Ecole d'été ENVI-STAT
Modélisation Statistique et Environnement

28 avril - 10 mai 2008 à Dakar
Ecole Supérieure Polytechnique ESP-UCAD

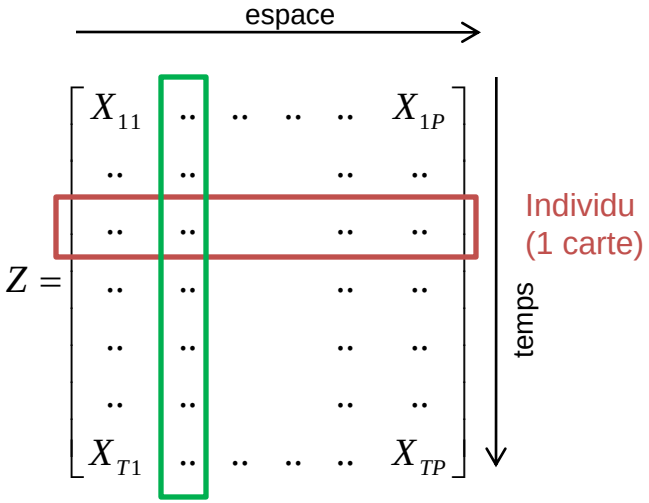
Application des méthodes de décomposition en composantes principales à la climatologie

Etude d'un exemple: les modes de variabilité de la température de surface de la mer dans l'océan Atlantique tropical et la variabilité atmosphérique associée.

Juliette Mignot – LOCEAN / IRD
Abdou Kane – LSCE/CEA
Abdou Gueye – LTI – LOCEAN / IRD

Définition de la matrice des données

On analyse un signal discrétisé Z_i^j où Z (déviation de moyenne temporelle nulle par rapport à $\overline{Z}$) est connu aux points Xj (j∈[1,P]) et aux temps ti (i ∈[1,T]). Typiquement, Z est une succession (série temporelle) de cartes de température à la surface de la mer (SST).



$$Z = \begin{bmatrix} X_{11} & .. & .. & .. & .. & X_{1P} \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ .. & .. & & & .. & .. \\ X_{T1} & .. & .. & .. & .. & X_{TP} \end{bmatrix}$$

La base canonique de R^P sur laquelle sont représentées les variables initialement est l'ensemble des vecteurs de type $e_i = [0 \quad 1 \quad \ldots \quad 0 \quad 0]$
C'est-à-dire que chaque vecteur correspond à 1 point d'espace pris individuellement.

Idée générale

L'idée de l'ACP (ou ici décomposition en EOF) est de déterminer une nouvelle base de R^P qui représente le mieux les liens (les covariances) entre les différents individus. On peut montrer que cette nouvelle base est obtenue par diagonalisation de la matrice de covariance.

Présentation des données

- Matrice Z: Données de température à la surface de la mer (SST) issues du jeu de réanalyses NCEP/NCAR:
http://www.cdc.noaa.gov/cdc/data.ncep.reanalysis.derived.surfaceflux.html

-Résolution temporelle:
Données mensuelles couvrant la période 1948/01 – 2007/12 (60 années)
T=12x60=720 par défaut. Dans la suite, on travaillera essentiellement
• soit avec les moyennes mensuelles saisonnières, donc T=12 (moyenne sur les 60 années),
• soit avec les moyennes annuelles, donc T=60 (moyenne sur les 12 mois de chaque année)

- Résolution spatiale:
Le jeu de données initial couvre la Terre entière: [88.542N - 88.542S, 0E - 358.125E] soit 192x94 points de grille spatiaux (lonxlat)

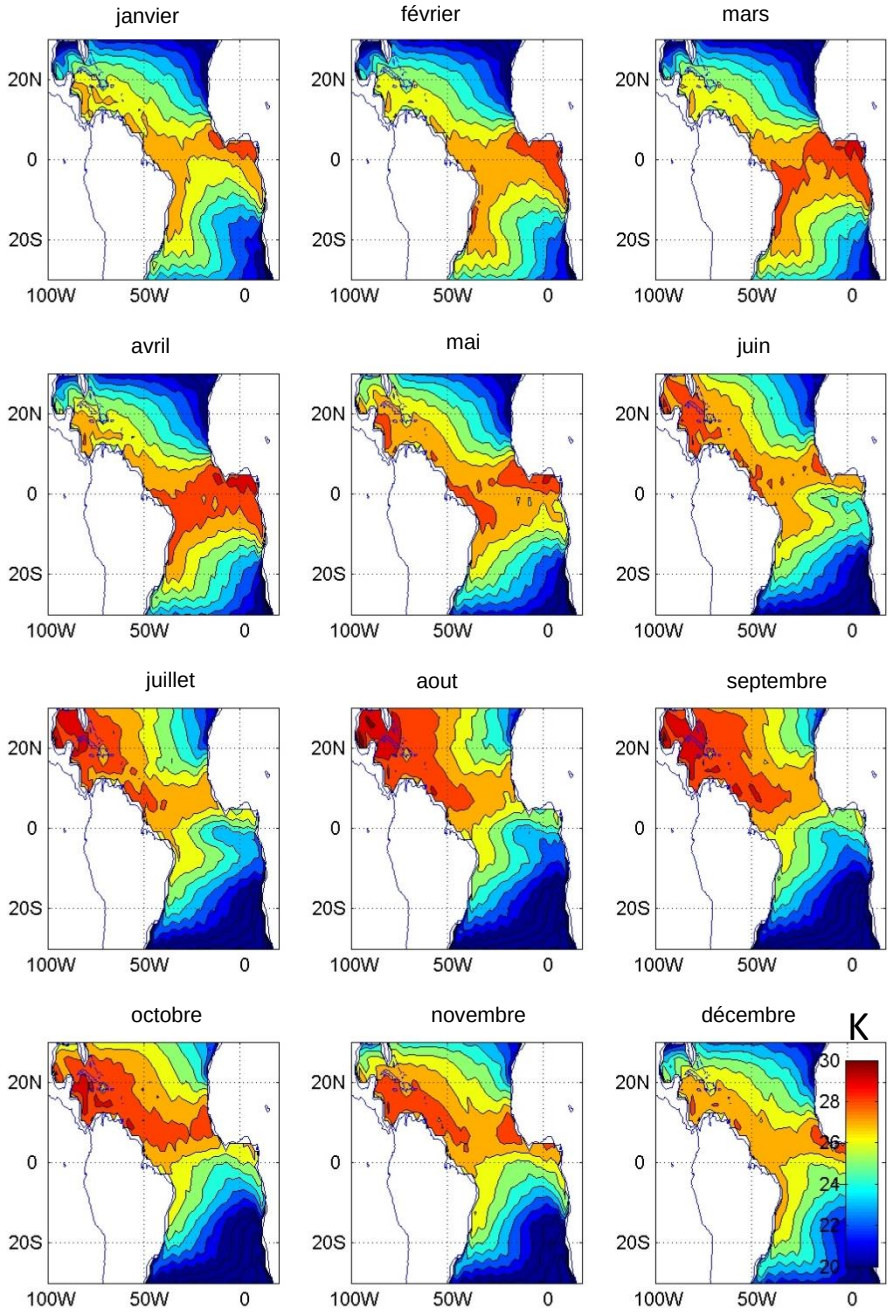
On travaille ici avec un sous ensemble couvrant le domaine spatial [40S-40N, 100W-30E] (Atlantique tropical), soit 70x42 points de grille.
Donc P=70x42=2940

