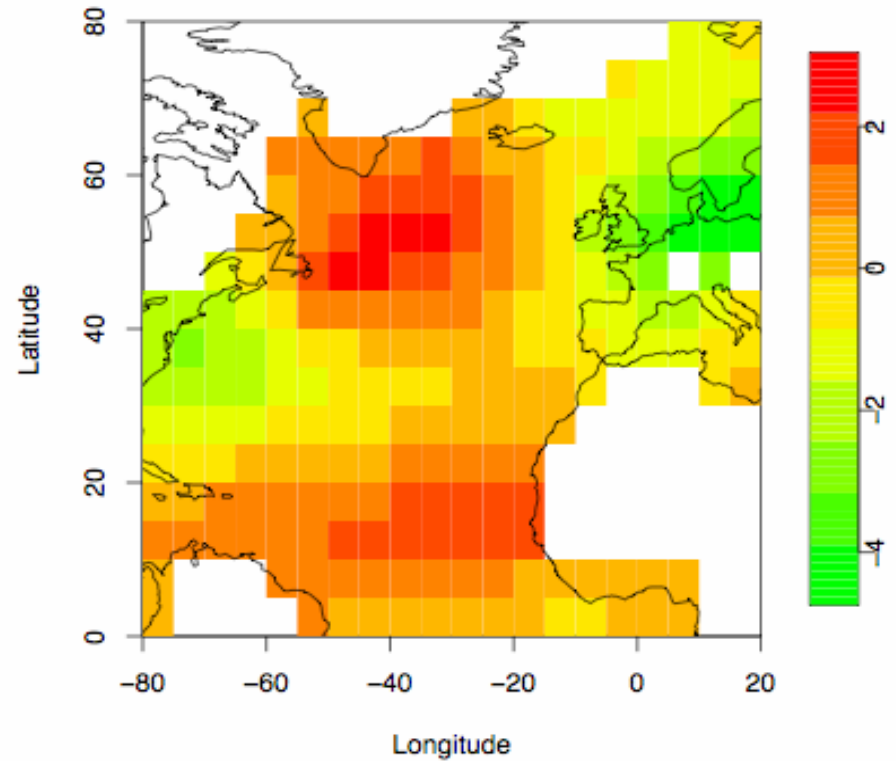
$R1=0.61$

SLP Pattern 1



29% de la variance

SST Pattern 1



17% de la variance

Interprétation

- Mise en évidence d'un mode couplé entre SST et SLP:
 - Mode NAO de SLP et mode en « fer à cheval » de SST
- Si les modes sont les mêmes au cours de l'année, la corrélation maximale est atteinte en hiver

