

UNIVERSITÉ DE BORDEAUX

MASTER 1 — 2024/2025

UE Analyse des données environnementales

---

# Rapport de Projet

Projections du climat futur

---

**Réalisé par :**

Marwa Dades

Amélie de Maillard

Mathis Rita

**Encadrant :**

Didier Swingedouw

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Contexte du projet . . . . .	2
1.2	Objectifs . . . . .	2
<b>2</b>	<b>Partie 1 : Manipulation des données</b>	<b>2</b>
2.1	Présentation générale des données . . . . .	2
2.2	Identification des matrices . . . . .	3
2.3	Modèles climatiques disponibles . . . . .	3
2.4	Périodes temporelles couvertes . . . . .	4
2.5	Analyse de l'évolution des anomalies de température . . . . .	4
2.6	Analyse statistique sur l'année 2099 . . . . .	5
2.7	Validation de l'hypothèse de normalité . . . . .	5
2.8	Lien avec les scénarios SSP . . . . .	6
<b>3</b>	<b>Partie 2 : Méthodes de réduction d'incertitude</b>	<b>7</b>
3.1	Transformation du problème en régression univariée . . . . .	7
3.2	Question Bonus . . . . .	7
3.3	Moyenne multi-modèle . . . . .	8
3.4	Moyenne pondérée . . . . .	8
3.5	Régression linéaire . . . . .	11
3.6	One-Step Kalman . . . . .	13
3.7	Comparaison des méthodes . . . . .	15
3.8	Validation croisée Leave-One-Out . . . . .	20
3.9	Modification du prédicteur $X$ . . . . .	21
3.10	Approche multivariée . . . . .	25
<b>4</b>	<b>Conclusion</b>	<b>34</b>

## 1. Introduction

### 1.1. Contexte du projet

Face à l'urgence climatique, il est essentiel d'améliorer les projections du climat futur afin d'anticiper au mieux les impacts environnementaux, sociaux et économiques. Ces projections sont réalisées à l'aide de modèles numériques du climat, regroupés dans le cadre des *Coupled Model Intercomparison Projects* (CMIP), dont la dernière version en date est le CMIP6. Chaque modèle fournit une projection différente, ce qui reflète l'incertitude liée à la modélisation du système climatique.

Trois sources majeures d'incertitude sont identifiées :

- L'incertitude liée aux émissions futures de gaz à effet de serre (différents scénarios SSP),
- La variabilité naturelle interne du système climatique,
- L'incertitude de la réponse modélisée par les différents modèles.

Dans le cadre de ce projet, nous nous concentrerons sur cette dernière source d'incertitude, en analysant comment les modèles climatiques du CMIP6 simulent l'évolution de la température moyenne globale à long terme. Pour cela, des données simulées et observées de température moyenne globale (exprimées en anomalies) ont été mises à disposition, couvrant la période de 1850 à 2099.

### 1.2. Objectifs

L'objectif principal de ce projet est de proposer des méthodes statistiques pour améliorer la précision des projections climatiques futures, en réduisant l'incertitude liée à la modélisation. L'approche classique consiste à prendre la moyenne des simulations de tous les modèles (moyenne multi-modèle), mais cette méthode ne prend pas en compte la qualité ou la similarité des modèles.

Nous explorons ainsi plusieurs méthodes pour améliorer les projections :

- Une moyenne pondérée des modèles basée sur leur performance passée et leur indépendance,
- Une régression linéaire entre les simulations passées et futures,
- Une méthode inspirée du filtre de Kalman pour intégrer l'incertitude observationnelle,
- Des analyses multivariées pour exploiter davantage d'informations temporelles.

L'enjeu est de proposer une projection future de la température plus fiable et plus précise, tout en évaluant la robustesse de chaque méthode via validation croisée. Ce travail s'inscrit dans une démarche scientifique rigoureuse visant à affiner les outils d'aide à la décision en matière de politique climatique.

## 2. Partie 1 : Manipulation des données

### 2.1. Présentation générale des données

Pour ce projet, plusieurs ensembles de données ont été mis à disposition afin d'analyser l'évolution de la température moyenne globale dans le cadre des projections climatiques. Ces données incluent :

- Des anomalies de température simulées sur la période passée,
- Des anomalies de température simulées pour le futur,
- Des anomalies de température observées (passé),
- Une incertitude associée aux observations passées.

Chaque série temporelle est associée à un ensemble de labels permettant d'identifier :

- Les différents modèles climatiques (issus du projet CMIP6),
- Les années passées (1850–2021),
- Les années futures (2022–2099).

Les données sont stockées sous forme de matrices NumPy et ont été chargées dans notre environnement de travail Python.

## 2.2. Identification des matrices

Afin de reconnaître à quoi correspond chaque matrice, nous avons examiné les dimensions des fichiers et leur structure. Nous avons pu associer les matrices de la manière suivante :

- `data_simulated_past` → Matrice simulant l'anomalie passée : `matrix_3`,
- `data_simulated_future` → Matrice simulant l'anomalie future : `matrix_1`,
- `data_observed_past` → Matrice contenant les observations passées : `matrix_2`,
- `data_observed_past_sigma` → Matrice d'incertitudes (écart-type) associées aux observations : `matrix_4`.

## 2.3. Modèles climatiques disponibles

Les simulations climatiques disponibles sont issues de 25 modèles numériques différents. Ces modèles sont notamment utilisés dans le cadre du CMIP6. Voici la liste des modèles intégrés dans notre analyse :

- |                 |                |
|-----------------|----------------|
| — ACCESS-CM2    | — GFDL-ESM4    |
| — ACCESS-ESM1-5 | — GISS-E2-1-G  |
| — AWI-CM-1-1-MR | — INM-CM5-0    |
| — BCC-CSM2-MR   | — IPSL-CM6A-LR |
| — CAMS-CSM1-0   | — KACE-1-0-G   |
| — CAS-ESM2-0    | — KIOST-ESM    |
| — CESM2         | — MCM-UA-1-0   |
| — CIESM         | — MIROC6       |
| — CMCC-ESM2     | — MRI-ESM2-0   |
| — CNRM-CM6-1    | — NESM3        |
| — CanESM5       | — TaiESM1      |
| — EC-Earth3     | — UKESM1-0-LL  |
| — FIO-ESM-2-0   |                |

## 2.4. Périodes temporelles couvertes

Les simulations ainsi que les observations s'étendent sur différentes périodes :

- **Données passées (observées et simulées)** : de 1850 à 2021,
- **Données futures (simulées)** : de 2022 à 2099.

Ces plages temporelles permettent de mener des analyses historiques comparatives, mais aussi de projeter des tendances futures selon différents scénarios d'émissions.

## 2.5. Analyse de l'évolution des anomalies de température

L'évolution des anomalies de température a été tracée pour l'ensemble des modèles climatiques ainsi que pour les données observées. Ces visualisations permettent de comparer directement les simulations avec la réalité mesurée.

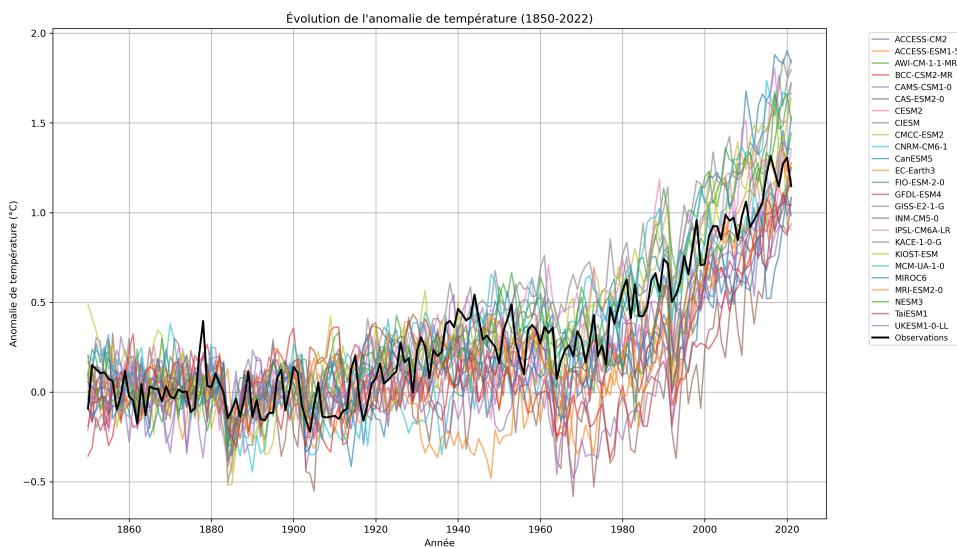


FIGURE 1 – Evolution des anomalies de température (1850-2022)

Globalement, les modèles simulent une hausse progressive de l'anomalie de température sur la période 1850–2022. On observe que la série d'observations s'inscrit dans la bande de dispersion des modèles, ce qui témoigne d'une certaine cohérence entre les données simulées et les données réelles. Cependant, on peut voir des écarts importants pour des simulations qui sont à prendre en compte. La courbe des observations reste centrées par rapport aux différentes simulations.

Pour affiner cette analyse, nous avons ensuite calculé la moyenne multi-modèle ainsi que son incertitude (écart type entre modèles) pour chaque année. Cette série moyenne a été comparée aux observations et à leur propre incertitude. On constate que les observations sont en général bien encapsulées dans l'enveloppe d'incertitude simulée, ce qui renforce la crédibilité globale des modèles.

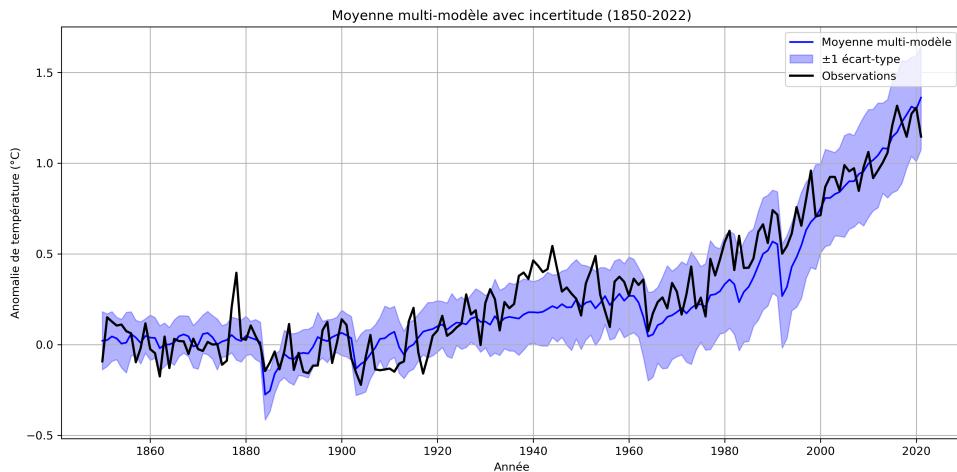


FIGURE 2 – Moyenne multi-modèle avec incertitude (1850-2022)

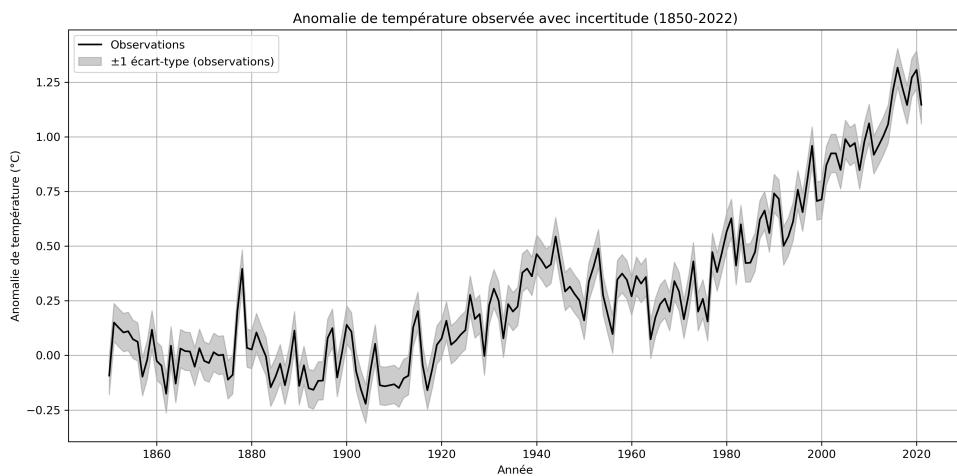


FIGURE 3 – Anomalie de température observée avec incertitude (1850-2022)

## 2.6. Analyse statistique sur l'année 2099

L'année 2099 a été choisie comme horizon pour évaluer les projections climatiques futures. Nous avons extrait, pour chaque modèle climatique, l'anomalie de température prédictive en cette année. En faisant l'hypothèse que la distribution des données simulées suit une loi normale (hypothèse raisonnable à ce stade), nous avons pu calculer les intervalles de confiance suivants :

- Intervalle à 68% : [4.40 ; 6.66] °C
- Intervalle à 95% : [3.27 ; 7.78] °C
- Intervalle à 99% : [2.14 ; 8.91] °C

Ces valeurs donnent une idée de l'incertitude associée aux projections, en prenant en compte uniquement la dispersion inter-modèle.