- On dispose également des flux atmosphériques associés (chaleur, eau douce, moment)

Exemple d'application

Visualisation de la variabilité saisonnière de la SST dans l'Atlantique tropical: cartes mensuelles moyennes

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

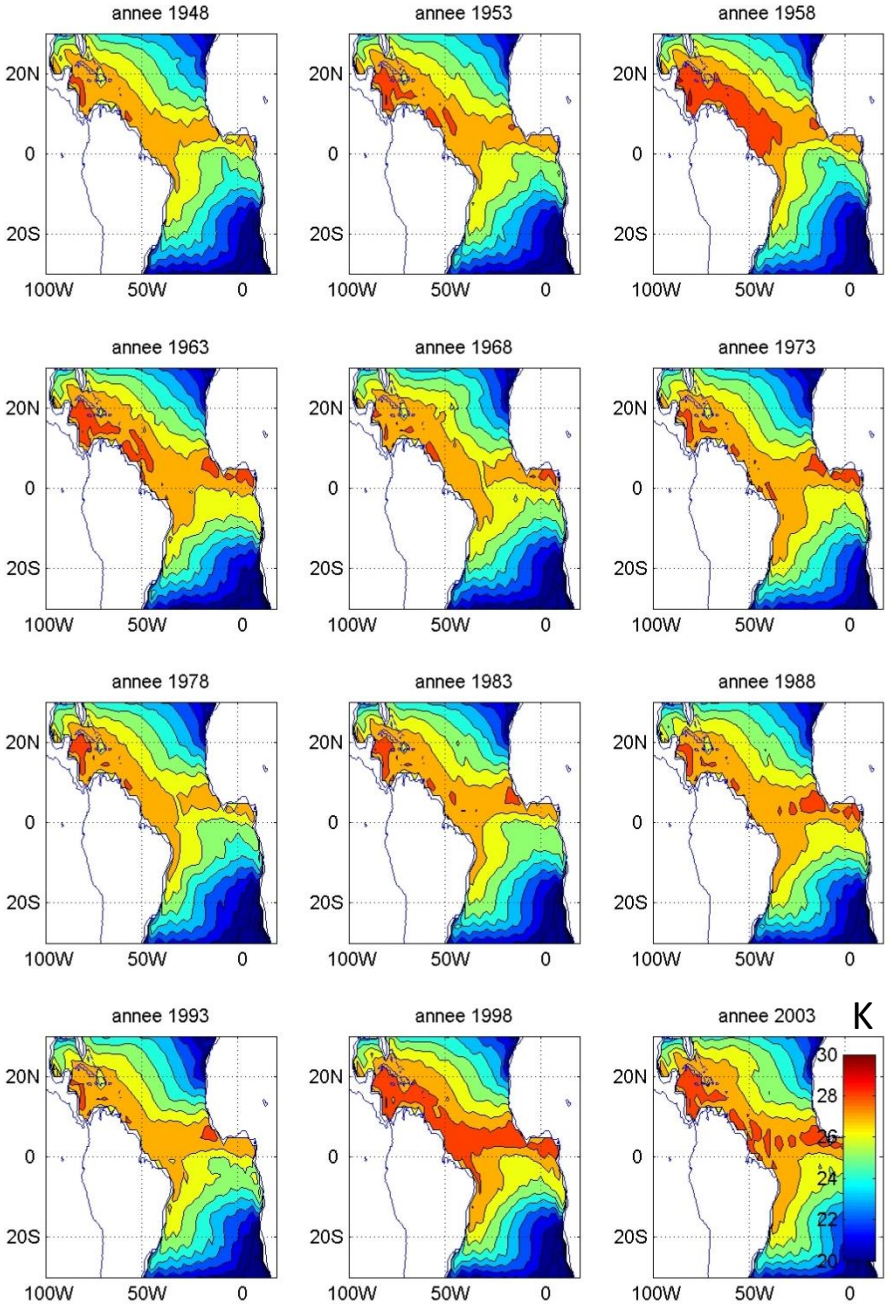


Programme matlab fourni

Exemple d'application

Visualisation de la variabilité interannuelle de la SST dans l'Atlantique tropical: cartes annuelles moyennes