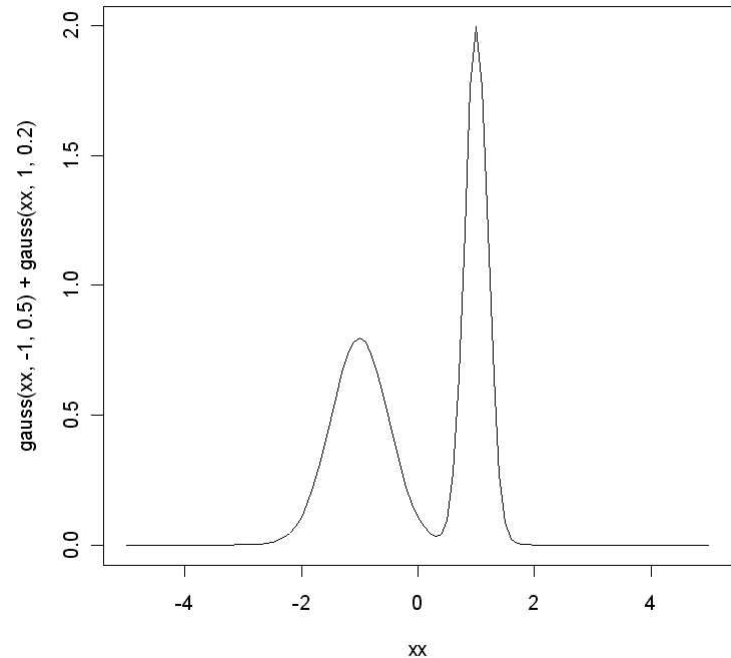
Remarques techniques

- Les techniques d'EOF ou de CCA ne sont pas spatiales au sens strict:
 - Les points sont traités indépendamment de leurs distances respectives (une permutation des points ne change ni les PCs ni les EOFs)
- Les critères de sélection des EOFs ne sont pas triviaux
 - En particulier dans le cas de champs non gaussiens comme les précipitations

Classification

La moyenne et l'écart-type peuvent être de mauvaises description d'une variable aléatoire.

Existence de bi-modalité de la distribution.



En 1 ou 2D, un histogramme suffit. S'il faut échantillonner P catégories dans chaque dimension, il faut de l'ordre de P^D observations en dimension D .

Exemple:

$D=1, P=20$: $O(20)$ obs

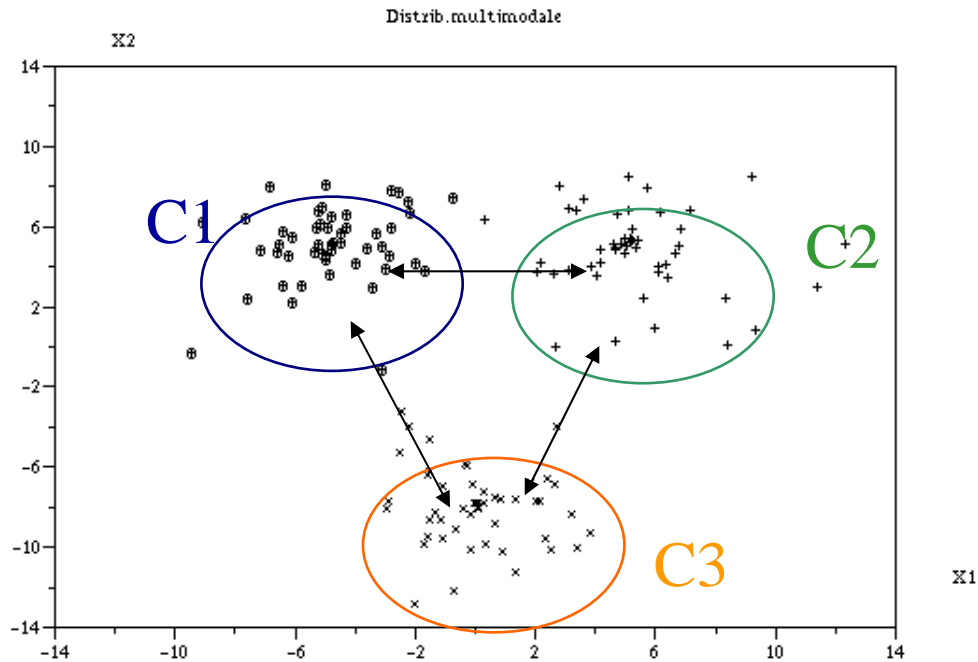
$D=2, P=20$: $O(400)$ obs

$D=5, P=20$: $O(3200000)$ obs!

Classification

- Un histogramme en grande dimension n'est donc pas un bon outil.
- On va s'intéresser à la détermination des maxima de la densité de proba.
- **Classification:** on possède une suite d'observations $\{X(t,x)\}$, et on cherche à savoir si cette suite s'agglomère autour d'un petit nombre d' « états » préférentiels, les modes de distribution de X .

Classification



Les questions qu'on se pose:

- Combien d'états pour bien représenter les observations?
- Persistance de chaque état?

Classification

On part d'une série multivariée \mathbf{X} . On agrège ses valeurs autour de K classes.