## 2.7. Validation de l'hypothèse de normalité

Afin de valider l'utilisation de la loi normale dans le calcul des intervalles ci-dessus, nous avons procédé à deux vérifications :

- **QQ-plot (Quantile-Quantile plot)** : le tracé indique un bon alignement des quantiles empiriques avec les quantiles théoriques d'une loi normale.

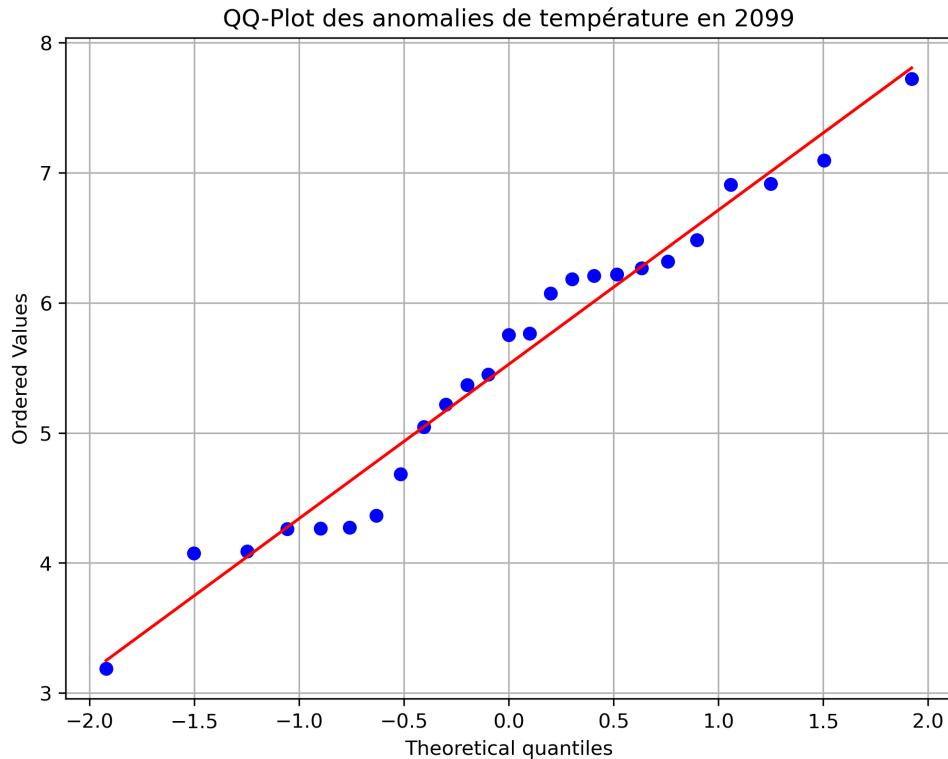


FIGURE 4 – QQ-Plot des anomalies de température en 2099

- **Test de Shapiro-Wilk** : la p-value obtenue est de 0.5636, supérieure au seuil de 0.05, ce qui permet de ne pas rejeter l'hypothèse de normalité.

Ainsi, l'hypothèse gaussienne est raisonnable dans ce contexte.

## 2.8. Lien avec les scénarios SSP

Les simulations utilisées dans ce projet sont réalisées sous un scénario d'émissions de gaz à effet de serre donné. Ces scénarios sont appelés *SSP* (Shared Socio-economic Pathways), sont des narratifs, traduits en ensembles d'hypothèses socio-économiques (Population, Éducation, Urbanisation, PIB). Ces narratifs décrivent des évolutions alternatives de la société future en l'absence de changement climatique ou de politique climatique. Cinq narratifs ont été construits par le GIEC, chacun étant numéroté de 1 à 5. Parmi eux, on distingue :

- **SSP1-1.9 et SSP1-2.6** : scénario optimiste avec forte action climatique (baisse des émissions, réchauffement limité),
- **SSP2-4.5** : scénario moyen (monde moyen, réchauffement modéré),
- **SSP3-7.0** : scénario pessimiste (conflits, peu de coopération internationale, fortes émissions),
- **SSP5-8.5** : scénario catastrophe (économie fossile extrême, émissions très élevées).

On a le tableau suivant qui correspond aux niveaux de réchauffement par scénario et par horizon (en °C, « best estimate ») :

	Court terme : 2021–2040	Moyen terme : 2041–2060	Long terme : 2081–2100
SSP1-1.9	1.5	1.6	1.4
SSP1-2.6	1.5	1.7	1.8
SSP2-4.5	1.5	2.0	2.7
SSP3-7.0	1.5	2.1	3.6
SSP5-8.5	1.6	2.4	4.4

Source : <https://www.carbone4.com/publication-scenarios-ssp-adaptation>

En comparant les anomalies projetées en 2099 (souvent comprises entre 4°C et 6°C), il semble probable que les données simulées correspondent au scénario SSP5-8.5, qui est le plus pessimiste. Ce scénario suppose une poursuite massive des émissions sans politique de limitation.

Cette première analyse exploratoire pose les bases nécessaires à une meilleure compréhension des données utilisées pour affiner les projections climatiques dans la suite du projet.

### 3. Partie 2 : Méthodes de réduction d'incertitude

#### 3.1. Transformation du problème en régression univariée

Pour simplifier l'analyse et appliquer des méthodes statistiques interprétables, le problème est transformé en une version univariée. On cherche ici à prédire l'anomalie de température moyenne sur la période 2090–2099 (variable  $Y$ ) à partir de l'anomalie moyenne sur la période 1950–2000 (variable  $X$ ).

- $X_{\text{simu}}$  : anomalie moyenne passée simulée (1950–2000), pour chaque modèle climatique,
- $Y_{\text{simu}}$  : anomalie moyenne future simulée (2090–2099), pour chaque modèle,
- $X_{\text{obs}}$  : anomalie moyenne passée observée (1950–2000), issue des données historiques.

Ces variables sont extraites à partir des matrices de simulation et d'observation, en sélectionnant les années concernées puis en calculant la moyenne annuelle correspondante. On obtient ainsi trois vecteurs :

- $X_{\text{simu}} \in \mathbb{R}^{M \times 1}$ ,
- $Y_{\text{simu}} \in \mathbb{R}^{M \times 1}$ ,
- $X_{\text{obs}} \in \mathbb{R}^{1 \times 1}$ .

Ce format est parfaitement adapté à l'utilisation de méthodes telles que la régression linéaire, la moyenne pondérée ou le filtre de Kalman.

#### 3.2. Question Bonus

Nous avons pu voir que l'écart type empirique de  $Y_{\text{simu}}$  est : 1.036. Ainsi nous avons une incertitude de 1.036°C autour de la moyenne des anomalies de température projetées pour 2099, selon les modèles simulés.

**Comment utiliser aussi  $X_{\text{simu}}$  ou encore  $X_{\text{obs}}$  pour diminuer l'incertitude ? Proposez différentes pistes et/ou testez les.**

Afin de diminuer l'incertitude dans les prévisions futures, nous pouvons utiliser  $X_{\text{simu}}$  (les anomalies passées simulées) et  $X_{\text{obs}}$  (l'anomalie observée passée), surtout si on veut prendre en compte des relations entre les données passées et futures pour améliorer la prédiction des anomalies futures et réduire les erreurs.

- Utiliser  $X_{\text{simu}}$  et  $X_{\text{obs}}$  dans un modèle de régression multiple :

Nous pouvons modéliser une relation entre les anomalies passées ( $X_{\text{simu}}$  ou  $X_{\text{obs}}$ ) et les anomalies futures ( $Y_{\text{simu}}$ ), pour utiliser ces relations pour réduire l'incertitude de la prédiction des anomalies futures. La régression multiple permet de combiner plusieurs facteurs ( $X_{\text{simu}}$ ,  $X_{\text{obs}}$ ) pour prédire  $Y_{\text{simu}}$ .

- Faire une combinaison des modèles climatiques :

Nous pouvons pondérer les modèles en fonction de leur performance. Nous pouvons pondérer les valeurs de  $Y_{\text{simu}}$  en fonction de leur précision passée sur les anomalies observées. Cela pourrait réduire l'incertitude. La pondération permet de donner plus de poids aux modèles les plus fiables et ainsi obtenir une prévision plus stable. Cela pourrait réduire l'incertitude en ne laissant pas les modèles moins fiables perturber la moyenne globale.

### 3.3. Moyenne multi-modèle

Avant de tester des approches avancées, la première estimation naturelle consiste à utiliser la moyenne des valeurs simulées futures  $Y_{\text{simu}}$  par l'ensemble des modèles climatiques :

$$\hat{Y}_{\text{multi-modèle}} = \frac{1}{M} \sum_{i=1}^M Y_i \quad (1)$$

L'incertitude associée est quantifiée par l'écart type des valeurs  $Y_i$  :

$$\sigma_{\text{multi-modèle}} = \sqrt{\frac{1}{M} \sum_{i=1}^M (Y_i - \hat{Y})^2} \quad (2)$$

Résultats :

- Moyenne multi-modèle :  $\hat{Y} = 5.221 \text{ } ^\circ\text{C}$ ,
- Incertitude (1  $\sigma$ ) :  $1.036 \text{ } ^\circ\text{C}$ .

Cette estimation constitue notre référence de départ. Les méthodes suivantes chercheront à améliorer cette prévision, notamment en réduisant l'incertitude sans compromettre la validité statistique du modèle.

### 3.4. Moyenne pondérée

La méthode de moyenne pondérée cherche à affiner la projection climatique en pondérant différemment les modèles selon deux critères :

- **Critère de performance** : les modèles les plus proches de l'observation  $X_{\text{obs}}$  sont favorisés,
- **Critère d'indépendance** : les modèles très similaires entre eux sont pénalisés pour éviter la redondance.

Cette approche est inspirée de l'article de Brunner et al. (2019). Chaque modèle  $i$  reçoit un poids  $w_i$  défini par :

$$w_i = \frac{\exp\left(-\frac{D_i^2}{\sigma_D^2}\right)}{1 + \sum_{j \neq i}^M \exp\left(-\frac{S_{i,j}^2}{\sigma_S^2}\right)} \quad (3)$$

où :

- $D_i = |X_i - X_0|$  : la distance entre le modèle  $i$  et les observations  $X_0$ , ce qui mesure la performance du modèle par rapport aux données observées passées.
- $S_{i,j} = |X_i - X_j|$  : la distance entre les modèles  $i$  et  $j$ , ce qui mesure leur indépendance.
- $\sigma_D$  et  $\sigma_S$  : paramètres d'échelle contrôlant respectivement l'importance des critères de performance et d'indépendance.

Les poids sont ensuite normalisés pour que  $\sum_i w_i = 1$ . La projection pondérée s'écrit :

$$\hat{Y}_{\text{pondérée}} = \sum_{i=1}^M w_i Y_i \quad (4)$$

et l'estimateur de la variance obtenue est :

$$\hat{\sigma}_{\text{pondérée}}^2 = \sum_{i=1}^M w_i (Y_i - \hat{Y})^2 \quad (5)$$

### Application avec paramètres fixés

Pour une première application, nous fixons les paramètres à  $\sigma_D = 0.8$  et  $\sigma_S = 0.2$ . Le calcul donne :

- Moyenne pondérée :  $\hat{Y} = 5.356$  °C,
- Incertitude associée :  $\sigma = 1.030$  °C.

Cette méthode améliore légèrement l'incertitude par rapport à la moyenne multi-modèle ( $1.030 < 1.036$ ), ce qui montre que pondérer les modèles peut être pertinent, même si la moyenne centrale reste proche.

### Analyse de l'impact des paramètres

Les paramètres  $\sigma_D$  et  $\sigma_S$  permettent de moduler la sensibilité du poids à la performance et à l'indépendance :

- $\sigma_D \rightarrow 0$  : seuls les modèles proches de l'observation sont sélectionnés,
- $\sigma_S \rightarrow 0$  : les modèles les plus différents sont favorisés,
- $\sigma_D, \sigma_S \rightarrow \infty$  : tous les modèles sont considérés équivalents, on retrouve la moyenne simple.

**Cas particuliers :**

- Pour ne considérer que la performance :  $\sigma_S \gg 1$ ,
- Pour ne considérer que l'indépendance :  $\sigma_D \gg 1$ .