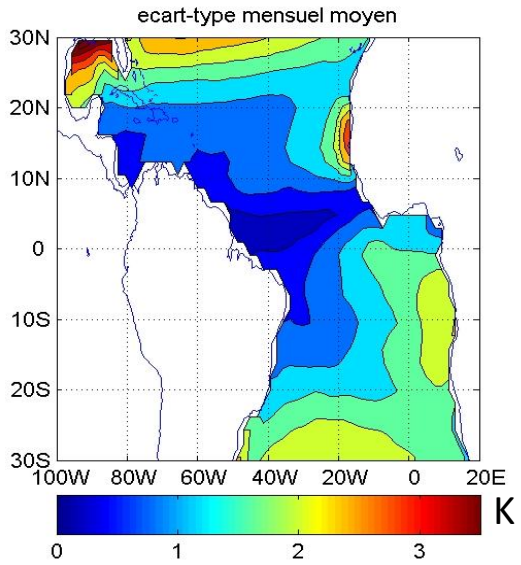
dim(Z)=[T x P]
Ici, T=60 (nombre d'années)

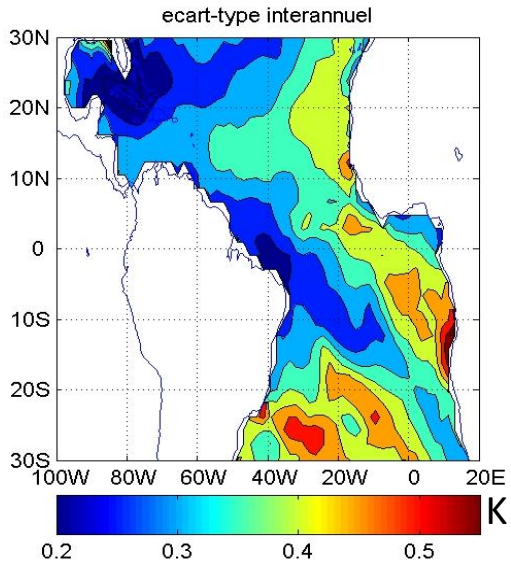


Modifier le programme matlab précédent pour visualiser la variabilité interannuelle de la SST

Exemple d'application

Visualisation de la variabilité de la SST dans l'Atlantique tropical: écart-type





dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

dim(Z)=[T x P]
Ici, T=60 (nombre d'années)

Programmer le calcul de l'écart-type mensuel moyen et interannuel à l'aide du programme matlab précédent et de la fonction « std » (voir >> help std)

L'écart-type quantifie la variabilité en chaque point d'espace pris séparément. L'objectif de la décomposition en EOF est d'obtenir des informations sur la co-variabilité des différents points d'espace entre eux.

Matrice de covariance

On définit la matrice de covariance (ou matrice d'inertie) par $C_{il} = \frac{1}{P}\sum_i Z_i^j Z_i^l$

Ou encore $C = {}^tZZ$. *C* est donc une matrice de dimensions *PxP*

La trace de la matrice de covariance est aussi appelée moment d'inertie totale. C'est une mesure de la dispersion du nuage des individus (mesures de SST) par rapport à son centre de gravité. La recherche de directions u_k maximisant l'inertie du nuage (donc tels que $Max({}^tu_k{}^tZZu_k)$) se traduit en algèbre linéaire par une recherche de vecteurs propres de la matrice de covariance C.

Diagonalisation de la matrice de covariance

C est une matrice symétrique de taille PxP définie positive, dont les valeurs propres λ_i sont réelles et positives.
Les vecteurs propres associés constituent une nouvelle base de R^P qui maximise la représentation de la variance/covariance entre les individus. Ils sont appelés les Fonctions Empiriques Orthogonales (EOF) pour *Z*.

Composantes principales

On définit les composantes principales de Z comme la décomposition (normalisée) du signal décrit par Z sur les EOF (vecteurs propres) En:

$$P_n^j = \sum_i Z_i^j E_{n,i} \qquad P_n = {}^{P_n}\!/_{std(P_n)}$$

Les composantes principales sont des vecteurs R^T . Ce sont de « nouvelles variables » dont les composantes donnent les coordonnées du nuage de point (des individus) sur les nouvelles directions principales (vecteurs propres, EOF)

Interprétations

La composante principale Pn est le coefficient de l'EOF En et dépend du temps (indice j) mais pas de l'espace. En est une carte spatiale indépendante du temps.

On a ainsi obtenu une décomposition de Z sous la forme $Z = \sum_n P_n E_n$
Où les EOF En sont orthogonales entre elles et les composantes principales Pn sont orthogonales entre elles.

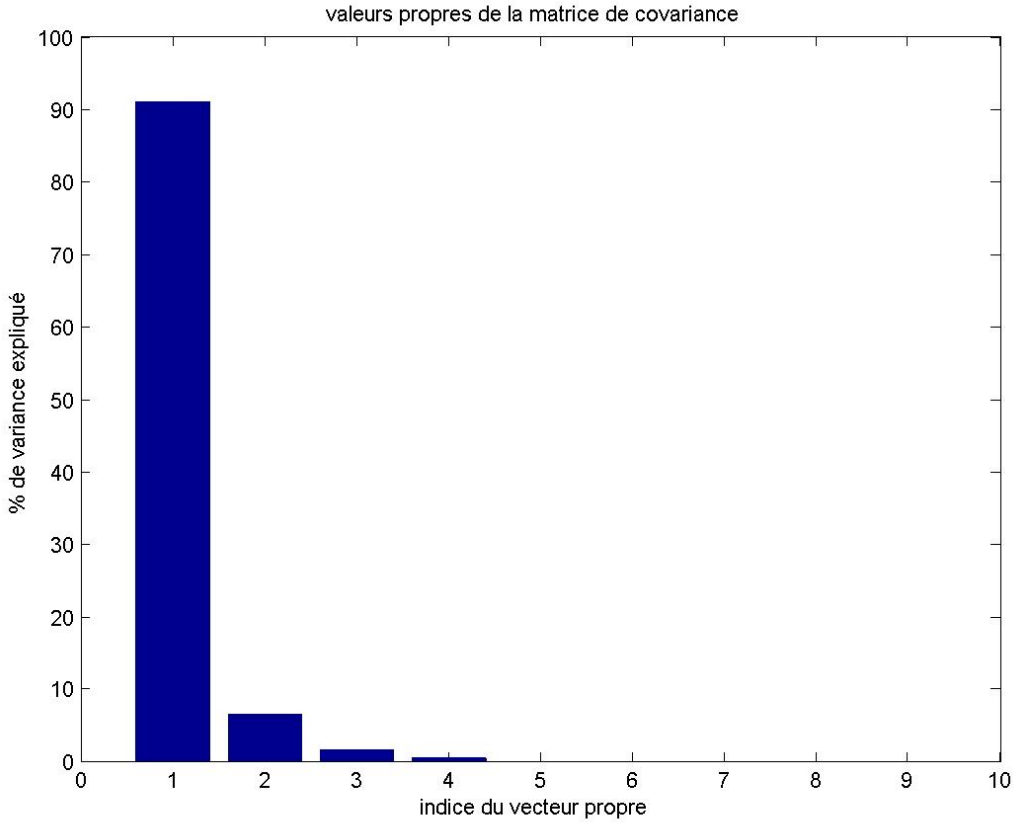
On peut par ailleurs montrer que l'on peut considérer $\lambda_m / \sum_n \lambda_n$ comme la part de la variance associée à l'EOF m.

Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

Programme matlab fourni



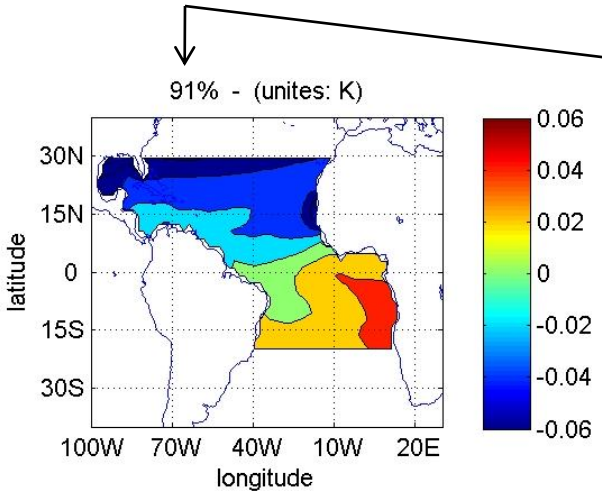
Répartition des valeurs propres: le 1er vecteur propre explique 91% de la variance saisonnière.
Les suivants expliquent une proportion largement plus faible, 6% pour le second et moins de 2% pour les suivants. Au moins dans un premier temps, on limite donc l'analyse au 1er vecteur propre.

Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

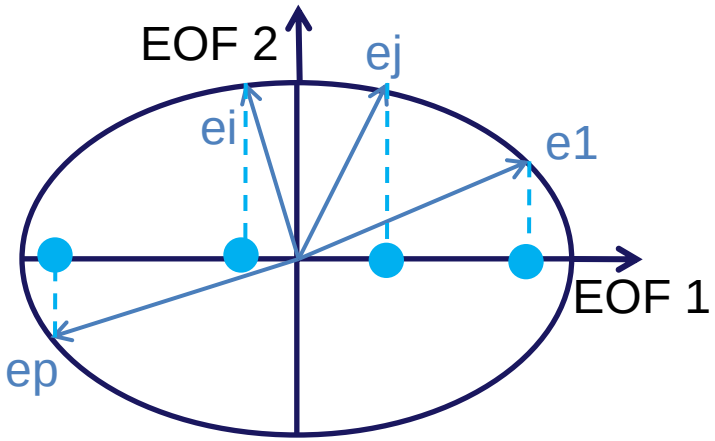
Programme matlab fourni



Pourcentage de variance expliquée

La **carte** illustre le poids (codé en couleur) de chacun des vecteurs de la base initiale (points d'espace) sur ce vecteur propre EOF1

Les données étant géographiques et nécessitant donc déjà 2 dimensions pour la visualisation, on ne représente qu'un mode à la fois.

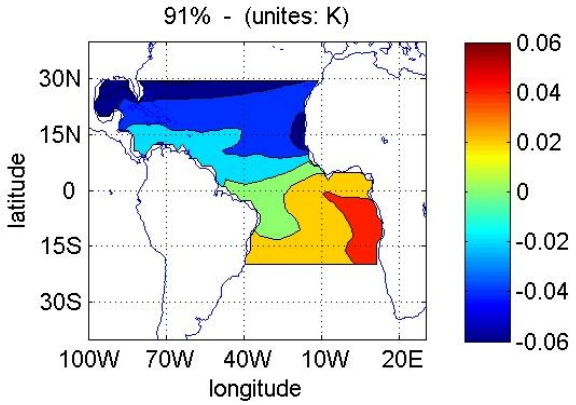


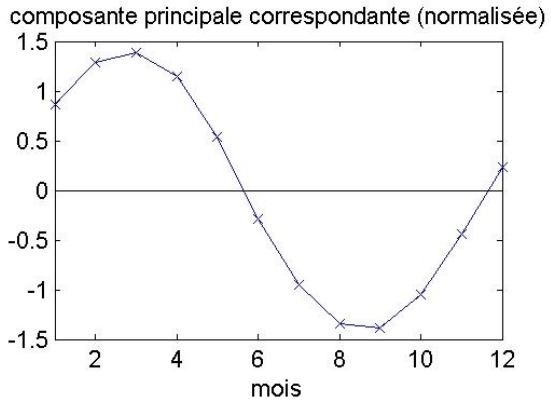
Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