Exemples de techniques:

- Algorithme *k-means*: on cherche K classes en groupant les observations, par itérations, en minimisant les variances de chaque groupe (e.g. Michelangeli et al., *J. Atmos. Sci.*, 1995)
- *Mixture modeling*: on représente le « nuage » de points par une juxtaposition de Gaussiennes multivariées, de tailles et d'orientations différentes (Smyth et al., *J. Atmos. Sci.*, 1999)

Dans tous les cas, il faut chercher le nombre optimal de groupes de manière plus ou moins heuristique!

L'algorithme k-means

- Choix a priori d'une décomposition en K classes
- Initialisation (aléatoire ou non) de K centres de classes (*centroïdes*)
- Itérations
 - Attribution d'une classe à chaque observation du champ $X(x,t)$ par minimisation d'une distance ad hoc (e.g. Euclidienne)
 - Calcul des centres de chaque classe
 - Itération jusqu'à convergence des centres des classes

Les classes (ou nuées, clusters, régimes...) sont donc déterminées de manière itérative et dynamique

Les classes représentent les « maxima » de la fonction de probabilité des observations dans un espace à grande dimension

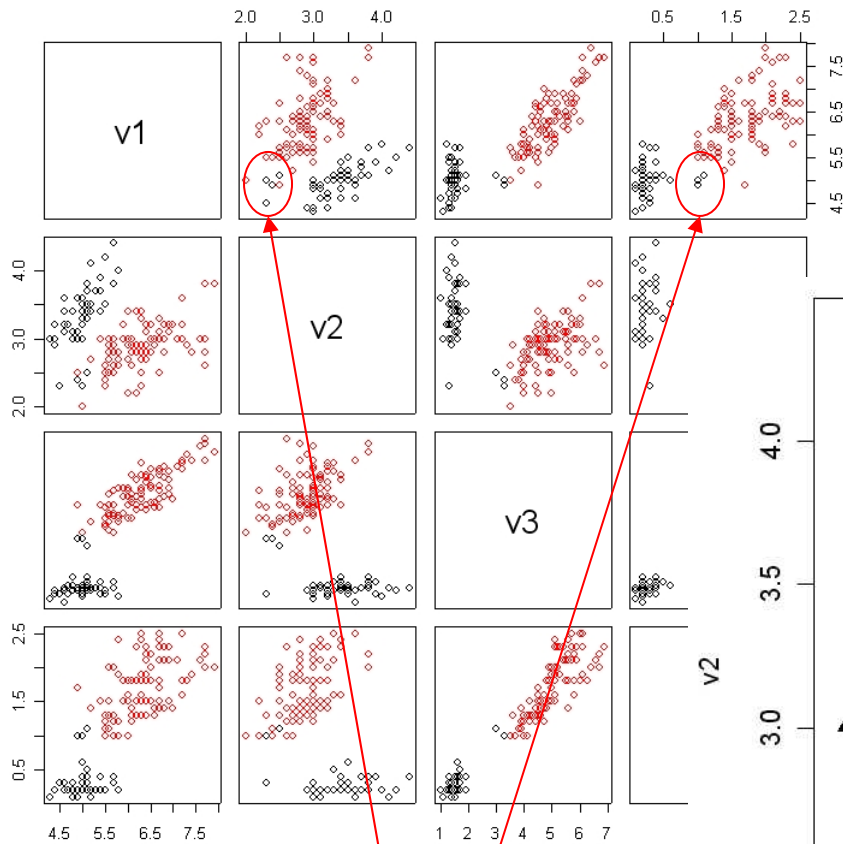
Remarques sur k-means

- **Nécessité de réduire l'espace des phases**
 - E.g. analyse en EOFs
- Importance du temps dans la classification
 - le mélange des données n'influence pas la décomposition, en principe! En pratique, il faut considérer que le résultat (les centroïdes) dépend de l'initialisation « aléatoire » de l'algorithme, et du parcours du champ.
 - Critère de classifiabilité sur le caractère reproductible des centroïdes
- Le choix de K (nombre de classes) est la partie la plus difficile de la décomposition.
 - Critères « pseudo objectifs » de classifiabilité du bruit rouge
 - Connaissance *a priori* du nombre de régimes
- Description en terme de sauts d'états vs. oscillations:
 - « granularité » vs. « ondes »