## Optimisation des paramètres via validation croisée

Pour améliorer le choix des paramètres, nous avons implémenté une validation croisée Leave-One-Out (LOO). L'idée est de retirer successivement un modèle, d'estimer sa prédiction avec les autres, puis de calculer l'erreur quadratique. On cherche alors à minimiser l'erreur moyenne sur tous les modèles.

Un balayage sur une grille de valeurs de  $\sigma_D$  et  $\sigma_S$  nous a permis de trouver :

- $\sigma_D^* = 0.536$ ,
- $\sigma_S^* = 0.536$ ,
- Erreur quadratique moyenne minimale :  $\mathbb{E}[(\hat{Y}_k - Y_k)^2] = 1.152$ .

Cela indique une robustesse de la méthode malgré une incertitude encore présente entre les modèles.

## Étude de sensibilité des paramètres

Nous avons ensuite étudié comment varie l'incertitude estimée en fonction des valeurs de  $\sigma_D$  et  $\sigma_S$  :

- En fixant  $\sigma_D$  et en faisant varier  $\sigma_S$ , on constate que l'incertitude est plus faible pour de petites valeurs de  $\sigma_S$ , ce qui favorise la diversité des modèles.

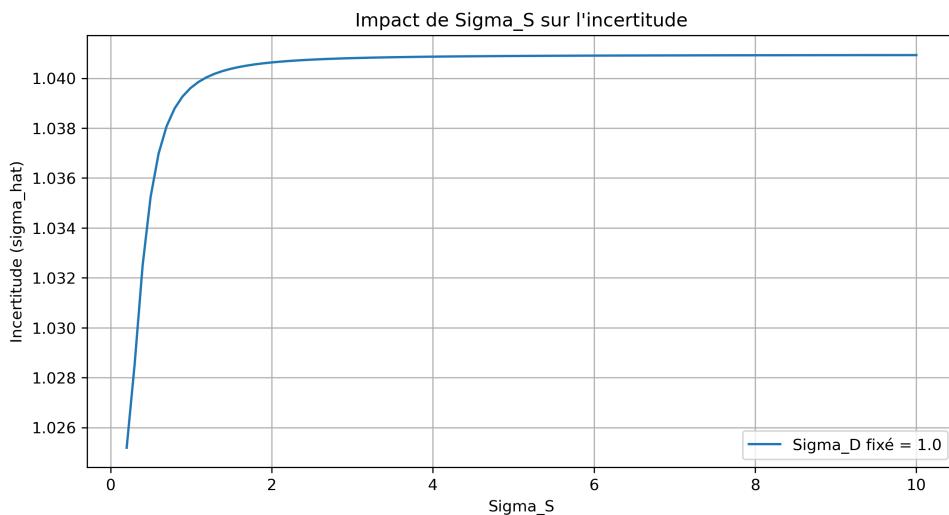
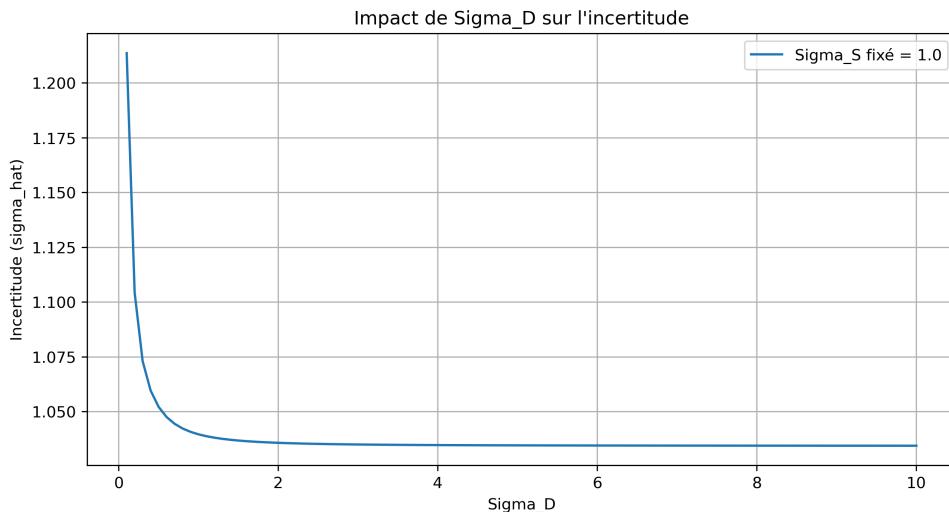


FIGURE 5 – Impact de  $\sigma_S$  sur l'incertitude

- En fixant  $\sigma_S$  et en faisant varier  $\sigma_D$ , l'incertitude diminue lorsque les modèles performants sont davantage valorisés.

FIGURE 6 – Impact de  $\sigma_D$  sur l'incertitude

Ces observations confortent l'idée qu'une combinaison équilibrée entre performance et indépendance permet de minimiser l'incertitude, en exploitant la complémentarité des modèles.

### Conclusion sur la moyenne pondérée

Cette méthode montre un potentiel intéressant : même si l'amélioration de la moyenne centrale est limitée, l'incertitude peut être réduite de manière significative en choisissant judicieusement les paramètres. Cela justifie son usage comme alternative crédible à la moyenne brute des modèles.

## 3.5. Régression linéaire

Une autre méthode d'estimation consiste à modéliser une relation linéaire entre les anomalies de température passées et futures simulées. Cette approche repose sur l'hypothèse qu'il existe une corrélation entre le comportement passé et futur des modèles climatiques.

La relation est modélisée comme suit :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

où :

- $X$  est l'anomalie moyenne passée (1950–2000),
- $Y$  est l'anomalie moyenne future (2090–2099),
- $\beta_0$  et  $\beta_1$  sont les coefficients de la régression,
- $\varepsilon$  est le bruit (erreur résiduelle).

Le modèle est ajusté à l'aide des paires  $(X_{\text{simu}}, Y_{\text{simu}})$ , puis on utilise  $X_{\text{obs}}$  pour estimer  $Y$ .

### Résultats de la régression :

- Prédiction  $\hat{Y}_{\text{obs}} = 5.298^{\circ}\text{C}$ ,

— Incertitude associée (1 écart type) : 1.097 °C.

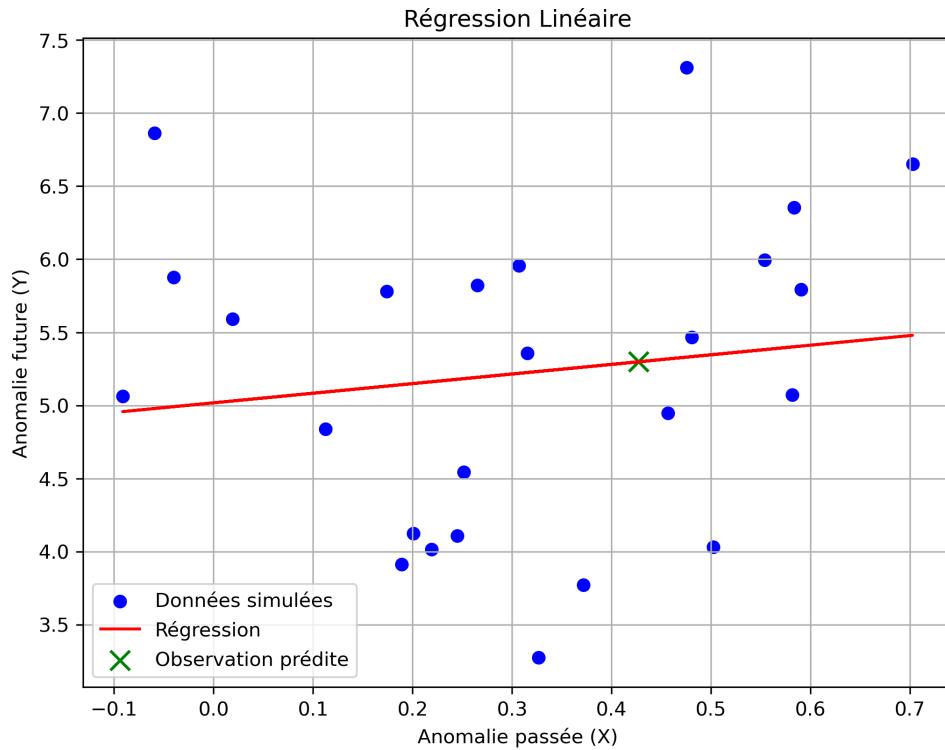


FIGURE 7 – Régression linéaire

### Interprétation des coefficients

Le modèle estimé est :

$$Y = 0.658 \cdot X + 5.017$$

- **Ordonnée à l'origine** ( $\beta_0 = 5.017$ ) : même en l'absence d'anomalie passée, le modèle prédit déjà environ +5°C d'anomalie en 2090–2099. Cela reflète une forte tendance au réchauffement, structurellement présente dans les modèles.
- **Pente** ( $\beta_1 = 0.658$ ) : chaque degré d'anomalie supplémentaire dans le passé est associé à une hausse d'environ +0.66°C dans le futur.

Cette faible pente montre que les différences entre modèles dans le passé influencent peu leur projection future. Cela suggère une convergence des modèles vers un futur réchauffé, indépendamment de leurs divergences passées.

### Comparaison de l'incertitude

L'incertitude obtenue ici est légèrement plus élevée que celle de la moyenne multi-modèle :

$$\sigma_{\text{régression}} = 1.097 \quad > \quad \sigma_{\text{multi-modèle}} = 1.036$$

La régression ne permet donc pas, dans ce cas, de réduire l'incertitude. Cela peut être dû à l'absence de réelle dépendance linéaire entre les données passées et futures.

## Évaluation de la qualité de la régression

Le coefficient de détermination  $R^2$  est utilisé pour évaluer la qualité d'un modèle linéaire. Il est défini par :

$$R^2 = 1 - \frac{\text{Variance des résidus}}{\text{Variance totale}}$$

- $R^2 = 1$  : régression parfaite,
- $R^2 = 0$  : la régression n'explique rien,
- $0 < R^2 < 1$  : qualité proportionnelle à la valeur.

**Résultat obtenu :**  $R^2 = 0.019$

Ce faible score ( $R^2 \approx 1.9\%$ ) indique que la régression linéaire ne capture quasiment aucune variabilité des anomalies futures. Cela suggère qu'il n'existe pas de lien linéaire fort entre les anomalies passées et futures simulées, ce que confirme également la forte dispersion du nuage de points observé.

## Conclusion sur la régression linéaire

La régression linéaire fournit une estimation simple mais peu efficace dans ce cas. Elle ne permet pas de réduire l'incertitude, et son pouvoir explicatif est très limité. Cela montre les limites d'une modélisation linéaire dans un système aussi complexe et potentiellement non linéaire que le climat global.

### 3.6. One-Step Kalman

La méthode *One-Step Kalman* est inspirée du filtre de Kalman classique, et permet ici d'améliorer la projection des anomalies de température futures en combinant de façon optimale les observations et les simulations climatiques, tout en tenant compte des incertitudes.

#### Principe de la méthode

On suppose que les anomalies passées  $X$  et futures  $Y$  suivent une loi jointe gaussienne. Sous cette hypothèse, la meilleure prédition de  $Y$  conditionnellement à une observation  $X_0$  est donnée par :

$$\hat{Y} = \mu_Y + \frac{\rho\sigma_Y\sigma_X}{\sigma_Y^2 + \sigma_X^2}(X_0 - \mu_X)$$

où :

- $\mu_X, \sigma_X$  : moyenne et incertitude (écart-type) de  $X$ ,
- $\mu_Y, \sigma_Y$  : moyenne et incertitude (écart-type) de  $Y$ ,
- $\sigma_B$  : incertitude (écart type du bruit) sur l'observation  $X_0$ ,
- $\rho = \frac{\text{cov}(Y,X)}{\sigma_Y\sigma_X}$  : corrélation entre  $X$  et  $Y$ .

L'incertitude étant la  $\sqrt{Var(Y|X_0)}$  alors l'incertitude associée à cette prédition est :

$$\sqrt{\hat{\sigma}_{\text{Kalman}}^2} = \sqrt{\left(1 - \frac{\rho^2}{1 + \frac{\sigma_B^2}{\sigma_X^2}}\right)\sigma_Y^2}$$

## Résultats numériques

En appliquant cette méthode sur nos données :

- Projection :  $\hat{Y} = 5.294 \text{ } ^\circ\text{C}$ ,
- Incertitude :  $\hat{\sigma}_{\text{Kalman}} = 1.027 \text{ } ^\circ\text{C}$ .

La méthode réduit donc légèrement l'incertitude par rapport à la moyenne multi-modèles ( $1.027 < 1.036$ ), ce qui montre son efficacité dans notre cas.