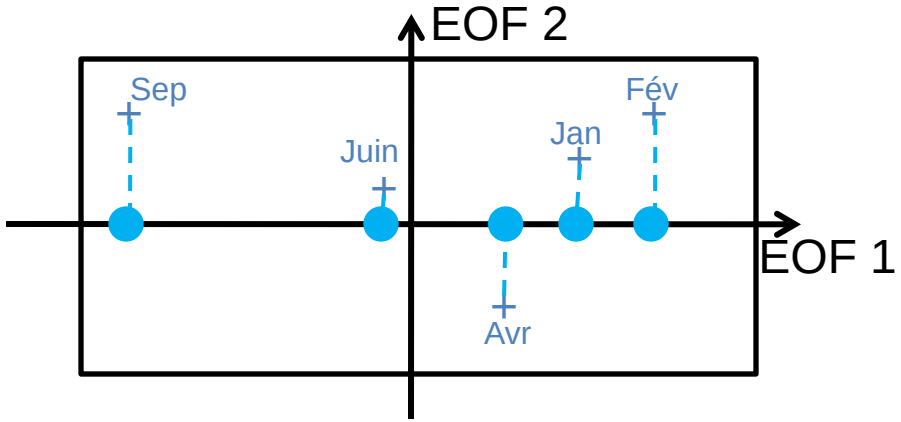
dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

Programme matlab fourni





La **série temporelle** correspond à la projection des individus de la matrice Z initiale sur le vecteur propre EOF 1

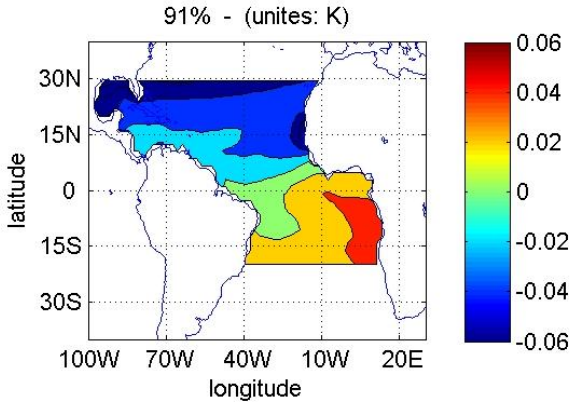


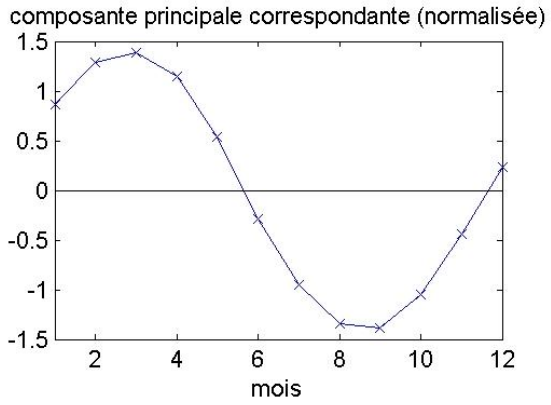
Exemple d'application (1)

Décomposition des moyennes **mensuelles saisonnières** en EOF:

dim(Z)=[T x P]
Ici, T=12 (nombre de mois de l'année)

Programme matlab fourni





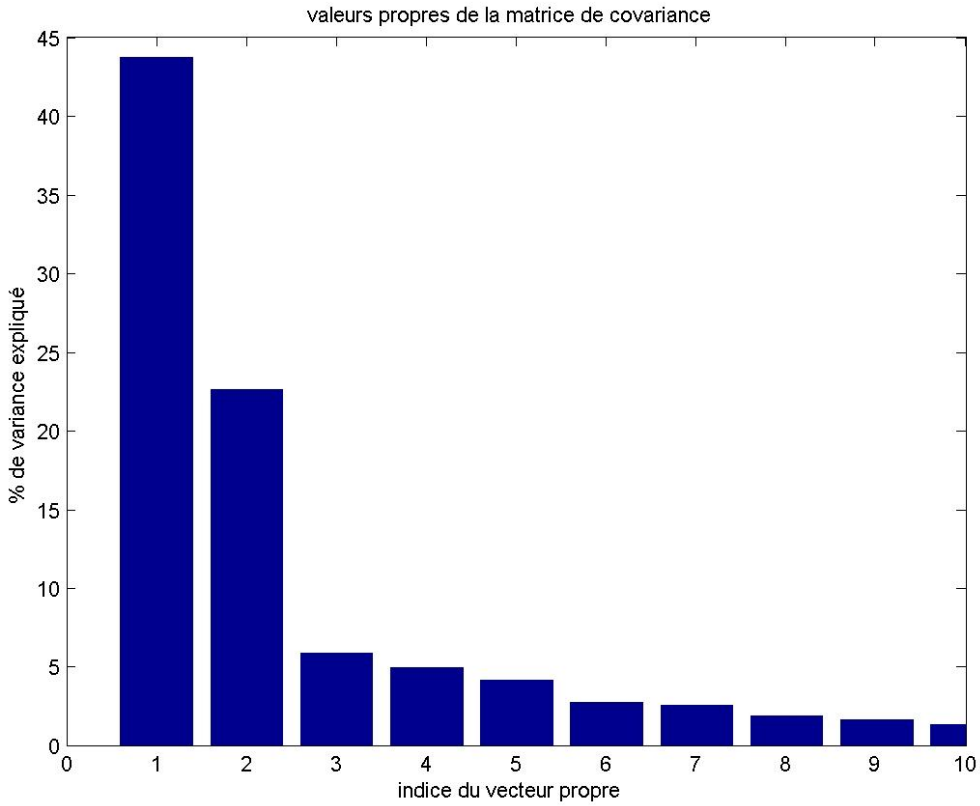
91% de la variabilité saisonnière moyenne de la SST est expliquée par un refroidissement relatif dans l'hémisphère nord associé à un réchauffement au sud
La série temporelle illustre la modulation saisonnière de ce mode (sur les 12 mois de l'année): ce mode est en phase positive durant l'hiver boréal et négative (il fait + chaud au nord et + froid au sud) durant l'hiver austral

Exemple d'application (2)

Décomposition des moyennes **annuelles** en EOF:

dim(Z)=[T x P]
Ici, T=60 (nombre d'années)

Modifier le programme matlab précédent pour calculer les modes de variabilité interannuelle de la SST dans l'Atlantique tropical



Répartition des valeurs propres: les valeurs propres sont mieux réparties qu'en saisonnier. On voit néanmoins que hors mis les 2 premières valeurs propres, les autres sont assez faibles et proches. Les 2 premiers vecteurs propres expliquent à eux deux les 2/3 de la variance totale de la SST interannuelle sur le bassin.