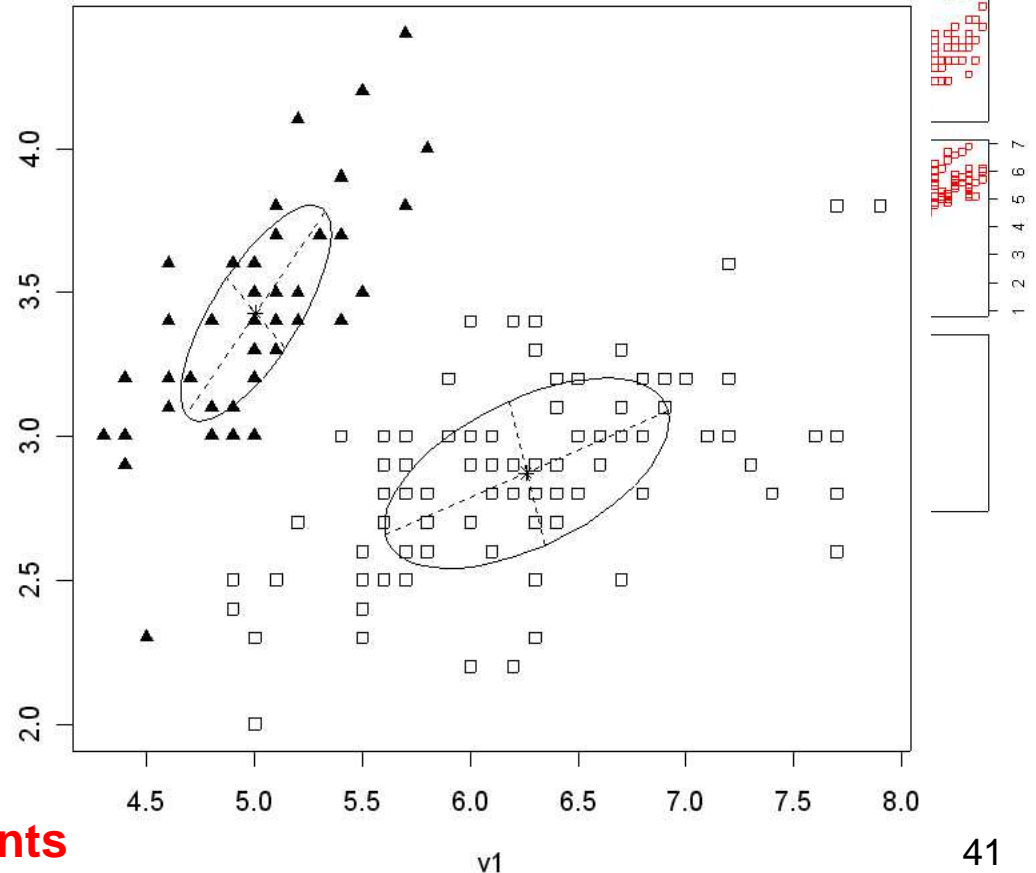
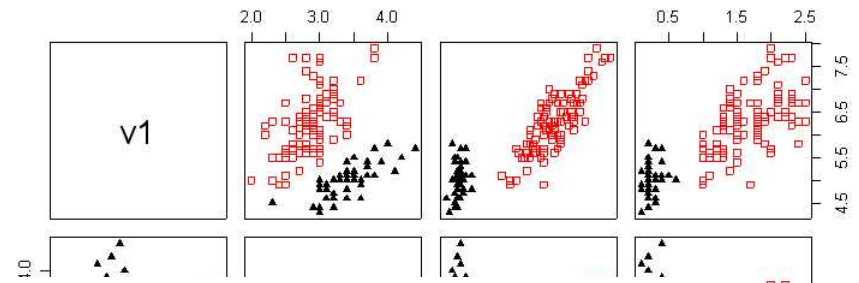
Mixture Modeling