## Évolution de l'incertitude avec la corrélation

Théoriquement, l'incertitude évolue selon :

$$\hat{\sigma}_{\text{Kalman}}^2 = \left( 1 - \frac{\rho^2}{1 + \frac{\sigma_B^2}{\sigma_X^2}} \right) \sigma_Y^2$$

On observe que :

- Plus la corrélation  $\rho$  entre  $X$  et  $Y$  est forte, plus l'incertitude est réduite.
- Si  $\rho \rightarrow 1$ , l'incertitude tend vers zéro (en l'absence de bruit).
- Si  $\rho \rightarrow 0$ , l'incertitude reste celle de la moyenne simple sur  $Y$ .

Un graphique de l'incertitude en fonction de  $\rho$  pour différents niveaux de bruit  $\sigma_B$  illustre cette dynamique : plus l'observation est précise (bruit faible), plus la corrélation aide à réduire l'incertitude.

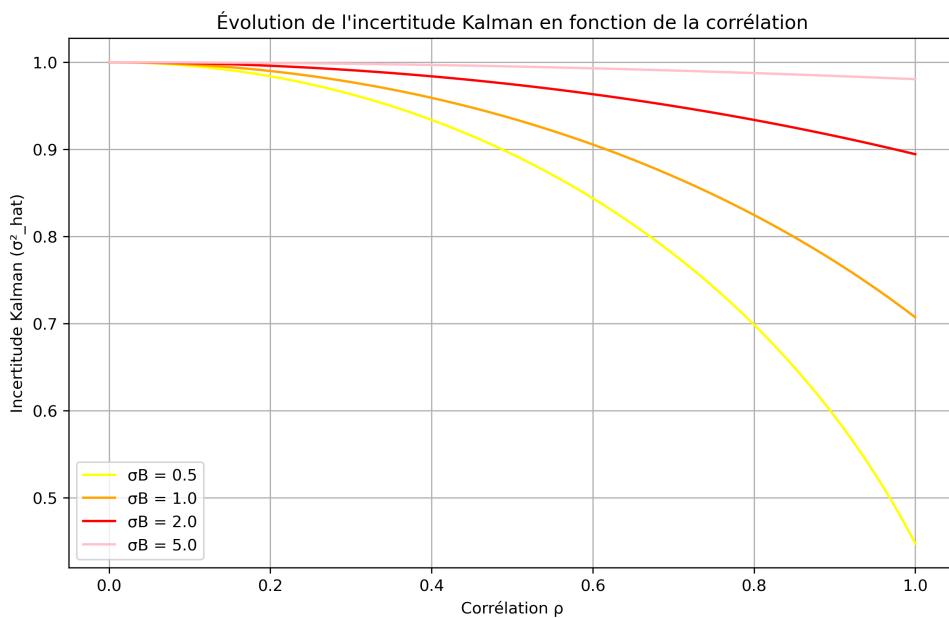


FIGURE 8 – Évolution de l'incertitude Kalman en fonction de la corrélation

## Évolution de l'incertitude avec le bruit d'observation

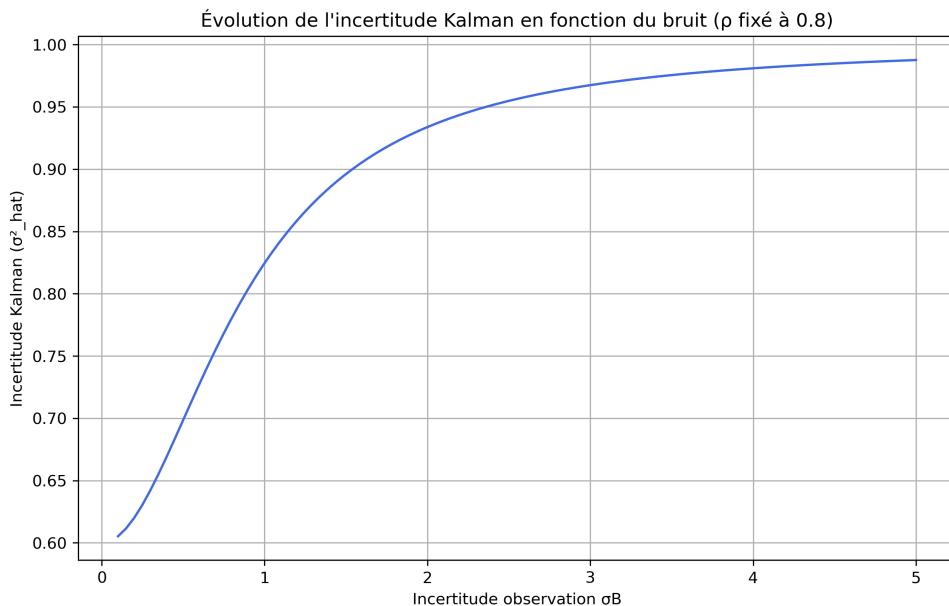


FIGURE 9 – Évolution de l'incertitude Kalman en fonction du bruit à  $\rho$  fixé

Lorsque l'on fixe  $\rho$  (ex :  $\rho = 0.8$ ) et que l'on fait varier  $\sigma_B$ , on observe que :

- Plus le bruit d'observation  $\sigma_B$  est faible, plus l'incertitude est réduite efficacement,
- Lorsque  $\sigma_B$  devient grand, l'incertitude augmente et la méthode de Kalman perd en efficacité.

Ce qui est logique car dans le calcul de l'incertitude nous avons, pour  $\rho^2$  le dénominateur  $1 + \frac{\sigma_B^2}{\sigma_X^2}$ , donc :

- Si le bruit  $\sigma_B$  est faible (observation très précise) cela implique que  $\sigma_B^2 \ll \sigma_X^2$  donc le dénominateur est proche de 1,  $\rho$  est "conservé" et  $\hat{\sigma}^2$  diminue, donc l'incertitude  $\hat{\sigma}$  est réduite.

- Si le bruit  $\sigma_B$  est grand (observation peu fiable) cela implique que  $\sigma_B^2 \gg \sigma_X^2$  donc le dénominateur devient grand, on divise  $\rho^2$  par un grand nombre alors  $\hat{\sigma}^2$  augmente, donc l'incertitude  $\hat{\sigma}$  est plus grande.

Plus l'observation est précise par rapport à la variabilité des modèles, plus l'incertitude est réduite.

Ces observations montrent que la méthode One-Step Kalman est particulièrement avantageuse lorsque les observations sont précises et que la corrélation entre passé et futur est relativement forte.

### 3.7. Comparaison des méthodes

Nous avons appliqué et comparé quatre méthodes différentes pour projeter l'anomalie de température moyenne future (2090–2099) :

- Moyenne multi-modèles,
- Moyenne pondérée,
- Régression linéaire,
- One-Step Kalman.

## Résultats synthétiques

Le tableau suivant résume les résultats obtenus :

Méthode	Projection moyenne ( $\hat{Y}$ )	Incertitude ( $\hat{\sigma}$ )
Moyenne multi-modèles	5.221 °C	1.036 °C
Moyenne pondérée	5.356 °C	1.030 °C
Régression linéaire	5.298 °C	1.097 °C
One-Step Kalman	5.294 °C	1.027 °C

Un graphique comparatif a également été réalisé pour visualiser simultanément la projection centrale ( $\hat{Y}$ ) et l'incertitude associée ( $\hat{\sigma}$ ) pour chaque méthode.

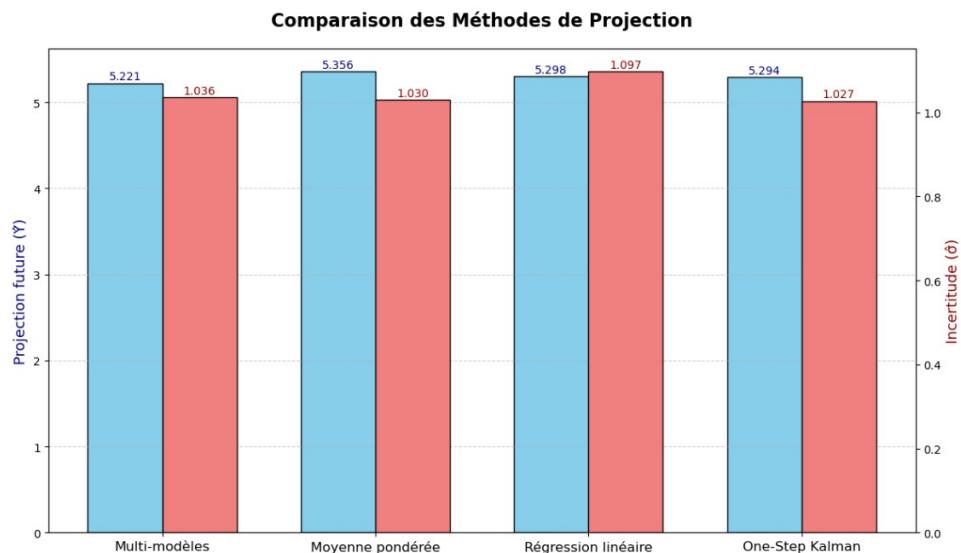


FIGURE 10 – Comparaison des Méthodes de Projection

## Analyse comparative

Grâce à ce graphique nous pouvons comparer chacun des modèles. Pour cela, nous allons nous intéresser aux incertitudes. En effet, nous remarquons que l'incertitude de référence de la moyenne multi-modèles est de 1.036. Le but, était de trouver, avec d'autres méthodes, une incertitude plus faible, nous constatons :

- **Moyenne pondérée** : légère amélioration ( $1.030 < 1.036$ ). Cette réduction dépend fortement du choix des paramètres  $\sigma_D$  et  $\sigma_S$  de pondération.
- **Régression linéaire** : augmentation de l'incertitude ( $1.097 > 1.036$ ). Cela s'explique par la mauvaise qualité de la relation linéaire entre anomalies passées et futures (confirmée par un  $R^2$  très faible).
- **One-Step Kalman** : meilleure réduction de l'incertitude (1.027), meilleure méthode parmi celles testées. Cela s'explique par la prise en compte explicite des incertitudes et de la corrélation entre passé et futur.

## Conclusions principales

La méthode *One-Step Kalman* apparaît comme la plus robuste et la plus efficace pour réduire l'incertitude sur la projection future. Elle combine de manière optimale l'information passée et présente, tout en tenant compte des erreurs de mesure.

À l'inverse, la régression linéaire simple, en raison de l'absence de relation forte entre passé et futur dans les données, est peu performante ici.

Enfin, la moyenne pondérée montre un potentiel intéressant, mais dépendant d'un choix judicieux des paramètres d'ajustement.

Cette comparaison souligne l'importance de choisir une méthode de projection adaptée à la structure réelle des données, en évitant des hypothèses trop fortes (comme la linéarité stricte), et en valorisant l'information disponible sur les incertitudes.

## Analyse des hypothèses sous-jacentes aux méthodes

Chaque méthode étudiée repose sur certaines hypothèses implicites ou explicites concernant la nature des données, notamment sur la linéarité, l'indépendance des erreurs ou la distribution (gaussianité). Ces hypothèses impactent fortement la robustesse et la validité des projections obtenues.

### Moyenne multi-modèles

La moyenne multi-modèles consiste à prendre la moyenne simple des projections issues de tous les modèles climatiques disponibles. Cette approche suppose :

- **Hypothèses :**
- Tous les modèles ont une qualité similaire,
- Les erreurs des modèles sont indépendantes.

Cependant, aucune hypothèse forte n'est faite sur la linéarité ou la gaussianité des données. Les limites sont que cette approche ne tient pas compte du fait que certains modèles pourraient mieux représenter le climat passé que d'autres. Elle peut diluer la qualité de la prévision en mélangeant de bons et de mauvais modèles sans distinction.

### Moyenne pondérée

La moyenne pondérée introduit des poids différenciés pour chaque modèle, selon sa performance et son indépendance. Cette méthode repose sur l'idée que les modèles plus fiables pour le passé seront également plus fiables pour le futur, et qu'une diversité de modèles réduit le risque d'erreur systématique.

- **Hypothèses :**
- Stabilité de la qualité des modèles du passé vers le futur,
- Diversité entre les modèles (l'indépendance doit être significative).

Cependant, on retrouve quelques limites à cette méthode. Tout d'abord, la capacité d'un modèle à bien simuler le passé n'implique pas nécessairement une bonne performance sur des conditions climatiques futures différentes. De plus, le choix des paramètres de pondération ( $\sigma_D$  et  $\sigma_S$ ) influence fortement les résultats, et reste délicat car il peut induire un surapprentissage.

