Projet industriel : Méthode Optimal Fingerprint et Extensions

Projet industriel MAIN4 - Rapport

Référent : H. Durand

Le 19 novembre 2024

Table des Matières

1 Introduction

1.1 Thématique

1.2 Problématique

Peut-on détecter qu'il existe un changement climatique ? Est-ce dû à l'Homme ?

Sous-problématique : Compréhension et amélioration de la méthode Optimal Fingerprint

- Qu'est-ce que la méthode Optimal Fingerprint (OF), et comment s'applique-t-elle aux données climatiques ?
- Quels sont les résultats obtenus par cette méthode dans le passé, et quelles en sont les limites (notamment pour les précipitations, la détection à l'échelle locale/régionale) ?
- Comment peut-on améliorer cette approche en utilisant des techniques récentes comme le machine learning ou l'inférence causale ?
- Quels nouveaux cas d'usage peuvent être imaginés pour ces extensions (détection régionale, "time of emergence", etc.) ?

1.3 Définitions

2 Théorie de la méthode Optimal Fingerprint

2.1 Préliminaire statistique

1. **Test statistique** : Un test statistique permet de prendre une décision entre deux hypothèses. À partir de l'hypothèse statistique et d'un échantillon de données, on doit répondre à une certaine problématique.
2. **Hypothèse nulle** H_0 : C'est l'hypothèse qu'on considère par défaut et dont on cherche à vérifier la vraisemblance.
3. **Hypothèse alternative** H_1 : Elle traduit une incompatibilité avec H_0.
4. **Erreur de type 1** : C'est la probabilité de rejeter à tort H_0.
5. **Erreur de type 2** : C'est la probabilité de conserver à tort H_0.
6. **Puissance d'un test** : C'est la probabilité de rejeter H_0 quand H_1 est vraie. C'est donc l'aptitude d'un test à rejeter une hypothèse fausse.

2.2 L'article de Zwiers

Ici, nous allons expliquer les notions utilisées dans l'article : *The detection of climate change.*

- **Idée principale** : On observe une quantité $T(x,t)$ avec x la localisation et t le temps observé. Cette quantité est appelée *signal climatique*, donnée par l'équation suivante :

$$T(x,t) = S + N$$

avec :

  - S : Un signal déterministe. Il représente la réponse du climat à un forçage externe spécifique (gaz à effet de serre, activité volcanique, etc.).
  - N : Un bruit aléatoire. Il est modélisé comme un processus stochastique de moyenne nulle. Il représente la variabilité naturelle du climat.

- **Le problème de la détection du changement climatique consiste à :**
  1. Identifier la présence du signal S dans les observations T.

2. Attribuer tout ou partie du signal détecté aux activités humaines.

— Pour répondre au premier problème (1), les chercheurs proposent un filtre spatio-temporel pour séparer le signal S du bruit N. Par exemple, Bell (1982) propose une variable de détection donnée par une moyenne pondérée spatio-temporelle.

$$A_t = \sum_{x=1}^{l} \sum_{\tau=1}^{m} w(x,\tau) T(x, t-\tau+1) \tag{1}$$

$$\text{avec} \quad \sum_{x=1}^{l} \sum_{\tau=1}^{m} w(x,\tau) = 1. \tag{2}$$

Le filtre défini par l'équation ci-dessus peut être exprimé en notation matricielle-vecteur comme suit :

$$A_t = \mathbf{w}^T \mathbf{T}_t$$

— Le filtre optimal est celui qui maximise le rapport signal / bruit défini par :

$$\gamma^2 = \frac{\mathbb{E}(A_t)^2}{\mathrm{Var}(A_t)} \quad \text{ou} \quad \gamma^2 = \frac{(\mathbf{w}^T \mathbf{S}_t)^2}{\mathbf{w}^T \Sigma_{N_t} \mathbf{w}},$$

où :

— $\mathbf{w}^T$: la transposée du vecteur des poids $\mathbf{w}$,

— $\mathbf{S}_t$: le vecteur du signal au temps t,

— Σ_{N_t} : la matrice de covariance du bruit N.

— **Hypothèses de la méthode** : La méthode vient originellement de Klaus Hasselmann. Il suppose que $T(x,t)$ i.i.d, c'est-à-dire des variables indépendantes et identiquement distribuées, et que N est gaussien.