- On considère que les observations sont représentables par une juxtaposition de K gaussiennes multivariées (de centres, d'orientation et de tailles différentes).
- Chaque gaussienne est définie par une moyenne (un vecteur) et une matrice de covariance.
- Pour un nombre de classes donné, on optimise la moyenne de chaque gaussienne et sa matrice de covariance, pour minimiser la distance aux données.
- On répète cette procédure pour plusieurs nombres de gaussiennes.

Exemple synthétique



K-means



Résultats potentiellement différents entre les méthodes!

Régimes de temps

- Un régime de temps est un état récurrent de la circulation atmosphérique à grande échelle.
- Application sur la hauteur du géopotential à 500 mb, sur les données quotidiennes d'hiver (DJF).
 - Circulation atmosphérique parallèle aux lignes d'iso-z500
 - Z500 moins bruité que SLP (effet orographique moins fort).
- Décomposition EOF/PC sur les premières 7 PCs (80% de la variance).

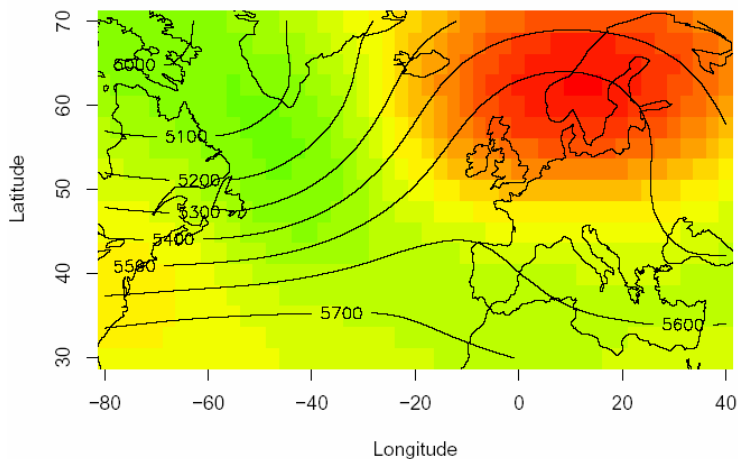
Limitations

- Choix « politiquement correct » de quatre régimes (Corti et al. Nature, 1999; Michelangeli et al. JAS, 1995; Kimoto and Ghil, JAS, 1993...).
- Critère de stabilité de 5 jours consécutifs passés dans un régime donné.
- Pas de classification en été...
- Les ré-analyses NCEP ne sont pas des *observations*.