## Régression linéaire

La régression linéaire cherche à établir une relation directe entre anomalies passées ( $X$ ) et futures ( $Y$ ) :

- **Hypothèses :**

- Existence d'une relation linéaire entre  $X$  et  $Y$ ,
- Résidus de la régression gaussiens, indépendants et homoscédastiques (variance constante).

Il y a également des limites à la régression linéaire. Pour commencer, Si la relation entre  $X$  et  $Y$  est non linéaire, la régression linéaire peut induire des biais importants. L'hypothèse de linéarité est souvent irréaliste en climatologie où les réponses du système peuvent être non linéaires. Ici, nous avons notamment une faible corrélation observée dans nos données ( $R^2 \approx 0.019$ ) donc une absence de lien fort entre passé et futur. Ensuite, la méthode est sensible aux valeurs extrêmes et suppose que l'incertitude d'observation est négligeable.

## One-Step Kalman

La méthode de Kalman repose sur la combinaison optimale de l'observation et de la simulation, en prenant en compte explicitement les incertitudes :

- **Hypothèses :**

- $X$  et  $Y$  suivent une loi jointe gaussienne,
- Incertitudes correctement estimées.

Il existe tout de même des limites. Tout d'abord, la gaussianité est une hypothèse forte : si les distributions de  $X$  ou  $Y$  sont asymétriques ou multi-modales, la méthode peut être sous-optimale. D'un autre côté, la précision dépend de la bonne estimation des variances associées aux observations et simulations. La méthode est sensible à la mauvaise estimation des variances ou de la corrélation  $\rho$ .

## Un modèle qui simule mieux le passé est-il nécessairement un modèle qui simule mieux le futur ?

Il est important de noter que la capacité d'un modèle climatique à reproduire fidèlement les observations passées ne garantit pas sa capacité à prédire le futur. Le futur climatique est associé à des conditions de forçage différentes (par exemple, augmentation rapide des gaz à effet de serre). De plus, certains modèles peuvent être adaptés pour bien représenter le climat historique sans pour autant capturer correctement les réponses à ces nouveaux forçages.

## Robustesse face à la paramétrisation

Chaque méthode explorée dans cette étude repose sur un ensemble de paramètres, qui peuvent être classés en trois grandes catégories :

1. Paramètres libres : ceux qu'on choisit ou ajuste (par exemple,  $\sigma_D$  et  $\sigma_S$  dans la moyenne pondérée),
2. Paramètres estimés par la méthode : comme les coefficients  $\beta_0$ ,  $\beta_1$  en régression linéaire,

3. Paramètres calculés en amont sur les données : tels que la moyenne  $\mu_X$ , la variance  $\sigma_X^2$ , ou encore la corrélation  $\rho$  dans la méthode de Kalman.

Certaines méthodes sont particulièrement sensibles à la manière dont les paramètres sont choisis. En effet, la moyenne pondérée, est très dépendante du choix de  $\sigma_D$  et  $\sigma_S$ . Des valeurs mal choisies peuvent conduire à sur-pondérer un petit groupe de modèles ou, au contraire, à lisser trop fortement les différences, annulant l'effet de pondération. De plus, la régression linéaire repose sur un ajustement automatique, mais suppose une stabilité de la relation linéaire entre  $X$  et  $Y$ , ce qui peut ne pas tenir dans un contexte de changement climatique.

Mais d'autres méthodes sont plus robustes. La moyenne multi-modèles n'implique aucun paramètre ajustable, ce qui en fait une méthode stable mais peu discriminante et le filtre de Kalman dépend de paramètres statistiques calculés sur l'ensemble des modèles, et donc moins arbitraires. Toutefois, sa robustesse dépend directement de la qualité des estimations de variances et de corrélation.

### Hypothèses nécessaires à l'estimation des paramètres

Chaque méthode fait des hypothèses implicites ou explicites pour estimer ses paramètres. La régression linéaire suppose une indépendance des résidus, leur normalité, et une relation linéaire globale, ce qui est rarement garanti en climatologie. Le choix des  $\sigma_D$  et  $\sigma_S$  dans la moyenne pondérée ne repose sur aucune théorie solide : on les ajuste souvent empiriquement, par validation croisée ou minimisation d'une incertitude. L'estimation de  $\rho$  (corrélation) dans la méthode de Kalman suppose une relation linéaire entre  $X$  et  $Y$ .

### Discussion sur le surapprentissage

Certaines méthodes présentent un risque de surapprentissage :

- **Moyenne pondérée** : si les paramètres  $\sigma_D$  et  $\sigma_S$  sont sur-optimisés pour les données passées, il y a un risque de biais et de perte de généralisation.
- **Régression linéaire** : surajustement possible si peu de données ou si la relation réelle n'est pas linéaire.

D'autres méthodes sont plus robustes :

- **Moyenne multi-modèles** : aucun paramètre ajusté, donc insensible au surapprentissage,
- **One-Step Kalman** : bien que reposant sur des statistiques calculées (moyennes, variances, corrélation), le risque est plus limité tant que les données sont fiables.

### Conclusion

La performance d'une méthode dépend autant de sa structure que de la qualité et de la robustesse de ses paramètres. Plus une méthode utilise de paramètres ajustables, plus elle peut théoriquement s'adapter aux données, mais plus elle est sujette au risque de surapprentissage. Il est donc crucial de valider toute méthode sur des données indépendantes ou à l'aide de techniques comme la validation croisée.

### 3.8. Validation croisée Leave-One-Out

La validation croisée est une méthode permettant d'évaluer la robustesse d'un modèle en simulant des conditions de test sur des données non vues. Cette approche est essentielle pour détecter la présence d'un éventuel sur-apprentissage et pour ajuster les paramètres d'un modèle de manière optimale.

#### Principes de la validation croisée K-Fold et Leave-One-Out

**Validation croisée K-Fold :** Dans la validation croisée K-Fold, l'ensemble des données est divisé en  $K$  sous-groupes (ou "folds"). Le modèle est entraîné sur  $K - 1$  folds et testé sur le dernier. Ce processus est répété  $K$  fois, chaque fold jouant tour à tour le rôle de jeu de test. La performance finale est obtenue en moyennant les erreurs des  $K$  tests.

**Validation Leave-One-Out (LOO) :** Le Leave-One-Out (LOO) est un cas particulier du K-Fold où  $K$  est égal au nombre total d'échantillons. Chaque observation est tour à tour utilisée comme jeu de test, tandis que le reste est utilisé pour l'entraînement. Si nous avons  $M = 25$  modèles climatiques, alors le LOO effectue 25 entraînements/tests. Cela maximise l'utilisation des données pour l'apprentissage, mais est plus coûteux en temps de calcul.

#### Pourquoi la validation croisée permet-elle de détecter le sur-apprentissage ?

Un modèle est en sur-apprentissage (overfitting) lorsqu'il s'ajuste trop précisément aux données d'entraînement, capturant le bruit au lieu des tendances générales. Cela conduit à une forte performance sur les données d'entraînement, mais à une mauvaise généralisation sur des données nouvelles en test.

La validation croisée permet de détecter ce phénomène en mesurant la performance d'un modèle sur des données non utilisées lors de l'apprentissage. Si un modèle affiche un écart important entre la performance en apprentissage et en validation comme de très bonnes performances sur les données d'entraînement, mais de mauvaises performances moyennes en validation croisée, cela signale un surajustement. De plus, elle aide à ajuster les paramètres (comme  $\sigma_D$ ,  $\sigma_S$  ou les coefficients en régression) en maximisant la performance sur des données non vues.

La validation croisée est indispensable pour évaluer la robustesse d'un modèle et éviter qu'il ne s'adapte trop aux données disponibles. Elle est particulièrement utile lorsqu'on dispose de peu de données.

#### Pourquoi utiliser le Leave-One-Out dans notre contexte ?

Dans notre cas, nous avons peu d'échantillons (modèles climatiques). Chaque donnée étant précieuse, il est essentiel d'utiliser la totalité des modèles pour l'entraînement et de tester sur un seul modèle à la fois.

Les avantages du LOO sont :

- **Utilisation maximale des données :** chaque modèle est utilisé pour l'entraînement  $M - 1$  fois, ce qui correspond à presque l'ensemble des données disponibles. Cela permet d'obtenir une estimation plus fiable des performances du modèle, un point particulièrement important lorsque les données sont rares.

- **Moins de perte d'information** : contrairement au K-Fold où 20-30% des données sont mises de côté à chaque itération, ici un seul modèle est exclu. Cela minimise les risques de sous-apprentissage liés à une réduction excessive des données d'entraînement.
- **Estimation plus stable** : chaque modèle climatique étant testé indépendamment, l'évaluation est plus fine. Cela est crucial dans un contexte où la variabilité entre les modèles climatiques est importante, et où l'on cherche à éviter le surapprentissage.
- **Absence de paramètre arbitraire** : le LOO a l'avantage de ne pas nécessiter le choix d'un paramètre  $K$  comme dans le K-Fold, ce qui peut être délicat avec un petit échantillon. Il utilise au contraire le nombre maximal d'itérations possibles, renforçant ainsi la robustesse de l'évaluation.

Ainsi, le Leave-One-Out est particulièrement adapté pour notre contexte où le nombre de modèles est limité. Il maximise l'exploitation des données, fournit une estimation stable des performances, et limite les biais liés à la partition. Dans le contexte des simulations climatiques avec peu de modèles disponibles, il représente donc une stratégie optimale.

### Application et comparaison des performances

Nous appliquons la validation Leave-One-Out sur les quatre méthodes de projection :

Méthode	Erreur RMSE (Leave-One-Out)
Moyenne multi-modèle	1.079
Moyenne pondérée	1.082
Régression linéaire	1.118
One-Step Kalman	1.085

### Analyse des résultats

Les erreurs obtenues montrent qu'aucune méthode ne surpassé de manière significative les autres en termes d'erreur expérimentale, ce qui met en évidence la difficulté du problème ainsi que le faible nombre d'échantillons disponibles. La moyenne multi-modèle, bien que naïve, représente une base fiable. Cela souligne également l'importance d'un bon choix des paramètres pour optimiser les méthodes avancées.

Nous avons été un peu surpris par le fait que la méthode One-Step Kalman ne présente pas la plus faible erreur RMSE, malgré le fait qu'elle soit celle associée à la plus faible incertitude dans nos prédictions. Cela peut s'expliquer par plusieurs facteurs : d'une part, le nombre réduit d'échantillons rend la validation Leave-One-Out particulièrement sensible aux variations locales dans les données ; d'autre part, il n'est pas exclu qu'une erreur d'implémentation ait pu affecter les performances de cette méthode. Une analyse plus poussée pourraient permettre de mieux comprendre cette divergence entre incertitude théorique et performance empirique.

### 3.9. Modification du prédicteur $X$

#### Changement de la période de moyennage

Au lieu de prendre comme prédicteur la moyenne des anomalies de température entre 1850 et 1900, nous avons testé différentes fenêtres temporelles :

- 1850-1900
- 1900-2000
- 1950-2000
- 1990-2000
- 2000-2015

**Impact sur l'incertitude des méthodes** Nous avons recalculé les incertitudes obtenues pour chacune des méthodes en fonction de la période choisie :

Période	Multi-modèle	Pondérée	Régression	Kalman
1850-1900	1.036	1.036	1.044	1.036
1900-2000	1.036	1.019	1.061	1.036
1950-2000	1.036	1.030	1.053	1.027
1990-2000	1.036	1.007	1.015	0.989
2000-2015	1.036	1.006	0.945	0.931

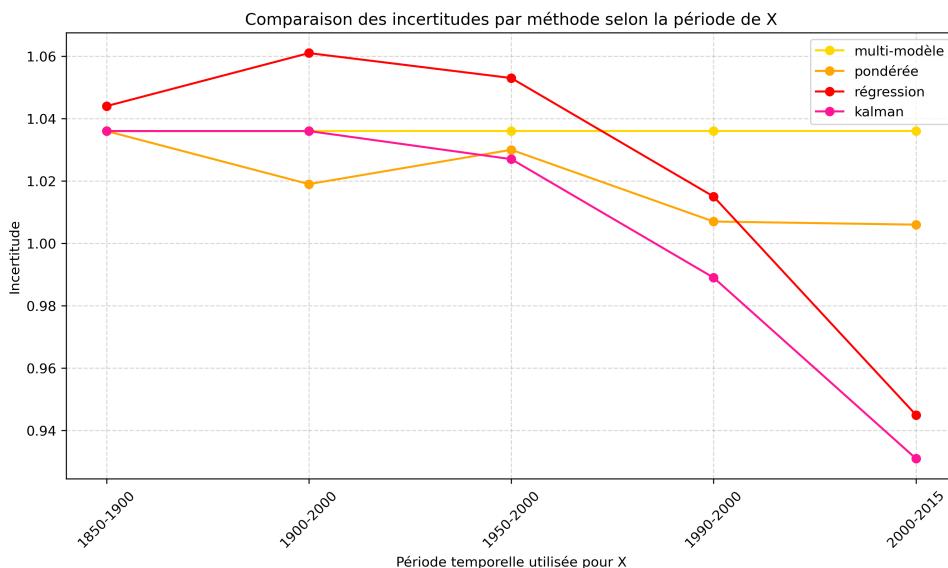


FIGURE 11 – Comparaison des incertitudes par méthode selon la période de X

**Analyse :** Les résultats montrent que plus la période de moyennage utilisée est récente, en particulier la fenêtre 2000–2015, plus l'incertitude associée aux prédictions est réduite. Ce phénomène s'observe notamment avec la méthode de Kalman, dont l'incertitude passe de 1.036 pour la période 1850–1900 à 0.931 pour la période 2000–2015.