— Pour déterminer si le signal S existe, on fait un test statistique avec l'hypothèse nulle $H_0 : T = N$ (il n'y a pas de signal) contre $H_1 : T = S + N$ (le signal existe). On calcule la statistique du test Z_t. On rejette l'hypothèse H_0 et on conclut qu'il y a un signal si Z_t est supérieur à un seuil critique.
Dans l'article, il fixe le seuil à $1,96 \sim 2$ parce qu'il fait un test à 5% d'erreur sur une gaussienne centrée.

$$Z_t = \frac{\mathbf{S}_t^T \Sigma_{N_t}^{-1} \mathbf{T}_t}{\sqrt{\mathbf{S}_t^T \Sigma_{N_t}^{-1} \mathbf{S}_t}} = \frac{\mathbf{S}_t^T \Sigma_{N_t}^{-1} \mathbf{T}_t}{\gamma} \geq 2$$

2.3 L'article d'Allen

Allen a décrit la théorie de l'OF dans son article *"Estimating signal amplitudes in optimal fingerprinting."* Il présente l'optimal fingerprint comme un modèle de régression linéaire. L'équation est la suivante :

$$y = \sum_{i=1}^{m} x_i \beta_i + v = X\beta + v$$

Où :

— y est le vecteur des observations climatiques.

— X est la matrice des "fingerprints" simulés par un modèle climatique. Chaque colonne de X représente le *pattern* de réponse à un forçage spécifique.

— β est le vecteur des amplitudes des "fingerprints" à estimer.

— v est le vecteur du bruit climatique, représentant la variabilité interne du système climatique non expliquée par les forçages.

On va tester pour chaque i l'hypothèse $H_0 : \beta_i = 0$. C'est-à-dire, on va tester si la variable x_i influe significativement sur les observations y ou pas. Si l'intervalle de confiance pour β_i ne comprend pas zéro, on rejette H_0. Cela signifie que le forçage x_i a probablement contribué au changement climatique observé.

Tous les résultats obtenus permettent donc d'identifier les forçages climatiques qui ont une influence significative sur le climat observé.

Pour conclure, Allen prend en compte l'incertitude d'échantillonnage dans les simulations. Cette approche est nommée *"total least squares" (TLS)*. Tandis que Zwiers développe l'approche de *"ordinary least squares" (OLS)*, qui suppose que le bruit est présent uniquement dans les observations.

Les deux méthodes sont équivalentes dans le but de détecter un changement climatique.

2.4 Démarche

2.4.1 Du théorique au pratique :

Pour des raisons de puissance , nous avons choisi l'approche de Zwiers tout en reprenant des notions abordés par Allen dans le calcul.
En effet, si les hypothèses sont correctes, la puissance du test est optimale avec cette méthode.
On considère T un vecteur aléatoire tel que ses colonnes représentent l'espace et ses lignes représentent le temps.

Nous allons utiliser les formules et notations suivantes dans notre code :

1. **Calcul de** β : Le vecteur des amplitudes à estimer, donné par la formule suivante :

$$\beta = C_N^{-1} X$$

avec :

— X : Le signal (noté S dans l'article),

— C_N : La matrice de covariance des anomalies (matrice de covariance du bruit dans l'article).

2. **Hypothèses du test statistique :**
H_0 : *"aucun changement climatique "* : $Z_t \leq 2$
H_1 : *"changement climatique observé"* : $Z_t > 2$
Nous effectuons le test à un niveau de confiance de 5%.

3. **Calcul de** Z : La statistique de test avec la formule suivante :

$$Z = \frac{Y(t)\beta}{\gamma}$$

Pour cela, nous allons commencer par calculer γ :

$$\gamma^2 = X^T C_N^{-1} X$$

2.4.2 Gestion de la taille des données :

Vu la taille des données sur lesquelles nous avons travaillé, nous avons eu recours à ce qu'on appelle **PCA** (**P**rincipal **C**omponents **A**nalysis).

L'idée principale de cette approche est de réduire la dimensionnalité des données tout en conservant au maximum l'information essentielle, en se basant sur les directions où la variance est la plus élevée.

Nous passons alors de données contenant plusieurs variables corrélées à un jeu de variables non corrélées appelées **composantes principales**.
Cette étape a été cruciale pour l'obtention de résultats valides.

3 Application de la méthode Optimal Fingerprint

3.1 Présentation du premier jeu de données

Notre client nous a fourni un jeu de données contenant 50 fichiers de type NetCDF (.nc). Ces derniers représentent différentes simulations climatologiques générées par des modèles climatiques. (1 seul modele)

Tous les fichiers sont multidimensionnels de dimension (time,lat,lon,bnds). La période couverte va du **16 janvier 1880** au **16 décembre 2022** soit 1716 mois.