Régimes de temps NCEP

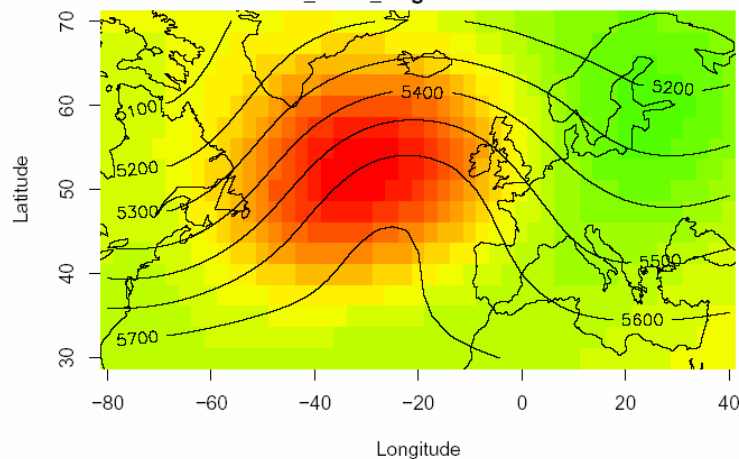
BLOCAGE

NCEP_Z500_regime kmeans 1



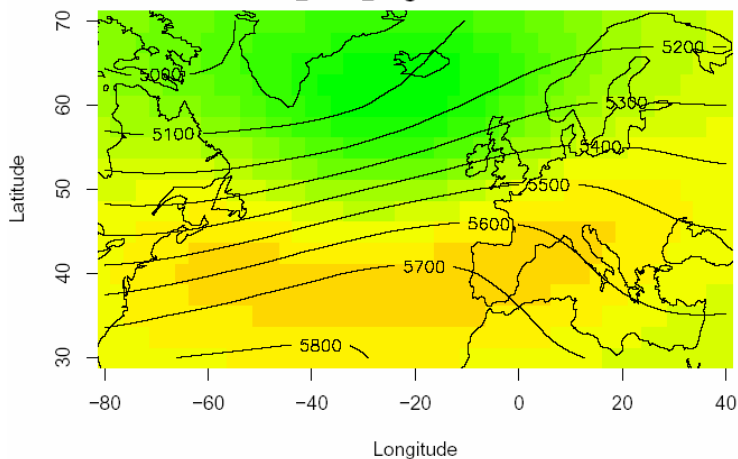
NA RIDGE

NCEP_Z500_regime kmeans 3



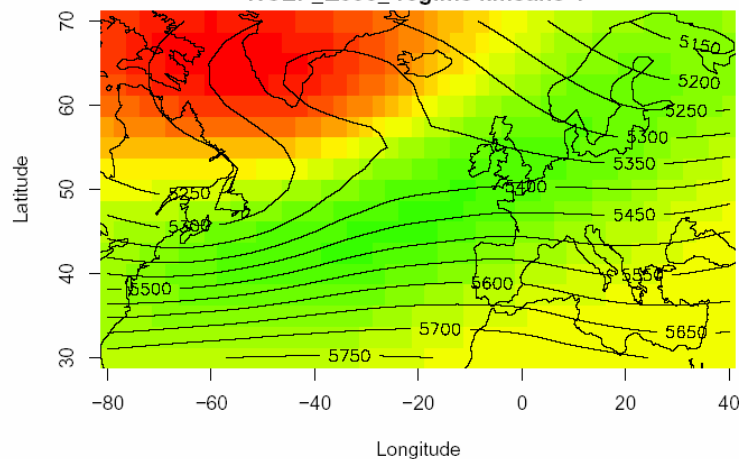
NAO +

NCEP_Z500_regime kmeans 2

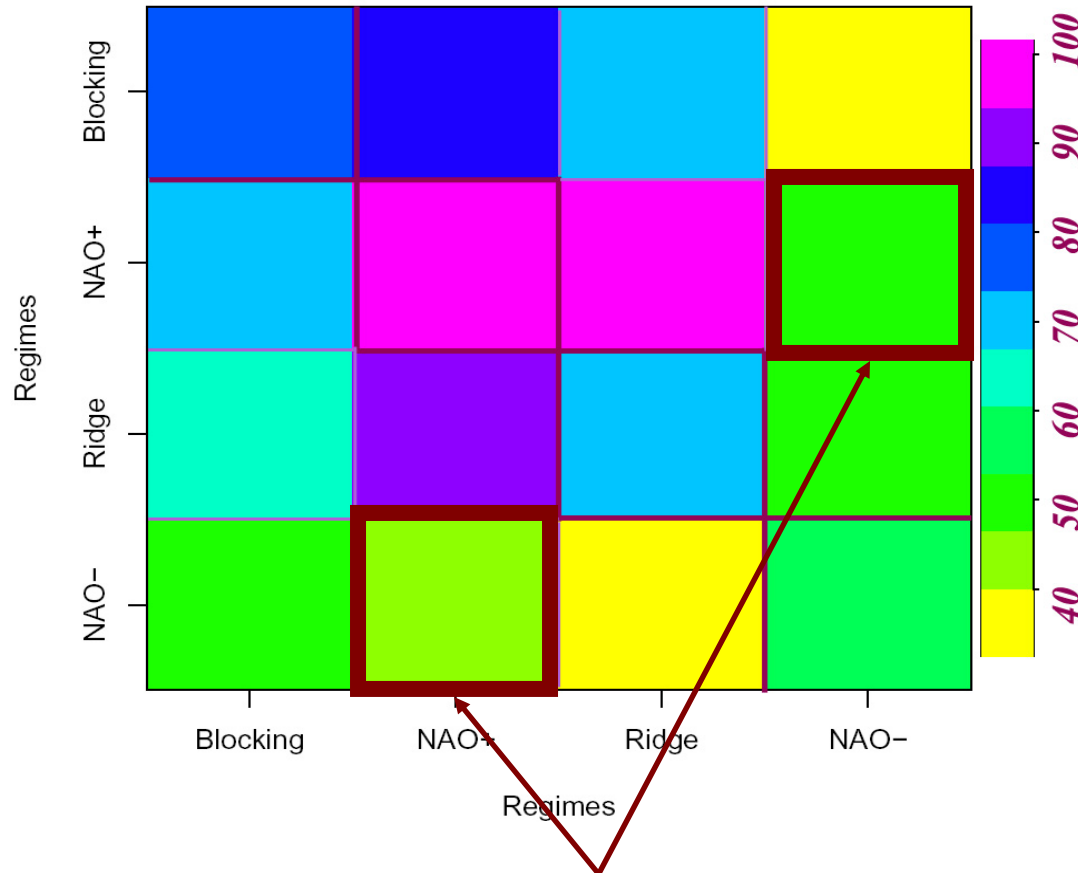


NAO -

NCEP_Z500_regime kmeans 4



Persistance et Transitions de Régime



On passe très rarement des régimes NAO+ à NAO- de manière directe!

Echelles de temps journalières vs. mensuelles

- Connections entre régimes de temps instantanés et indice NAO?
 - Relier *météorologie* et *climatologie*
- Comptage du nombre de jours passés dans chaque régime!