Ce gain peut s'expliquer par le fait que les périodes récentes capturent de manière plus précise les tendances climatiques actuelles, fortement influencées par l'augmentation des émissions de gaz à effet de serre. En intégrant cette dynamique récente, les méthodes bénéficient d'une meilleure représentativité des conditions contemporaines, ce qui se traduit par une diminution de l'incertitude.

### Impact sur l'erreur de validation croisée (Leave-One-Out)

Les erreurs expérimentales (RMSE) obtenues pour chaque période temporelle sont :

Période	Multi-modèle	Pondérée	Régression	Kalman
1850-1900	1.079	1.079	1.127	1.079
1900-2000	1.079	1.084	1.124	1.086
1950-2000	1.079	1.082	1.118	1.085
1990-2000	1.079	1.060	1.072	1.067
2000-2015	1.079	1.045	1.016	1.042

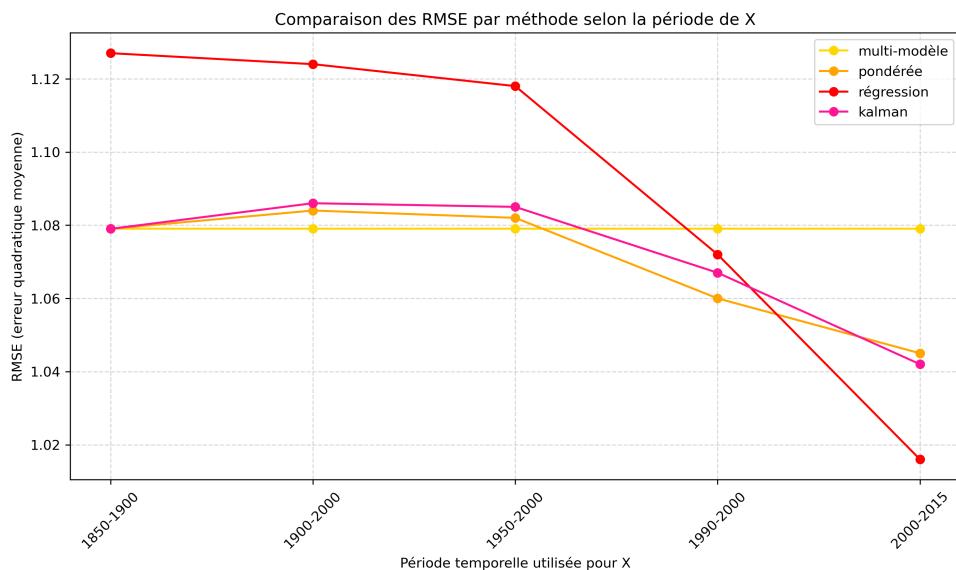


FIGURE 12 – Comparaison des RMSE par méthode selon la période de X

**Analyse :** Les résultats montrent une légère diminution des erreurs RMSE lorsque des périodes de moyennage plus récentes sont utilisées. Cette amélioration est particulièrement marquée pour la régression linéaire, qui bénéficie pleinement de l'utilisation des anomalies récentes comme prédicteur.

Ce gain de performance s'explique par le fait que les tendances climatiques actuelles, caractérisées par un réchauffement accéléré, sont mieux capturées sur des périodes récentes telles que 2000–2015. En s'alignant davantage sur les dynamiques contemporaines, ces fenêtres temporelles permettent aux modèles de mieux représenter la réalité climatique, améliorant ainsi la précision des prédictions.

### Utilisation d'autres variables prédictives

Utiliser uniquement l'anomalie de température comme prédicteur est limitatif. L'ajout d'autres variables prédictives, en plus de la température, permettrait d'améliorer la précision des modèles de prévision climatique. En effet, comme vu en cours, la température à un moment donné est influencée par une multitude de facteurs qui peuvent expliquer et affiner les prévisions. Si l'on comprend mieux l'impact de ces facteurs, il devient possible d'améliorer les prédictions de la température future, tout en réduisant les marges d'erreur.

Trois exemples de variables pertinentes :

- **Concentration en gaz à effet de serre (CO<sub>2</sub>, CH<sub>4</sub>)** : Ces gaz, comme le dioxyde de carbone (CO<sub>2</sub>) et le méthane (CH<sub>4</sub>), jouent un rôle clé dans l'augmentation des températures mondiales. Ils piègent la chaleur dans l'atmosphère, contribuant au réchauffement climatique. En quantifiant leur concentration dans l'atmosphère, on peut obtenir une estimation plus précise de la tendance future des températures.
- **Précipitations** : La pluviométrie influence la température de manière directe, par exemple à travers l'évaporation, ou indirecte en modifiant les conditions météorologiques locales. En connaissant les patterns de précipitations dans une région donnée, on peut affiner la prévision de la température, car une forte pluviométrie peut refroidir temporairement une zone, tandis qu'une sécheresse prolongée pourrait favoriser une montée des températures.
- **Couverture nuageuse** : La présence et le type de nuages affectent considérablement l'équilibre énergétique de la Terre. Les nuages peuvent soit retenir la chaleur (effet de serre), soit la réfléchir dans l'espace (effet de refroidissement). Une analyse de la couverture nuageuse, en particulier dans les zones sensibles, pourrait aider à mieux prédire les variations de température à court et à moyen terme.

Ces variables sont toutes prédictibles et quantifiables, et leur inclusion permettrait de modéliser plus précisément les phénomènes climatiques. En tenant compte de ces facteurs, les modèles climatiques pourraient mieux capturer les variations naturelles et humaines des températures, offrant ainsi des prévisions plus fiables pour l'avenir.

## Utilisation de la tendance comme prédicteur

Au lieu de prendre une moyenne sur une période, nous avons aussi testé un autre prédicteur : **la tendance linéaire** (pente de la régression linéaire de l'anomalie de température sur 1950-2000).

Les résultats obtenus sont :

### RMSE

- RMSE multi-modèle : **1.079**
- RMSE moyenne pondérée : **1.079**
- RMSE régression linéaire : **1.038**
- RMSE Kalman : **1.079**

### Incertitudes associées

- Incertitude multi-modèle : **1.036**
- Incertitude pondérée : **1.036**
- Incertitude régression linéaire : **0.972**
- Incertitude Kalman : **1.036**

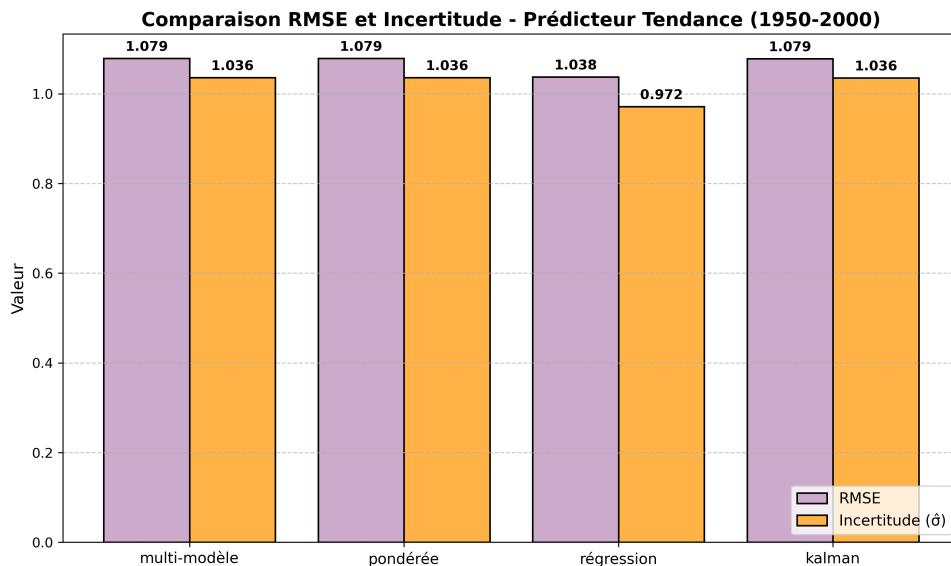


FIGURE 13 – Comparaison RMSE et incertitude des tendances comme prédicteur

**Analyse :** L'utilisation de la tendance (pente de régression) comme prédicteur, au lieu de la moyenne, permet une légère amélioration de l'incertitude dans le cas de la régression linéaire. Cette amélioration s'explique par le fait que la pente reflète plus fidèlement la dynamique récente du réchauffement climatique, contrairement à la moyenne qui lisse les variations temporelles.

La régression linéaire semble ainsi tirer profit de cette information supplémentaire, ce qui renforce sa capacité à capter les évolutions en cours. En revanche, pour les méthodes fondées sur des moyennes, qu'il s'agisse de la moyenne multi-modèle ou du filtre de Kalman, l'utilisation de la tendance n'apporte pas de gain significatif en termes de performance ou de réduction d'incertitude.

**Conclusion sur la modification de  $X$  :** Les analyses montrent que l'utilisation de périodes de référence plus récentes améliore systématiquement la réduction de l'incertitude, en phase avec la dynamique actuelle du climat. Par ailleurs, remplacer la moyenne par la tendance linéaire comme prédicteur s'avère particulièrement pertinent pour les méthodes de régression, car cela permet de mieux capter l'évolution récente du système climatique.

Enfin, l'introduction de nouvelles variables climatiques explicatives, telles que les concentrations de CO<sub>2</sub>, les précipitations ou encore la couverture nuageuse, constituerait une piste prometteuse pour affiner davantage les prédictions et enrichir la modélisation des mécanismes en jeu.

### 3.10. Approche multivariée

#### Introduction

Jusqu'ici, nous avons travaillé en **univarié** : un seul prédicteur (par exemple une moyenne ou une tendance de température) était utilisé pour estimer  $\hat{Y}$ . Une extension naturelle est de passer à une approche **multivariée**, c'est-à-dire d'utiliser simultanément plusieurs prédicteurs.

Dans notre étude, nous proposons de considérer chaque année passée comme un prédicteur distinct.

Puisque nous disposons de données entre 1850 et 2021, cela représente potentiellement 172 prédicteurs. Cependant, il est souvent pertinent de restreindre la période, par exemple aux années récentes.

**Remarque importante :** Dans un contexte classique, la multivariée consiste à utiliser des variables différentes (température, concentration en CO<sub>2</sub>, précipitations, etc.). Ici, nous nous limitons à une seule métrique (la température globale) mais pour plusieurs années.

### Extraction des prédicteurs multivariés

À titre d'exemple, nous extrayons les températures simulées pour la période **1990–2015**. Cela définit pour chaque modèle climatique un vecteur de dimension 25.

- $X_{\text{simu\_multi}}$  : Températures simulées entre 1990 et 2015 pour chaque modèle.
- $X_{\text{obs\_multi}}$  : Températures observées sur la même période.

### Formule de l'incertitude en régression multivariée

L'incertitude sur la prédiction  $\hat{Y}$  en multivarié est donnée par :

$$\hat{\sigma} = \sqrt{s^2 \left( 1 + \frac{1}{M} + (x_0 - \bar{X})^T (X^T X)^{-1} (x_0 - \bar{X}) \right)}$$

où :

- $x_0$  : vecteur des prédicteurs observés ( $X_{\text{obs\_multi}}$ ),
- $\bar{X}$  : moyenne des colonnes de  $X_{\text{simu\_multi}}$ ,
- $s^2$  : moyenne des carrés des résidus d'apprentissage,
- $M$  : nombre de modèles climatiques disponibles.

### Résultats sans réduction de dimension

Une régression multivariée simple, combinée à une validation croisée Leave-One-Out (LOO) a servi de référence et conduit aux résultats suivants :

- **RMSE** = 1.774
- **Incertitude sur la prédiction** ≈ 0.000

Le très faible niveau d'incertitude contrastant avec une RMSE élevée révèle un problème majeur de surapprentissage, dû à une dimension des prédicteurs élevée par rapport au nombre d'observations : 41 variables pour 25 modèles. Cela peut causer une forte instabilité des prédictions sur de nouvelles données.

