UE analyse des données environnementales : Contraintes observationelles pour les projections du climat futur

Didier Swingedouw & Valentin Portmann

Semaines du 4 mars et du 8 avril 2024

1 Introduction

1.1 Contexte

Les projections du climat futur, réalisées grâce aux modèles numériques de climat, sont capitales pour permettre aux preneurs de décisions d'anticiper les conséquences du changement climatique. Plusieurs de ces modèles climatiques sont regroupés en ce qui sont appelés les "Coupled Model Intercomparison Projects" (CMIPs), dont la dernière version est nommée CMIP6.

Cependant chaque modèle climatique fournit une projection du climat futur différente. Une facon d'utiliser cette base de modèles pour estimer une projection future est d'en faire la moyenne (moyenne multi-modèle). L'incertitude de cet estimateur peut alors être approximée par la dispersion (écart type) des projections des différents modèles. Cependant, il n'y a pas que cette incertitude qui importe. Au total, trois incertitudes sont admises en science climatique :

1. incertitude des émissions futures
2. variabilité interne du système climatique
3. incertitude de la réponse modélisée

L'incertitude liée au émissions futures (dit autrement, à quel point l'homme va continuer d'émettre des gaz à effet de serre, GES) est simplement représentée à l'aide de différents scénarios de développement socio-économique (SSP) allant d'une réduction drastique d'émission de GES à une forte augmentation si aucune réduction des émissons. Chaque modèle climatique de la base CMIP possède des résultats différents suivant le scénario choisis. Par la suite, on se placera dans un scénario d'émission fixe.

La variabilité interne est due à la propriété intrinsèquement chaotique du climat. Par exemple, celle ci désigne le fait que d'une année à l'autre, la température moyenne en France n'est pas la même et cela même si il n'y avait pas de réchauffement climatique.

Dans ce projet nous nous intéresseront à l'incertitude du point 3. L'objectif de ce projet est donc de diminuer l'incertitude de la température globale future estimée pour un scénario d'émission donnée.

1.2 Objectif

Une facon d'estimer la projection du climat futur est donc de réaliser une moyenne multi-modèles. Cependant tout les modèles ne se valent probablement pas. Il est intéressant de les comparer à la

réalité. Dans ce projet, vous allez coupler ces données simulées à des données réelles observées via différentes méthodes, pour estimer différemment qu'une moyenne multi-modèles, afin d'essayer de diminuer l'incertitude et d'affiner les projections futures.

1.3 Données

La variable étudiée est la température moyenne globale (moyennée dans le temps sur chaque année, et dans l'espace sur la totalité du globe). Nous travaillerons sur une anomalie de température, c'est à dire la température à laquelle on soustrait une température de référence. Cette réference est généralement prise avant que le réchauffement climatique ne commence vraiment à faire effet, en phase dite "pré-industrielle" définie ici entre 1850 et 1900.

Comme dit précédemment, nous utiliserons des données simulées et des données observées :

- 21 modèles climatiques, simulant chacun une série temporelle entre 1850 et 2099.
- 1 série temporelle d'observation entre 1850 et 2021, ainsi que son incertitude (liée aux appareils de mesure et à la couverture spatiale)

2 Cas univarié

2.1 Notations

On appelera Y la variable à prédire et X la variable utilisée pour prédire (prédicteur).
Exemple : X la moyenne entre 1950 et 2000, Y la moyenne entre 2091 et 2100.

	temps passés	temps futurs
Données simulées	X_i	Y_i
Données réelles	X_0	$\hat{Y}$

avec $i \in [\![1..M]\!]$, $M = 21$ le nombre de modèles climatiques.

2.2 Méthodes

2.2.1 Moyenne pondérée

Méthode dite de performance et d'indépendance, proposée dans l'article Brunner et al., 2019. L'idée est de trouver $\hat{Y}$ à partir des simulations futures Y_i, en les pondérants pour :

- privilégier celles dont la simulation passée X_i est proche de l'observation X_0 (critère de **performance**)
- privilégier celles dont la simulation passée X_i est éloignée des autres simulations passées X_j (critère d'**indépendance**)

Le calcul de l'estimateur de Y ($\hat{Y}$), est alors une moyenne pondérée des Y simulés (Y_i pour i allant de 1 à M) :

$$\hat{Y} = \sum_{i=1}^{M} w_i \, Y_i$$

$$w_i \propto \frac{exp(-D_i^2/\sigma_D^2)}{\sum_{j=1}^{M} exp(-S_{i,j}^2/\sigma_S^2)}$$

$$s.t. \sum_{i=1}^{M} w_i = 1$$

Où $D_i = |X_i - X_0|$ et $S_{i,j} = |X_i - X_j|$ et σ_D, σ_S sont des paramètres libres.