Les variables contenues sont :

— **tas** : variable principale représentant les températures en Kelvin à la surface terrestre de dimension *(time,lat,lon)*.

— **time_bnds** : définit les bornes temporelles pour chaque point dans la dimension *time*.

3.2 Présentation du code

3.2.1 Environnement de travail

— Nous avons travaillé tout au long du semestre sur Google Collab de manière interactive.

— Les librairies python auxquelles nous avons eu recours sont principalement : `numpy,matplotlib,pandas, xarray,cftime, cartopy,os,google.colab...`

3.2.2 Implémentation de l'algorithme

a) **Traitement des données :** pour chaque simulation, les étapes suivantes sont appliquées

— Nous avons réduit la granularité spatiale. Les données ont été rééchantillonnés à une résolution spatiale plus faible de l'ordre de mille points. Pour cela, nous avons implémenté une fonction `reduce_granularity` qui utilise `coarsen` de la bibliothèque **xarray**.

— Ensuite, nous avons séparé les données en deux périodes : passées de 1880 à 1950 et présentes de 1951 à 2022.

— Avec les données passées, nous avons réussi à calculer les saisonnalités (moyennes mensuelles pour chaque mois). Cette étape a été réalisée à l'aide de la fonction `groupby` sur l'axe `time.month` suivie du calcul de la moyenne.

— Après, nous avons calculé les anomalies passées ainsi que les anomalies présentes. Pour ce faire, les saisonnalités calculées ont été retarnchées des données.

— Pour finir, nous avons regroupé les anomalies par année puis avons calculé la moyenne des valeurs pour chaque groupe d'une année donnée.

$\Longrightarrow$ Ce traitement nous a permis d'obtenir des anomalies climatiques qu'on utilisera pour calculer le signal X ainsi que le bruit C_N.

b) **Calcul beta :**

— Nous avons implémenté une fonction `compute_covariance_and_signal` qui renvoie le signal ainsi que la matrice de covariance. Le signal est la différence entre les moyennes des anomalies passées et présentes.

— La fonction `calculate_beta` retourne alors le vecteur de dimension (lat*lon,1) aussi appelé détecteur optimal β

c) **Statistique de test :**

— Nous avons commencé par implémenter la fonction `calculate_gamma`.

— Ensuite, nous avons écrit la fonction `calculate_Z` qui retourne notre statistique de test sous forme de série temporelle.

$\Longrightarrow$ Maintenant que nous avons implémenté les fonctions nécessaires pour faire le test, nous allons passer au calibrage de ce dernier ainsi qu'à l'analyse des résultats obtenus.

3.2.3 Calibrage du test

a) **Travail sur les simulations** Le but de cette partie est de principalement tester la fiabilité de notre algorithme.
Nous l'avons alors entrainé sur le jeu données. L'idée est de prendre un fichier parmi les 50 qui nous servira de fichier test et utiliser les 49 restants pour calculer beta. On répète pour tous les autres fichiers.
Nous avons alors deux graphes par fichier : le premier avec les valeurs de Z au passé et le deuxième au présent. Voici un exemple :