### Approches alternatives pour la régression multivariée

Afin de remédier aux problèmes de surapprentissage observés avec la régression linéaire multivariée classique, nous avons exploré plusieurs méthodes de régression régularisée ou non linéaire, mieux adaptées aux situations où le nombre de prédicteurs est élevé par rapport au nombre d'observations.

- **La régression Ridge** ajoute une pénalité sur la norme des coefficients dans le but de limiter la variance du modèle, ce qui est particulièrement utile dans les contextes multivariés de haute dimension. Nous avons utilisé une validation croisée pour sélectionner le paramètre de régularisation optimal  $\lambda$  sur une grille logarithmique entre  $10^{-3}$  et  $10^3$ .
- **La méthode PLS (Partial Least Squares)** projette les données dans un espace de plus faible dimension tout en maximisant la covariance entre les prédicteurs et la variable cible. Elle est donc particulièrement adaptée aux problèmes multi-colinéaires. Le nombre de composantes a été fixé à 5, ce qui permet de projeter les données dans un espace latent réduit tout en maximisant la covariance entre prédicteurs et cible.
- **La méthode Random Forest** (méthode non linéaire) est basé sur un ensemble d'arbres de décision (ici 100), chaque arbre étant entraîné sur un bootstrap des données.

## Résultats comparatifs

Méthode	RMSE	Incertitude
Régression linéaire	1.774	0.000
Régression Ridge	1.240	0.800
Régression PLS	1.556	0.439
Random Forest	0.805	0.647

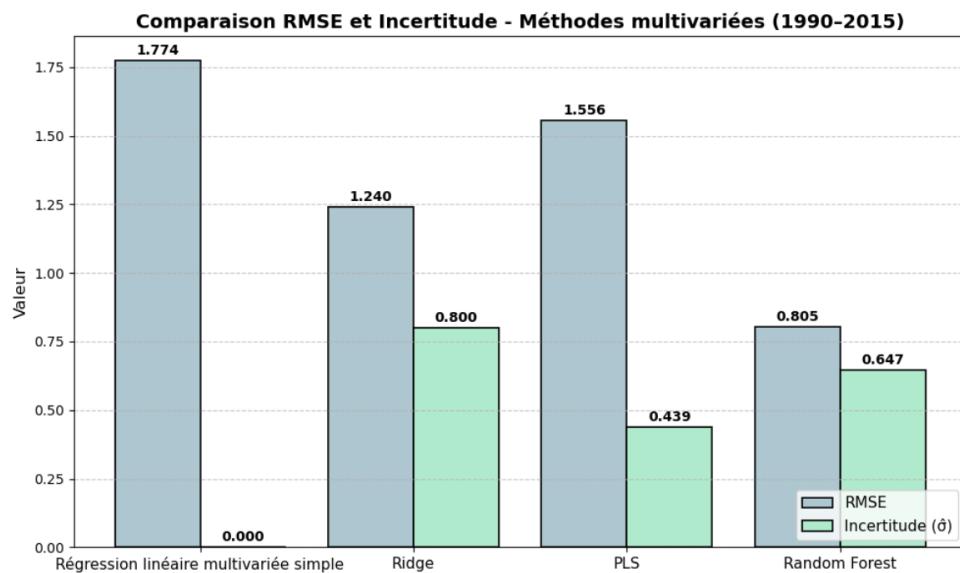


FIGURE 14 – Comparaison du RMSE et de l'incertitude - Méthodes multivariées

## Comparaison des méthodes multivariées avec les méthodes précédentes

Les approches multivariées telles que la régression Ridge, le PLS (Partial Least Squares) et la Random Forest se distinguent nettement des méthodes précédentes (moyenne multi-modèles, moyenne pondérée, régression linéaire simple, Kalman), tant du point de vue des hypothèses sous-jacentes que des résultats obtenus.

La régression Ridge repose sur une hypothèse de linéarité, à l'image de la régression linéaire classique, mais y ajoute une régularisation de type  $L_2$ . Cela permet de mieux contrôler les coefficients estimés en cas de colinéarité entre les variables explicatives, ce qui est fréquent dans les contextes multivariés. Cette régularisation améliore la robustesse du modèle et réduit le risque de surapprentissage, notamment lorsque le nombre de prédicteurs est élevé.

Le PLS constitue un compromis intéressant : il combine une réduction de dimension à une régression, en extrayant des composantes latentes qui maximisent la covariance entre les variables d'entrée et la variable cible. Cette approche permet de traiter efficacement la colinéarité tout en conservant une certaine lisibilité des résultats. Elle s'avère plus stable que la régression linéaire simple et plus performante que la régression Ridge dans certains cas.

La méthode Random Forest, quant à elle, ne repose sur aucune hypothèse de linéarité ou de distribution gaussienne. Elle est capable de capturer des relations complexes et non linéaires entre les variables, ce qui en fait un outil particulièrement puissant dans des contextes à forte dimensionnalité. Elle se démarque par ses performances prédictives supérieures, comme le montre la réduction du RMSE dans nos résultats, mais elle souffre d'une interprétabilité plus limitée par rapport aux approches linéaires.

Ainsi, ces méthodes multivariées permettent d'exploiter pleinement la richesse des données climatiques, tout en contrôlant les risques liés à la dimensionnalité élevée. Elles apportent une amélioration notable tant en termes de précision (réduction du RMSE) que de gestion de l'incertitude.

## Discussion sur les risques de la régression en haute dimension

Travailler dans un espace à grande dimension soulève plusieurs défis :

- **Surapprentissage (overfitting)** : plus on a de variables, plus le modèle risque de trop s'adapter aux données d'entraînement, au détriment de la généralisation.
- **Colinéarité** : les années de température sont très corrélées entre elles, ce qui rend les matrices mal conditionnées et les estimations instables.
- **Complexité computationnelle** : le coût de calcul augmente fortement avec la dimension.

## Trois solutions pour réduire la dimensionnalité

La réduction de dimension est une étape cruciale pour améliorer la stabilité des prédictions, limiter le surapprentissage et faciliter l'interprétation des résultats. Trois grandes approches complémentaires peuvent être mobilisées à cet effet :

1. **Extraction de caractéristiques : Analyse en Composantes Principales (ACP)** : L'ACP consiste à transformer un ensemble de variables corrélées en un nouveau jeu de variables non corrélées, appelées composantes principales, classées selon leur contribution à la variance totale. En ne conservant que les premières composantes, qui expliquent l'essentiel de la variance, on obtient une réduction drastique de la dimension tout en préservant l'information utile du jeu de données.
2. **Sélection de variables : filtrer les prédicteurs les plus pertinents** : Il est possible de conserver uniquement les variables les plus informatives sans transformer leur nature, grâce à des méthodes telles que :

- la sélection par importance (comme l'importance des variables dans une Random Forest),
- la sélection récursive par élimination (RFE),
- ou des critères statistiques (tests de corrélation, AIC, BIC).

Cette approche permet de supprimer les redondances tout en maintenant l'interprétabilité des variables originales.

**3. Régularisation : sélection implicite via le Lasso :** Les méthodes de régularisation, telles que le Lasso (pénalisation L1), induisent une sélection automatique des variables en contraignant certains coefficients à devenir exactement nuls. Cette approche a l'avantage de conserver les variables dans leur forme initiale, tout en assurant un contrôle strict de la complexité du modèle.

Ces stratégies de réduction de dimension permettent de construire des modèles plus robustes, mieux adaptés aux petits jeux de données, et moins sensibles au bruit statistique.

### Application de l'ACP (Analyse en Composantes Principales)

Pour surmonter certaines limites des approches multivariées, nous avons recours à l'Analyse en Composantes Principales (ACP). Cette méthode présente plusieurs avantages clés dans le cadre de notre étude.

Tout d'abord, l'ACP permet une réduction efficace de la dimensionnalité : elle projette les données d'origine dans un espace de dimension plus faible, tout en conservant la majeure partie de la variance initiale. Cela permet de simplifier les modèles sans perte significative d'information.

Ensuite, les composantes principales résultant de l'ACP sont orthogonales, ce qui implique une absence de corrélation entre elles. Cette propriété améliore la validité des hypothèses sous-jacentes aux modèles statistiques classiques, notamment dans le cadre des régressions.

Enfin, en ne retenant que les composantes qui expliquent l'essentiel de la variance (par exemple plus de 75%, 80% ou 95%), on élimine une partie du bruit présent dans les données. Cette sélection contribue à renforcer la stabilité et la robustesse des prédictions issues des modèles construits sur ces nouvelles bases.

## Corrélation entre les prédicteurs

Avant d'appliquer l'ACP, il est important d'examiner la structure de corrélation des prédicteurs multivariés. Dans notre cas, les températures annuelles successives (entre 1990 et 2015) sont fortement corrélées entre elles comme on peut le voir grâce à la matrice de corrélation ci-dessous :

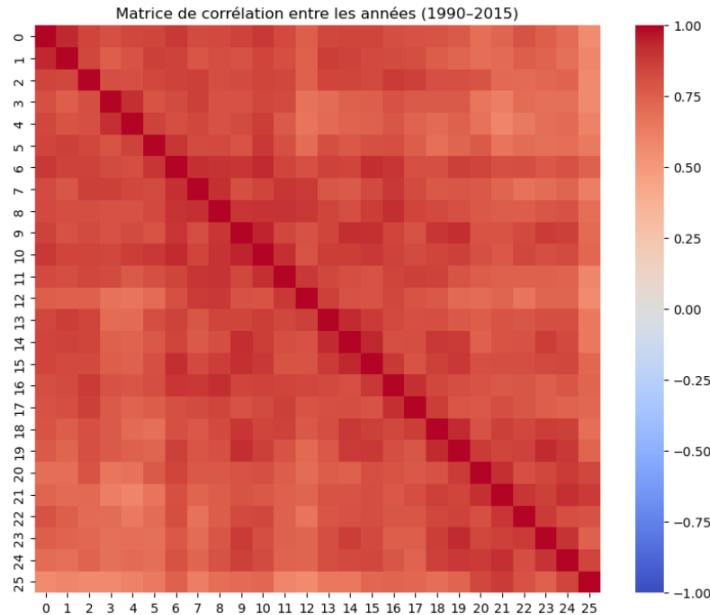


FIGURE 15 – Matrice de corrélation entre les années 1990-2015

Cela s'explique par la tendance climatique globale, mais également par des inerties naturelles dans les séries temporelles.

La matrice de corrélation calculée sur  $X_{simumulti}$  révèle des coefficients souvent supérieurs à 0.9 entre années proches. Ce phénomène de multicolinéarité pose problème dans les modèles de régression classiques, car il rend l'estimation des coefficients instable et amplifie la variance des prédictions.

Ainsi, l'ACP apparaît comme une solution naturelle pour :

- Résumer l'information redondante entre variables corrélées,
- Travailler dans un espace orthogonal plus adapté à la régression linéaire,
- Réduire le risque de surapprentissage.

## Choix du nombre de composantes principales

Le choix du nombre de composantes principales à retenir repose sur l'analyse de la variance expliquée cumulée par l'ACP. En fixant un seuil de 95% de variance expliquée, nous garantissons que l'essentiel de l'information contenue dans les données d'origine est conservé. Dans notre cas, ce seuil est atteint avec les 7 premières composantes principales. Ces dernières permettent donc de représenter fidèlement les prédicteurs tout en réduisant significativement la dimension du problème, ce qui contribue à la stabilité et à la performance des modèles utilisés par la suite.

## Résultats après réduction de dimension

Après réduction à 7 composantes principales, la régression multivariée combinée à une validation LOO donne :

- **RMSE** = 1.231
- **Incertitude sur la prédition** = 0.805

Le RMSE est considérablement réduit par rapport à l'approche multivariée brute, l'incertitude est réaliste (nettement supérieure à zéro). L'ACP permet donc d'augmenter la robustesse et la généralisation du modèle.