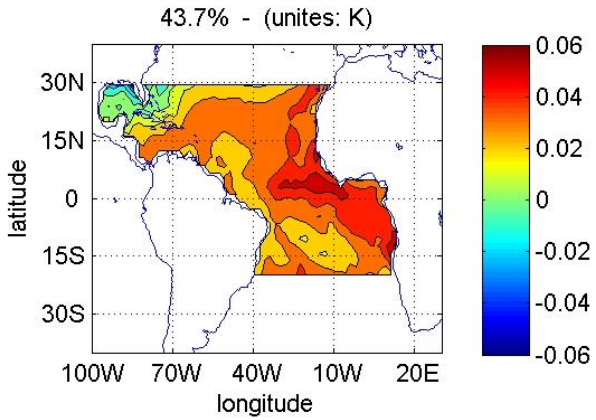
Exemple d'application (2)

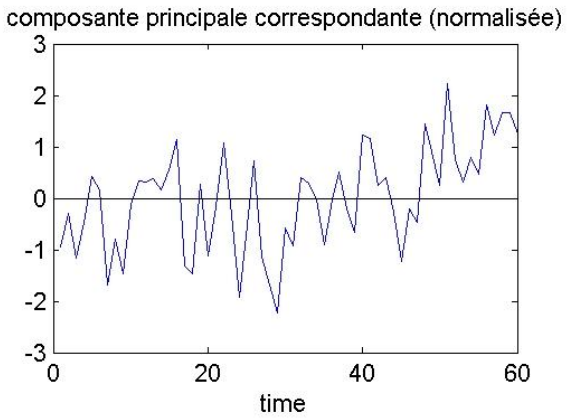
Décomposition des moyennes **annuelles** en EOF:

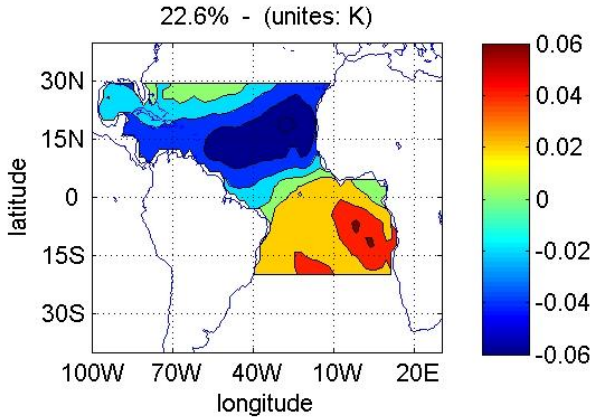
dim(Z)=[T x P]
Ici, T=60 (nombre d'années)

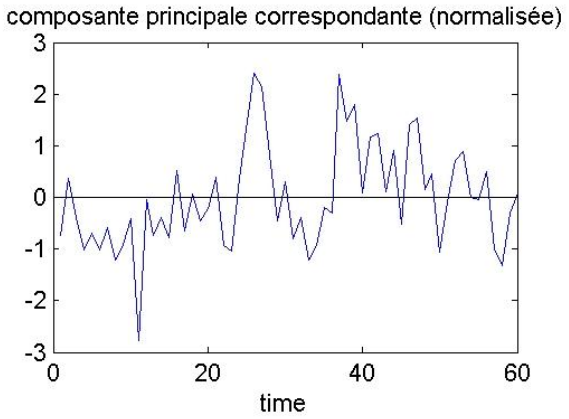
Modifier le programme matlab précédent pour calculer les modes de variabilité interannuelle de la SST dans l'Atlantique tropical

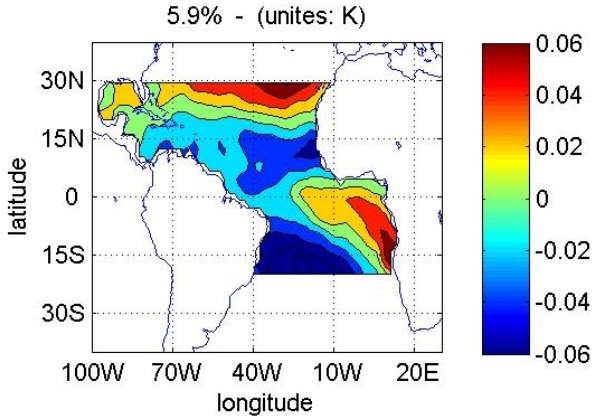
On ne montre ici que les 3 premiers vecteurs propres.

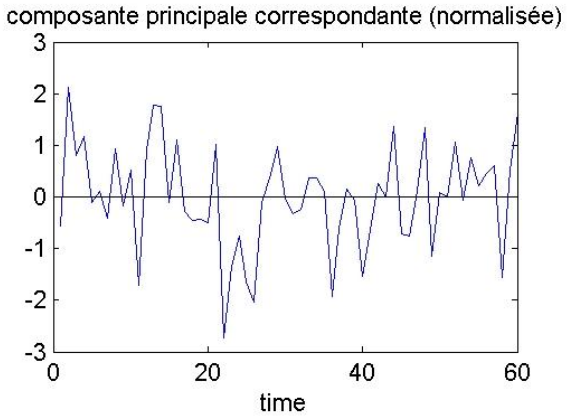












Remarques:

- Il faut se méfier d'accorder une signification physique particulière à une EOF indépendemment des autres. En effet, la décomposition présentée ci-dessus n'est pas intrinsèque mais dépend des choix arbitraires des produits scalaires (distance). De plus, la condition d'orthogonalité des EOFs contraint la structure des EOFs au-delà de la 2ème, limitant ainsi l'interprétation physique.