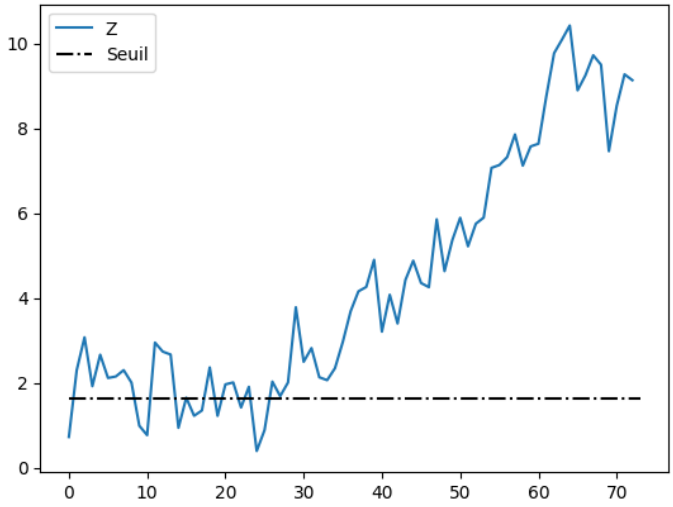

FIGURE 1 – Détection de changement climatique au présent

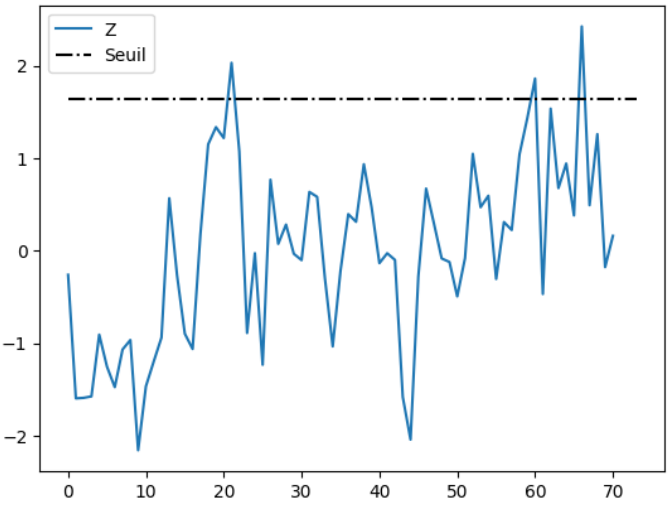

FIGURE 2 – Détection de changement climatique au passé

Nous avons eu des résultats similaires pour tous les fichiers, ce qui montre que notre algorithme détècte le changement climatique.

b) **Puissance et erreur type I**

Pour la partie d'avant, nous avons choisi un nombre de composantes aléatoire pour la PCA.

Cependant pour être sûr d'avoir le meilleur résultat, nous avons calculé la puissance du test ainsi que l'erreur de première espèce en faisant varier le nombre.

Le but est d'avoir la plus grande puissance possible tout en minimisant l'erreur de première espèce. Après analyse des résultats, nous avons opté pour n=70.

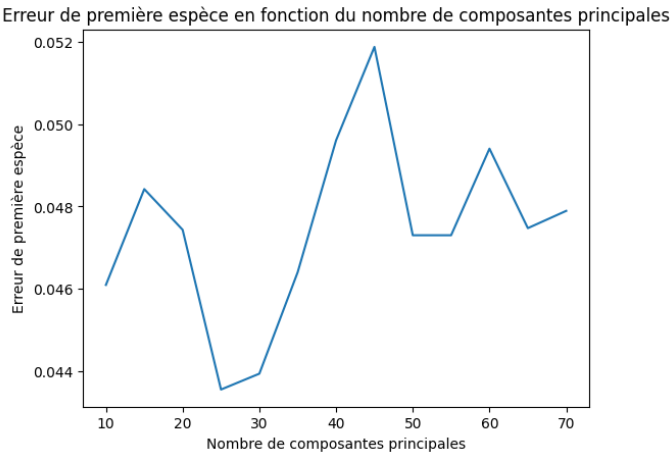

FIGURE 3 – Erreur première espèce en fonction de n

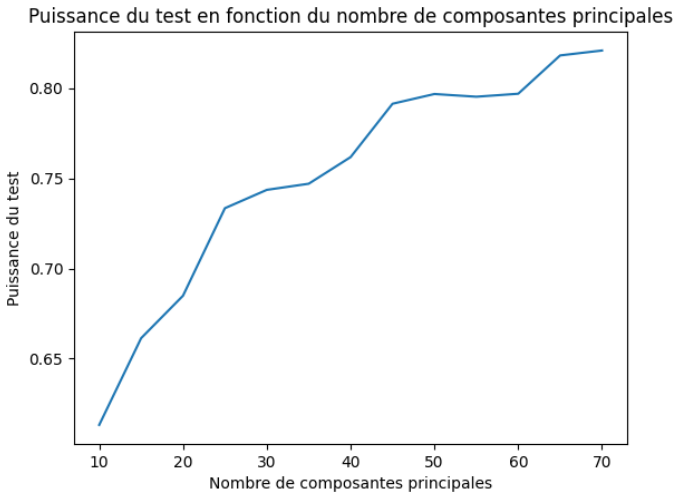

FIGURE 4 – Puissance du test en fonction de n

3.3 Application sur les données et résultats

a) Données données réelles recalibrés pour remplir les parties manquantes

b) Résultats ici y aura graphe et cartes

4 Phase 2 : Développement de nouvelles méthodes

5 Réflexion sur les aspects RSE

6 Conclusion

7 Bibliographie

8 Annexe