## Réduction à une seule composante principale

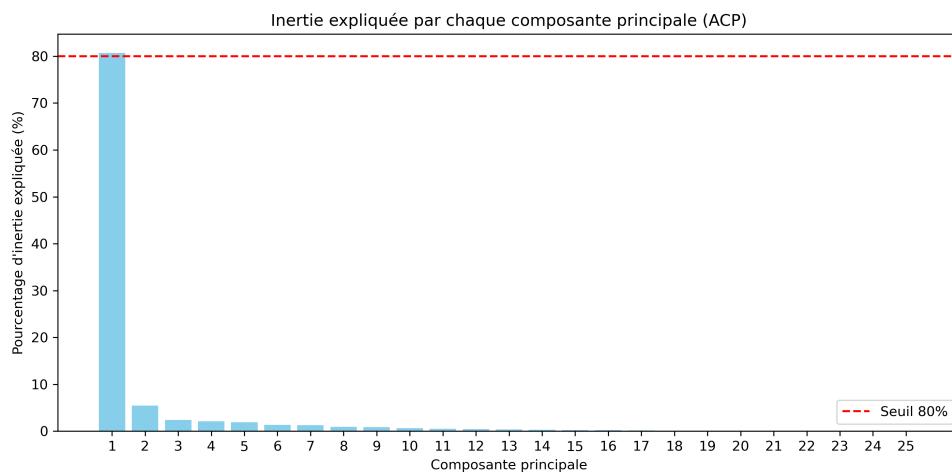


FIGURE 16 – Inertie expliquée par chaque composante principale (ACP)

En observant l'inertie expliquée par chaque composante, il apparaît que la première composante principale seule explique plus de 80% de la variance totale.

Dans un souci de simplification et pour éviter d'introduire du bruit inutile, nous décidons de ne conserver qu'une seule composante.

**Application :** Nous ajustons une nouvelle ACP avec  $k = 1$  composante principale :

- $X_{\text{simu\_1}}$  : Projection des simulations sur la première composante principale,
- $X_{\text{obs\_1}}$  : Projection des observations sur cette même composante.

## Résultats avec 1 composante principale

Après réduction à une seule composante, nous recalculons les performances de la régression multivariée avec validation Leave-One-Out :

- **RMSE** = 1.040
- **Incertitude sur la prédition** = 0.973

**Commentaires :** Le recours à une seule composante principale permet d'atteindre un RMSE proche du minimum observé, tout en conservant une structure de modèle extrêmement simple. Cette faible complexité constitue un avantage notable, notamment dans un contexte de données limitées, en réduisant les risques de surapprentissage.

Par ailleurs, l'incertitude associée aux prédictions reste à un niveau réaliste, comparable à celui obtenu avec des projections sur sept composantes principales. Cela indique que l'essentiel de l'information pertinente pour prédire la variable cible  $\hat{Y}$  est efficacement concentré dans la première composante principale.

Ces résultats soulignent la puissance de l'ACP dans ce contexte, en montrant qu'une réduction de dimension drastique peut être réalisée sans perte significative de performance.

### Comparaison des différentes méthodes avec 1 composante principale

Nous comparons ensuite les performances de quatre approches différentes en utilisant uniquement la première composante principale :

- Moyenne multi-modèle,
- Moyenne pondérée,
- Régression linéaire,
- Filtrage de Kalman one-step.

#### Résultats Leave-One-Out (RMSE) :

- **Moyenne multi-modèle** : RMSE = 1.079
- **Moyenne pondérée** : RMSE = 1.017
- **Régression linéaire** : RMSE = 1.040
- **Kalman one-step** : RMSE = 1.035

**Commentaires sur le RMSE :** Parmi les différentes méthodes évaluées, la moyenne pondérée s'impose comme la plus performante en termes d'erreur quadratique moyenne (RMSE). Elle parvient à combiner efficacement les informations disponibles tout en minimisant l'erreur de prédiction.

Kalman et la régression linéaire affichent des performances très proches, légèrement inférieures mais néanmoins compétitives. Ces deux approches bénéficient de mécanismes d'ajustement dynamiques ou structurels qui leur permettent de bien s'adapter aux données.

Enfin, bien que plus simple, la moyenne multi-modèle naïve reste relativement performante. Son léger retrait par rapport aux autres méthodes montre qu'elle constitue une base de référence robuste, particulièrement utile dans des contextes où la simplicité et l'interprétabilité sont prioritaires.

### Incertitudes associées aux prédictions

Nous estimons également l'incertitude de prédiction pour chaque méthode :

- **Moyenne multi-modèle** :  $\hat{\sigma} = 1.036$
- **Moyenne pondérée** :  $\hat{\sigma} = 0.712$
- **Régression linéaire** :  $\hat{\sigma} = 0.973$
- **Kalman one-step** :  $\hat{\sigma} = 0.952$

**Commentaires sur les incertitudes :** L'analyse des incertitudes montre une réduction significative lorsque l'on utilise la moyenne pondérée.

La régression linéaire et Kalman offrent eux aussi des gains notables par rapport à la moyenne simple, traduisant une meilleure capacité d'adaptation aux données observées. Ces approches réduisent l'ampleur des incertitudes tout en conservant une bonne stabilité.

Globalement, cette diminution de l'incertitude reflète une meilleure adéquation aux structures présentes dans les données, ainsi qu'un contrôle plus efficace du risque de sur-apprentissage.

## Prédictions finales pour l'anomalie de température future

Enfin, les prévisions  $\hat{Y}$  associées à chaque méthode sont :

- **Moyenne multi-modèle** :  $\hat{Y} = 5.221$
- **Moyenne pondérée** :  $\hat{Y} = 5.386$
- **Régression linéaire** :  $\hat{Y} = 5.354$
- **Kalman one-step** :  $\hat{Y} = 5.375$

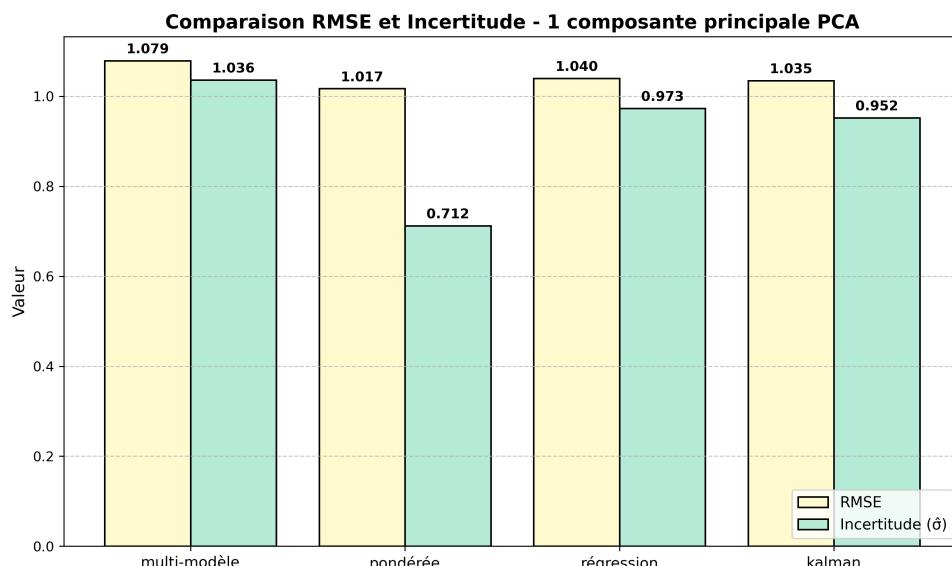


FIGURE 17 – Comparaison du RMSE et de l'incertitude - 1 compostante principale ACP

**Analyse finale :** L'ensemble des méthodes testées aboutit à des prévisions relativement proches, avec une anomalie de température estimée entre 5.2 et 5.4 degrés. Cette convergence générale témoigne de la cohérence des approches, malgré leur complexité variable.

Les méthodes plus élaborées, telles que la moyenne pondérée, la régression linéaire et Kalman, tendent à légèrement augmenter l'estimation par rapport à la moyenne simple.

Enfin, il est intéressant de noter que la prédiction fournie par Kalman est très proche de celle obtenue par la moyenne pondérée, suggérant une certaine robustesse de cette estimation, quel que soit le formalisme adopté.

## Résumé général

Dans un contexte de données climatiques multivariées, combiné à un nombre limité d'observations, la réduction de dimension via l'Analyse en Composantes Principales (ACP)

s'avère indispensable. Elle permet de limiter le surapprentissage tout en conservant l'essentiel de l'information. Dans notre cas, une seule composante principale suffit généralement à capturer la majeure partie de la dynamique climatique observée, ce qui simplifie l'analyse sans perte significative de précision.

Par ailleurs, les méthodes avancées telles que la moyenne pondérée, la régression linéaire et Kalman présentent des performances supérieures à la moyenne simple, aussi bien en termes de précision que de réduction de l'incertitude associée aux prédictions. Ces approches tiennent compte de l'information structurelle ou dynamique entre les modèles, ce qui améliore leur capacité à généraliser.

Dans notre configuration expérimentale, la moyenne pondérée apparaît comme le meilleur compromis entre robustesse et précision. Elle combine efficacement les avantages des différentes sources d'information tout en conservant une simplicité d'implémentation et une stabilité face aux variations des données.

## 4. Conclusion

Dans ce projet, nous avons étudié différentes méthodes d'agrégation et de prédiction sur un ensemble de modèles climatiques simulés, dans le but d'estimer une anomalie de température future à partir d'observations partielles du passé.

Notre travail a débuté par la mise en place de méthodes univariées classiques, utilisant des informations simples extraites des séries temporelles passées, comme la moyenne ou la tendance linéaire. Nous avons ensuite évalué la performance de plusieurs techniques prédictives :

- La moyenne multi-modèle simple,
- Une moyenne pondérée selon la proximité au modèle observé,
- La régression linéaire,
- Un filtrage de Kalman one-step.

L'évaluation rigoureuse des performances via validation croisée Leave-One-Out a permis de mettre en évidence plusieurs points fondamentaux :

- Avec peu d'échantillons disponibles, il est essentiel d'utiliser une validation fine pour éviter le surapprentissage.
- La moyenne multi-modèle, bien que simple, fournit un point de départ robuste mais améliorable.
- Les approches pondérées ou basées sur la régression permettent une réduction de l'erreur quadratique moyenne et de l'incertitude associée.

Nous avons ensuite étendu notre analyse vers un cadre multivarié, en considérant simultanément plusieurs années de températures passées comme prédicteurs. Cette richesse accrue d'information a permis d'envisager des méthodes d'apprentissage statistique plus avancées, telles que la régression Ridge, la régression par moindres carrés partiels (PLS) et les forêts aléatoires (Random Forest). Ces techniques ont été choisies pour leur capacité à gérer des problèmes de colinéarité, de forte dimension, ou encore de non-linéarité. Ridge apporte une régularisation efficace dans les modèles linéaires en pénalisant les coefficients trop instables, PLS extrait des composantes latentes corrélées à la variable cible tout en réduisant la dimension, et Random Forest offre une approche non paramétrique robuste, capable de capturer des interactions complexes entre variables. Ces méthodes ont

ainsi permis d'améliorer les performances de prédiction par rapport aux approches plus simples, en particulier dans les situations où la relation entre variables était moins linéaire ou fortement bruitée.

Toutefois, l'augmentation du nombre de prédicteurs s'est accompagnée d'un risque accru de surapprentissage, notamment en raison du faible nombre d'échantillons disponibles. Pour maîtriser cette complexité, nous avons eu recours à une réduction de dimension par Analyse en Composantes Principales (ACP).

Grâce à l'ACP, nous avons montré qu'une seule composante principale suffisait dans notre cas pour conserver plus de 80% de la variance totale, ce qui a conduit à une forte amélioration des performances prédictives. L'utilisation d'une seule variable issue de l'ACP a permis d'obtenir des résultats comparables, voire supérieurs, à ceux obtenus en univarié classique, tout en bénéficiant d'une incertitude réduite.

Finalement, la comparaison finale des méthodes sur cette composante principale a confirmé que les approches pondérées, la régression linéaire, Kalman, ainsi que les techniques avancées comme Ridge, PLS et Random Forest permettaient une meilleure précision que la moyenne multi-modèle, tout en offrant des estimations d'incertitudes réalistes et plus resserrées.

Pour finir, ce projet met en lumière plusieurs enseignements clés pour la modélisation climatique, comme l'importance cruciale de la validation croisée pour éviter les biais d'optimisme. Nous avons pu voir la nécessité d'adapter la complexité des modèles au volume d'information réellement disponible. Il met en avant l'apport déterminant de méthodes de réduction de dimension comme l'ACP en contexte multivarié. Il souligne aussi l'intérêt de combiner plusieurs méthodes pour obtenir des prédictions à la fois précises et robustes.

Pour ouvrir le sujet, plusieurs pistes pourraient prolonger ce travail. Nous pourrions par exemple introduire de nouvelles variables prédictives, telles que la concentration en gaz à effet de serre ou la couverture nuageuse, pour enrichir l'information disponible. Il est également possible d'étendre l'approche à des horizons temporels plus longs, pour anticiper l'évolution climatique au-delà du XXI<sup>e</sup> siècle.

Ainsi, ce projet constitue une base solide pour aller plus loin dans la prédiction climatique fiable, fondée sur des méthodes statistiques rigoureuses et une exploitation optimale des données disponibles.