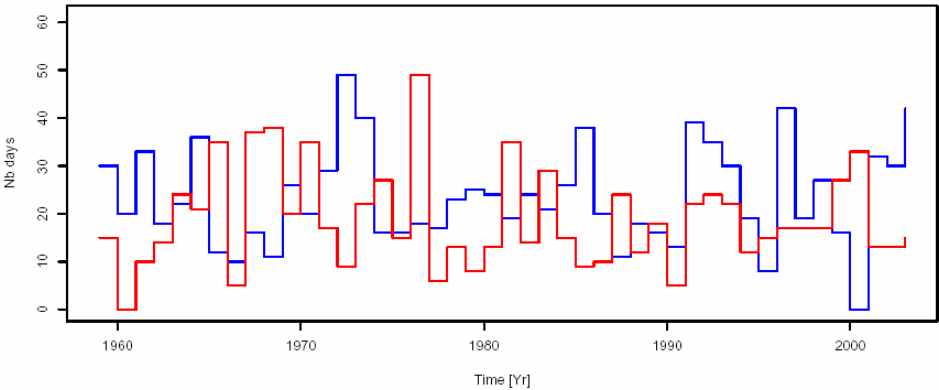
Blocking and Atlantic ridge

NAO+ and NAO-

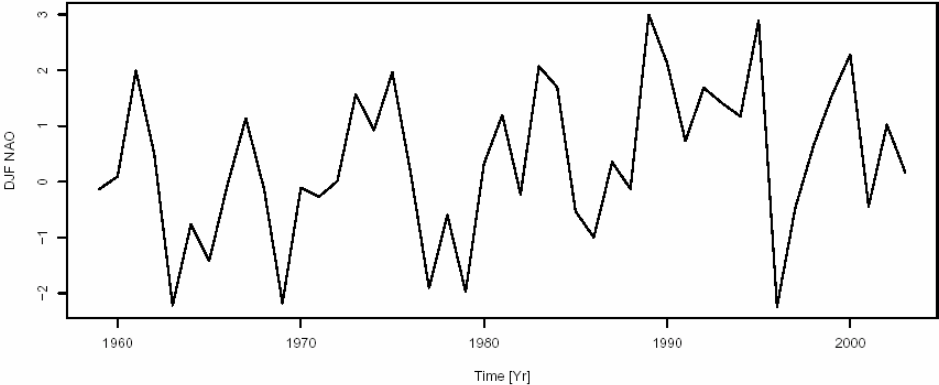
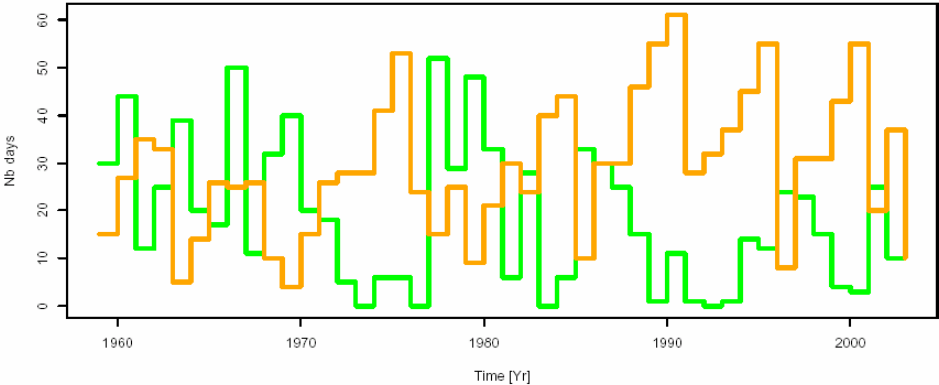
$r = 0.8$

$r = -0.7$

CRU NAO index
(winter average)



Reg. 3 & 4



Morale

- Nous avons projeté un champ quotidien hivernal de géopotential à 500mb sur 7 EOFs représentant 80% de la variance
- Nous avons décomposé ces 7 EOFs en 4 régimes de temps, dont les deux phases de la NAO
- Le nombre de jours passés dans chaque phase correspond à l'indice mensuel de la NAO
- On a besoin des deux autres régimes de temps (blocage et dorsale nord Atlantique) pour passer d'une phase NAO à l'autre.

La variance explique-t-elle tout?

- Toutes les méthodes et concepts exposés jusqu'à présent sont basés sur une décomposition de la variance (des cycles et des régimes).
- Que se passe-t-il quand le système considéré n'est sensible qu'à de grandes variations d'un paramètre?
- L'analyse de la variance se concentre sur les valeurs les plus communes. L'analyse des valeurs rares lui échappe!

Références

- P.A. Michelangeli, R. Vautard, B. Legras, Weather regimes: Recurrence and quasi-stationarity, *J. Atmos. Sci.*, 52: 1237—1256, 1995
- J.P. Peixoto et A.H. Oort, *Physics of Climate*, American Institute of Physics, New York, 1992
- H. von Storch et F. W. Zwiers, *Statistical Analysis in Climate Research*, Cambridge University Press, Cambridge, 1999
- D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, 1995
- <http://www.ipsl.jussieu.fr/CLIMSTAT/>
- <http://www.atmos.ucla.edu/tcd/ssa>
- <http://www.r-project.org>