L'estimateur de la variance obtenue sera donc : $\hat{\sigma}^2 = \sum_{i=1}^{M} w_i \, (Y_i - \hat{Y})^2$

2.2.2 Régression linéaire

Cette méthode suppose qu'il y a un corrélation entre la température simulée passée et future, représentée dans la diversité de modèles climatiques. On propose alors le modèle suivant :

$Y = \beta_0 + \beta_1 \, X + \epsilon$

β_0 et β_1 sont deux paramètres estimés à partir des données simulées, puis l'observation est utilisée pour prédire : $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} \, X_0$. L'incertitude (écart type) peut être estimé à partir de l'incertitude de cette régression cumulée à celle apportée par l'observation.

$$\sigma^2_{regression} = s^2 \left[1 + \frac{1}{M} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{M} (Xi - \overline{X})^2} \right]$$

avec $s^2 = \frac{1}{M-2} \sum_{i=1}^{M} (Y_i - f(X_i))^2$. Voir annexe le détail des calculs.

2.2.3 One-step Kalman

La méthode décrite dans l'article "Bowman et al., 2018", est équivalente à un filtre de Kalman à une étape. Un filtre de Kalman permet de combiner un modèle théorique à des observations, et met à jour l'estimation à chaque nouvelle observation. Dans notre cas, on ne met à jour qu'une seule fois ("one-step"). Cette méthode fait un compromis entre simulation et observation, en prenant en compte leurs incertitudes respectives.

Voici un récapitulatif des résultats à utiliser :

$\hat{Y} = E[Y|X_0] = \mu_Y + \frac{\rho \sigma_Y \sigma_X}{\sigma_Y^2 + \sigma_X^2}(X_0 - \mu_X)$,
$\hat{\sigma}^2 = Var(Y|X_0) = (1 - \frac{\rho^2}{1 + (\sigma_B^2 / \sigma_X^2)})\sigma_Y^2$.

avec :

- μ_X et μ_Y, les espérances de X et Y estimées par une moyenne sur les modèles climatiques,
- σ_X et σ_Y, les incertitudes (écarts types) de X et Y estimées sur les modèles climatiques,
- σ_B l'écart type du bruit d'observation (incertitude de X_0),
- $\rho = \frac{Cov(Y,X)}{\sigma_X \, \sigma_Y}$ la corrélation entre X et Y, estimée sur les modèles climatiques.

3 Excercices

Pour la suite voici quelques packages python utiles :

- scikit-learn
- numpy
- matplotlib.pyplot
- pandas

3.1 Mise en forme des données

Transformez les données tel que :

- X : moyenne sur 1950-2000
- Y : moyenne sur 2090-2099

3.2 Moyenne multi-modèles

Quelle est la moyenne multi-modèle de Y et son incertitude à un écart type ? Cette incertitude sera la référence pour la suite : il faudra essayer de faire moins.

3.3 Moyenne pondérée

Implémentez la méthode de moyenne pondérée. Choisissez un paramétrage (σ_D, σ_S), et testez le.

Comparez son incertitude avec celle obtenue avec la moyenne multi-modèles. Y-a-t-il une amélioration ?

Etudiez l'impact des paramètres sur la projection et son incertitude. Comment ne prendre en compte que le critère de performance, ou que le critère d'indépendance ?

Comment régler ces paramètres ? Proposez deux méthodes et testez les (astuce : si vous êtes en manque d'idée, vous pouvez vous pencher sur une méthode de validation croisée).

3.4 Régression linéaire

Réalisez une régression linéaire entre X et Y sur les données simulées, puis prédire à partir de l'observation.

Comparez son incertitude avec celle obtenue avec la moyenne multi-modèles. Y-a-t-il une amélioration ?

Interpréter les coefficients de la régression linéaire. Que signifient-ils ?