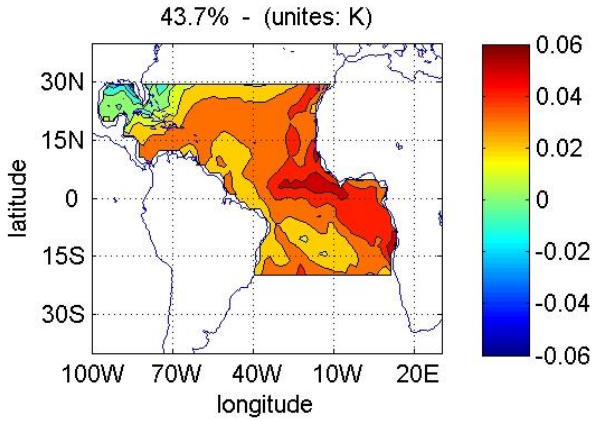
-Le principal intérêt des EOF apparaît lorsque les valeurs propres m décroissent rapidement avec m, la variance étant concentrée dans un petit nombre de modes. Dans ce cas, le signal se retrouve réduit à un petit nombre de degrés de libertés, ce qui facilite grandement l'analyse ultérieure. Souvent, la réduction en EOF est une étape préliminaire à une analyse statistique plus élaborée.

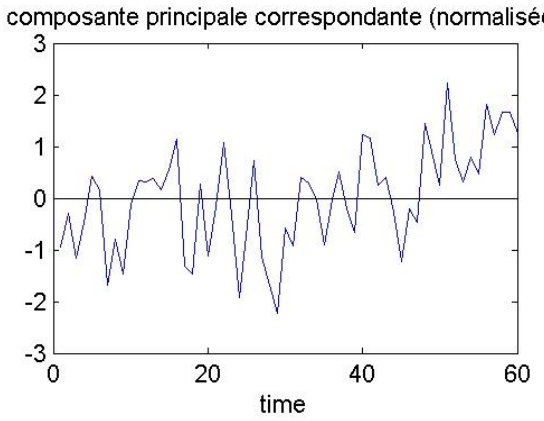
– La décomposition en EOF est bien adaptée à l'´etude des propriétés stationnaires de la variabilité. Plusieurs généralisations existent pour l'étude de modes propagatifs.

Exemple d'application (2)

Interprétation de la 1ère EOF

Maximum de variance expliquée situé à l'équateur et au sud le long de la côte africaine. Même signe d'anomalie dans tout le bassin. La série temporelle est caractérisée par une variabilité interanuelle relativement basse fréquence (oscillation bi-annuelle) superposée à une dérive positive.

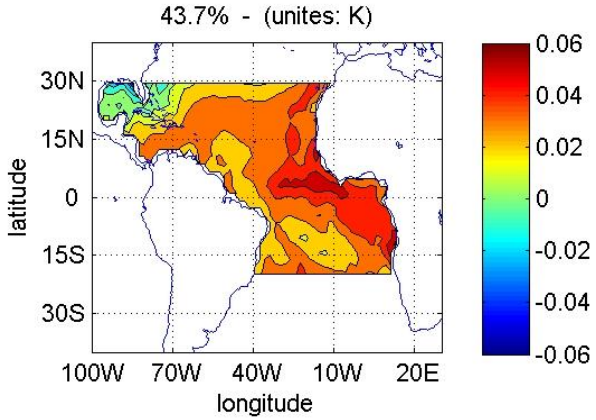


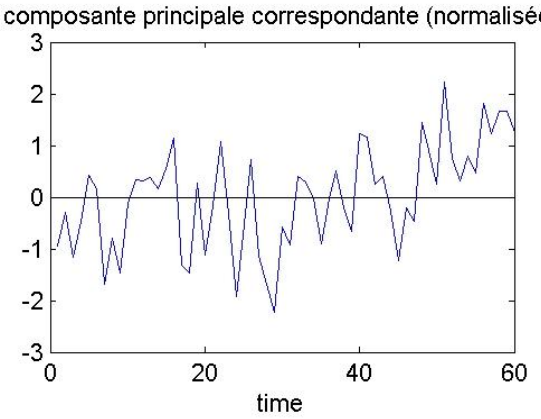


Calculer la régression linéaire de la série temporelle correspondant aux point d'espace [2N-15W] et [20N-95W] respectivement avec la composante principale. Commenter en lien avec la structure de ce vecteur propre.

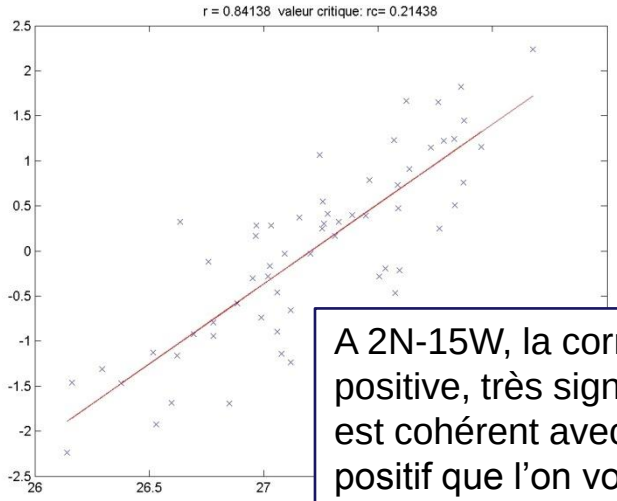
Exemple d'application (2)

Interprétation de la 1ère EOF

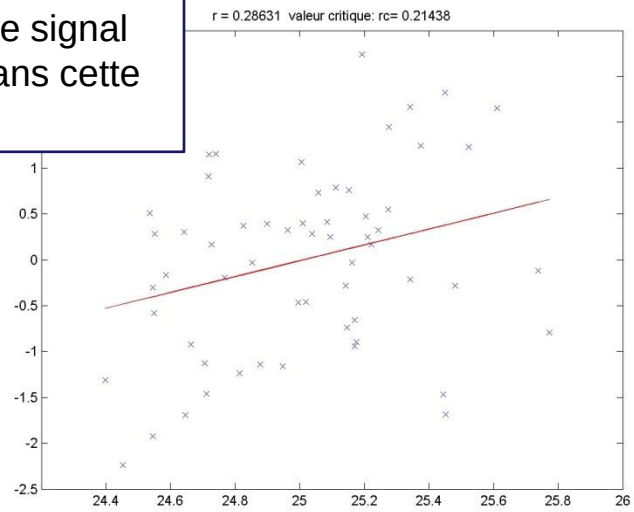




A 20N-95W, la corrélation est faiblement positive, à la limite de la significativité. Ceci est cohérent avec le faible signal associé à ce mode dans cette région.