La qualité de régression est-elle bonne ? Appuyez vous sur une métrique pour répondre à cette question.

3.5 One-step Kalman

Implémentez et testez cette méthode pour estimer la projection.

Comparez son incertitude avec celle obtenue avec la moyenne multi-modèles. Y-a-t-il une amélioration ?

En se basant sur la relation théorique, comment évolue l'incertitude en fonction de la corrélation ? Comment évolue l'incertitude en fonction du rapport signal bruit ?

3.6 Comparaison moyenne multi-modèles, moyenne pondérée, régression linéaire et Kalman

Comparez et interpréter les performances des différentes méthodes.

Quelles sont les différentes hypothèses faites au sein de chaque méthode sur la nature des données (linéarité, indépendance, gaussianité) ? Un modèle qui simule mieux le passé est-il nécessairement un modèle qui simule mieux le futur ? Discutez des limites de ces différentes approches.

Chaque méthode nécessite son lot de paramètres. Certains sont libres (par exemple σ_S et σ_D), d'autres sont calculés par la méthode (par exemple β_0), ou encore calculés en amont sur les données (par exemple μ_X ou encore ρ).
Discutez de la robustesse de ces méthodes face à leurs paramétrisation, ainsi que des hypothèses nécessaires au calcul de leurs paramètres. Y-a-t-il un risque de surapprentissage ?

3.7 Validation croisée

L'erreur théorique d'une méthode peut être différente de celle expérimentale calculée par validation croisée. Cependant, cette dernière se révèle très importante pour détecter la présence de sur-apprentissage, et/ou pour paramétrer une méthode. Il existe différentes approches de validation croisée, notamment les approches appelées Leave-One-Out (LOO) ainsi que K-fold. Le Leave-One-Out étant un cas particulier du K-fold. Quel est le principe de fonctionnement de la validation croisée K-fold et Leave-One-Out ? Pourquoi permet-elle de détecter la présence de sur-apprentissage ?

Dans notre cas de figure où nous avons peu d'échantillons (modèles climatiques), il est préférable d'utiliser un Leave-One-Out plutôt qu'un K-fold. Pourquoi ?

Utlisez une validation croisée de type Leave-One-Out pour calculer une erreur expérimentale par méthode. Comparez et expliquez ces performances.

3.8 Modification du prédicteur X

3.8.1 Autre prédicteur univarié

Testez ces différentes méthodes pour d'autres prédicteurs X, par exemple un découpage plus récent (moyenne entre 2000-2015) ou encore un découpage plus ou moins restreint dans le temps (moyennage sur 10 ans *vs* moyennage sur 100 ans). Comment évoluent les performances ? Comment expliquez vous ces variations ?

Comment l'utilisation d'une autre variable prédictrice que la température peut permettre d'améliorer les résultats ? Lesquelles seraient pertinentes ? Trouvez 3 variables qui auraient un intérêt.

Plutôt que prendre la moyenne, utilisez la tendance comme prédicteur. La tendance est définie comme la pente trouvée par régression linéaire sur la série temporelle d'un modèle climatique.

3.8.2 Multivarié

Jusqu'ici nous travaillons en univarié : un seul prédicteur pour estimer $\hat{Y}$. Une possibilité d'extension est de travailler en multivarié, en utilisant par exemple chaque année passé comme prédicteur. Etant donné qu'il y a des observations entre 1850 à 2021, cela fait potentiellement 172 prédicteurs. Mais une plage plus restreinte peut être bien sur sélectionnée.
Remarque : l'approche multivariée consiste plus communément à prendre comme prédicteurs différentes métriques (température globale, concentration en gaz à effet de serre, couverture nuageuse, *etc*), mais dans notre cas de figure nous allons essayer en prenant comme prédicteurs les différentes années pour une seule métrique, à savoir la température globale.

Adaptez certaines de ces méthodes, proposer et testez une diversité de nouvelles méthodes de votre choix pour pouvoir réaliser la projection en multivarié. Certaines de ces méthodes pourraient être non linéaire (Random Forest).

Comment avez vous paramétré ces méthodes ? Comparez les avec les méthodes précédentes, en terme d'hypothèses et de résultats. Quel problème peut poser cette approche multidimensionnelle ? Quel est le risque à apprendre une fonction de régression dans un espace à grande dimension ? Expliquez 3 solutions différentes pour réduire le nombre de variable (mots clés pour la recherche internet : feature extraction, feature selection).