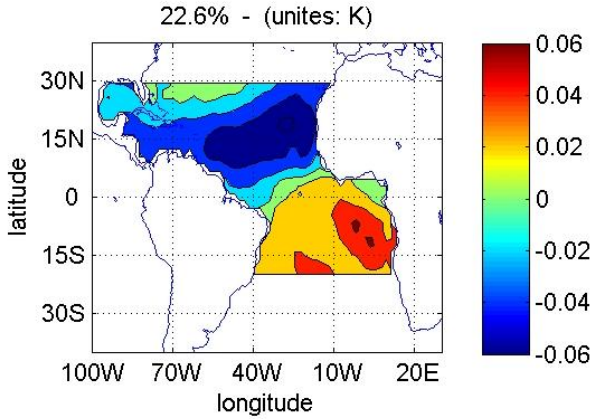
A 2N-15W, la corrélation est positive, très significative. Ceci est cohérent avec le fort signal positif que l'on voit sur la carte à cet endroit.

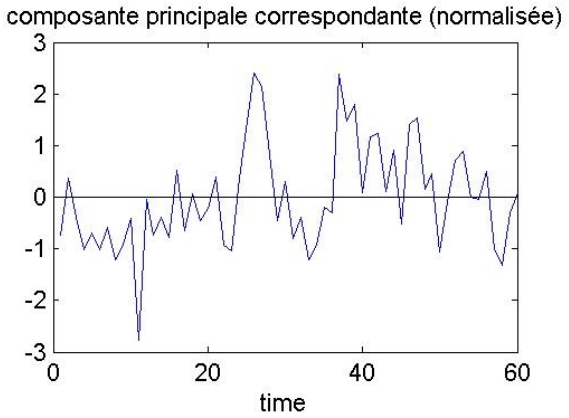


Interprétation de la 2ème EOF

maximum de variance dans le bassin tropical Nord et dans le bassin Sud avec des signes opposés.
baptisé mode « méridien » ou « dipôle » (Ruiz-Barradas et al., 2000).
Remarque: sa réalité physique est discutable (Houghton et Tourre 1992, Nobre et Shukla 1996, Enfield et al. 1998).

Série temporelle relativement basse fréquence

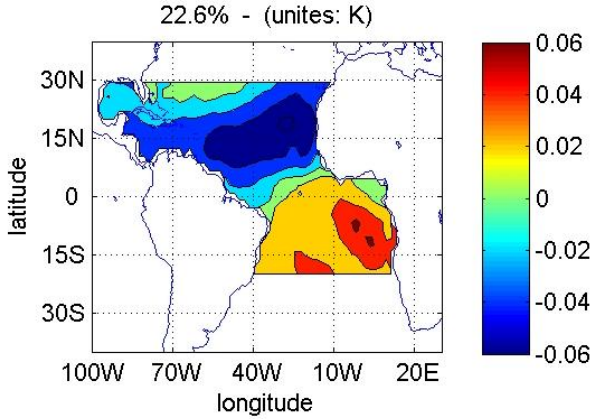


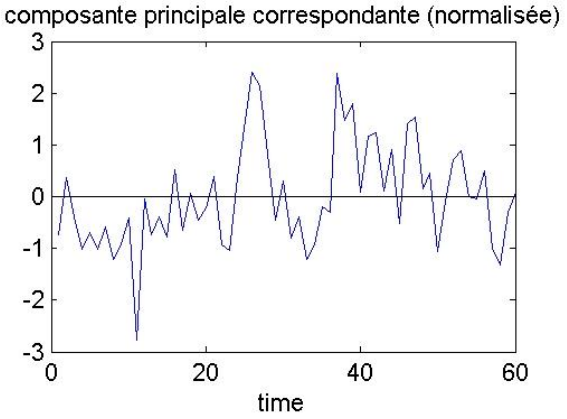


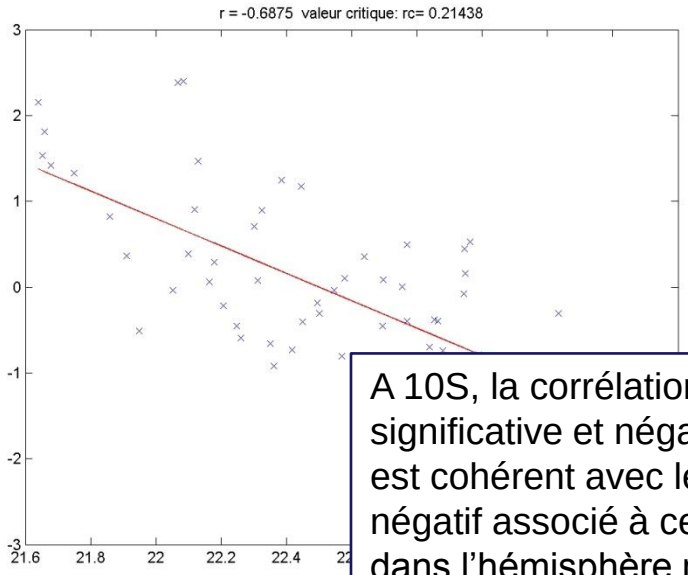
Calculer la régression linéaire des séries temporelles correspondant aux point d'espace [20N-20W] et [10S-5E] respectivement avec la composante principale. Commenter en lien avec la structure de ce vecteur propre.

Exemple d'application (2)

Interprétation de la 2ème EOF







A 10S, la corrélation est significative et négative. Ceci est cohérent avec le fort signal négatif associé à ce mode dans l'hémisphère nord

A 20N, la corrélation est significative et positive. Ceci est cohérent avec le fort signal positif associé à ce mode dans l'hémisphère nord