Réalisez une analyse en composante principale sur ces données multivariées. Les données sont-elles très corrélées ? Combien faut-il de composantes principales pour les expliquer (sélectionner le critère le plus approprié) ? Appliquez les différentes méthodes de réduction du nombre de variable pour n'obtenir qu'une seule variable. Testez les méthodes univariés vu précédemment dessus, et analyser les résultats et performances.

References

Bowman, K. W., Cressie, N., Qu, X., & Hall, A. (2018). A hierarchical statistical framework for emergent constraints: Application to snow-albedo feedback. *Geophysical Research Letters*, *45*(23), 13, 050–13, 059. https://doi.org/https://doi.org/10.1029/2018GL080082

Brunner, L., Lorenz, R., Zumwald, M., & Knutti, R. (2019). Quantifying uncertainty in european climate projections using combined performance-independence weighting. *Environmental Research Letters*, *14*. https://doi.org/10.1088/1748-9326/ab492f

4 Annexe : démonstration de la variance de l'erreur en régression linéaire

Prédiction d'une nouvelle donnée

$$Y = \beta_0 + \beta_1 X_0 + \epsilon$$
$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_0$$

Erreur de prédiction : $e = Y - \hat{Y}$

Espérance de l'erreur de prédiction : $\sigma^2_{regression} = E[e] = (\beta_0 + \beta_1 X_0 + E[\epsilon]) - \left(E[\hat{\beta_0}] + E[\hat{\beta_1}] X_0\right)$

$$E[e] = (\beta_0 + \beta_1 X_0 + 0) - (\beta_0 + \beta_1 X_0)$$
$$E[e] = 0$$

Variance de l'erreur de prédiction : $Var(e) = Var(Y) - 2\,Cov(Y, \hat{Y}) + Var(\hat{Y})$

$$Var(e) = \sigma^2 \left[1 + \frac{1}{M} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{M} (X_i - \overline{X})^2}\right]$$

Démonstration :

$$\begin{aligned}
Cov(Y,\hat{Y}) &= E\left[(Y - E[Y])\,(\hat{Y} - E[\hat{Y}])\right] \\
&= E\left[(\epsilon)\,(\hat{\beta_0} + \hat{\beta_1}\,X_0 - \beta_0 - \beta_1\,X_0)\right] \\
&= E\left[\epsilon\right]\;E\left[\hat{\beta_0} + \hat{\beta_1}\,X_0 - \beta_0 - \beta_1\,X_0\right] \text{ (erreur } \epsilon \text{ indépendante)} \\
&= 0
\end{aligned}$$

$$Var(Y) = \sigma^2$$

$$\begin{aligned}
Var(\hat{Y}) &= Var(\hat{\beta_0} + \hat{\beta_1}\,X_0) \\
&= Var(\hat{\beta_0}) + X_0^2\,Var(\hat{\beta_1}) + 2\,X_0\,Cov(\hat{\beta_0},\hat{\beta_1}) \\
&= Var(\hat{\beta_0}) + X_0^2\,Var(\hat{\beta_1}) + 2\,X_0\,Cov(\overline{Y} - \hat{\beta_1}\,\overline{X},\hat{\beta_1}) \\
&= Var(\hat{\beta_0}) + X_0^2\,Var(\hat{\beta_1}) - 2\,X_0\,\overline{X}\,Cov(\hat{\beta_1},\hat{\beta_1}) \\
&= Var(\hat{\beta_0}) + (X_0^2 - 2\,X_0\,\overline{X})\,Var(\hat{\beta_1}) \\
&= \sigma^2\left[\frac{1}{M} + \frac{\overline{X}^2}{\sum_{i=1}^{M}(X_i - \overline{X})^2}\right] + \sigma^2\left[\frac{X_0^2 - 2\,X_0\,\overline{X}}{\sum_{i=1}^{M}(X_i - \overline{X})^2}\right] \\
&= \sigma^2\left[\frac{1}{M} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{M}(X_i - \overline{X})^2}\right]
\end{aligned}$$

Estimateur sans biais de σ^2 :

$$s^2 = \frac{1}{M-2}\;\sum_{i=1}^{M}(Y_i - \hat{Y_i})^2$$