Université de Bordeaux

Modèles climatiques et observations : vers plus de précision...

Félicia DOSSOU
Arthur LE CAMUS

2024/2025

Résumé — Ce document porte sur la réduction des incertitudes dans les projections climatiques futures à partir de méthodes statistiques appliquées à des simulations CMIP6. Il repose sur un jeu de données observées (1850–2021) et simulées (2022–2099), et mobilise plusieurs approches statistiques, dont la moyenne d'ensemble des modèles climatiques, une pondération naïve fondée sur la distance aux observations passées, une régression linéaire émergente reliant climat passé et futur, ainsi qu'un filtre de Kalman formulé dans un cadre probabiliste gaussien.

L'objectif est d'identifier les méthodes les plus robustes pour contraindre les projections futures à l'aide des données du passé.

L'ensemble des annexes est fourni dans les fichiers accompagnant ce rapport. Chaque méthode est détaillée dans un fichier distinct nommé `methode_x.ipynb`, où `x` correspond au numéro de la méthode utilisée. La partie dédiée aux intervalles de confiance ainsi qu'à l'analyse descriptive est rassemblée dans un fichier spécifique intitulé `sujet_1.ipynb`.

Table des matières

1 Introduction

Dans le contexte du changement climatique, il est essentiel de mieux cerner les projections futures de température afin de guider les politiques publiques et les stratégies d'adaptation. Cependant, une source importante d'incertitude réside dans la diversité des modèles climatiques utilisés (ensemble CMIP6 [1]), chacun reposant sur ses propres hypothèses et paramètres.

Ce projet vise à construire un modèle statistique permettant d'estimer au mieux les projections climatiques futures à partir d'une contrainte observationnelle. Pour cela, nous exploitons les liens statistiques entre les observations du climat passé et les simulations issues de modèles numériques. L'objectif est de combiner ces simulations et de les pondérer en fonction de leur cohérence avec les observations, afin de réduire l'incertitude associée aux projections futures.

Afin d'établir une relation entre une variable passée X et une variable future Y, afin d'estimer la distribution $p(Y|X_0)$, nous allons implémenter et comparer quatre méthodes de réduction d'incertitude afin d'évaluer la robustesse des résultats obtenus pour les températures futures. Nous discuterons ensuite des limites de ces approches en contexte climatique.

Notation

On appellera Y la variable à prédire et X la variable utilisée pour prédire (prédicteur).

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

avec $i \in [\![1..M]\!]$, $M = 25$ le nombre de modèles climatiques.

2 Description des données

Les données utilisées dans ce projet sont fournies sous forme de fichiers `.npy`, organisés selon le format suivant :

- **matrix_1.npy** (25 x 78) : températures futures simulées par 25 modèles climatiques pour la période 2022–2099.
- **matrix_2.npy** (1 x 172) : observations passées moyennées globalement sur la période 1850–2021.
- **matrix_3.npy** (25 x 172) : températures passées simulées par les mêmes 25 modèles sur la période 1850–2021.
- **matrix_4.npy** (172 x 1) : incertitudes (erreur standard) associées aux observations passées.

Les séries temporelles sont accompagnées de trois fichiers de labels : **label_models.npy** contient les noms des modèles, **label_past_times.npy** les années de la période historique (1850–2021), et **label_future_times.npy** celles de la période future (2022–2099). Cette structure permet d'associer chaque simulation à un modèle et à une année précise.

Dans un premier temps, pour chaque année entre 1850 et 2021, interessons nous à la moyenne, au minimum et au maximum des températures simulées par les 25 modèles climatiques. On peut les comparer aux observations, accompagnées de leur incertitude fournie.

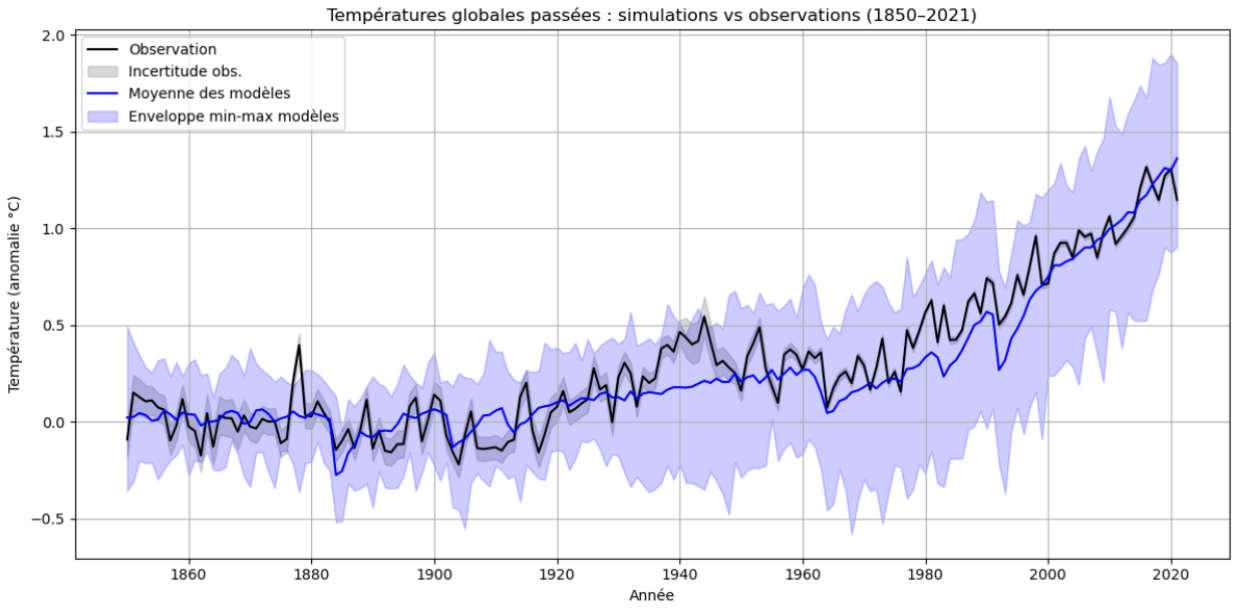

FIGURE 1 – Visualisation des températures simulées et observées (1850–2021)

Le graphique montre que les anomalies de température simulées par les différents modèles climatiques sont globalement cohérentes avec les observations, en particulier à partir du milieu du XXe siècle. Les incertitudes, bien que présentes, n'altèrent pas la tendance nette au réchauffement observée dans toutes les courbes.

3 Méthode 0 : Moyenne multi-modèles

Cette première méthode repose sur l'utilisation directe des projections climatiques fournies par les 25 modèles CMIP6, sans pondération ni ajustement. Pour chaque année future t entre 2022 et 2099, la température estimée est calculée comme la moyenne arithmétique des modèles :

$$\hat{Y}(t) = \frac{1}{M} \sum_{i=1}^{M} Y_i(t)$$

où $Y_i(t)$ est la projection du modèle i et $M = 25$ le nombre total de modèles. L'incertitude associée est quantifiée par l'écart-type intermodèle ($\pm\sigma_Y(t)$) ainsi que par l'enveloppe définie par les valeurs minimale et maximale des simulations. Cette méthode simple sert de référence pour évaluer les gains apportés par les approches statistiques plus élaborées.

3.1 Résultats

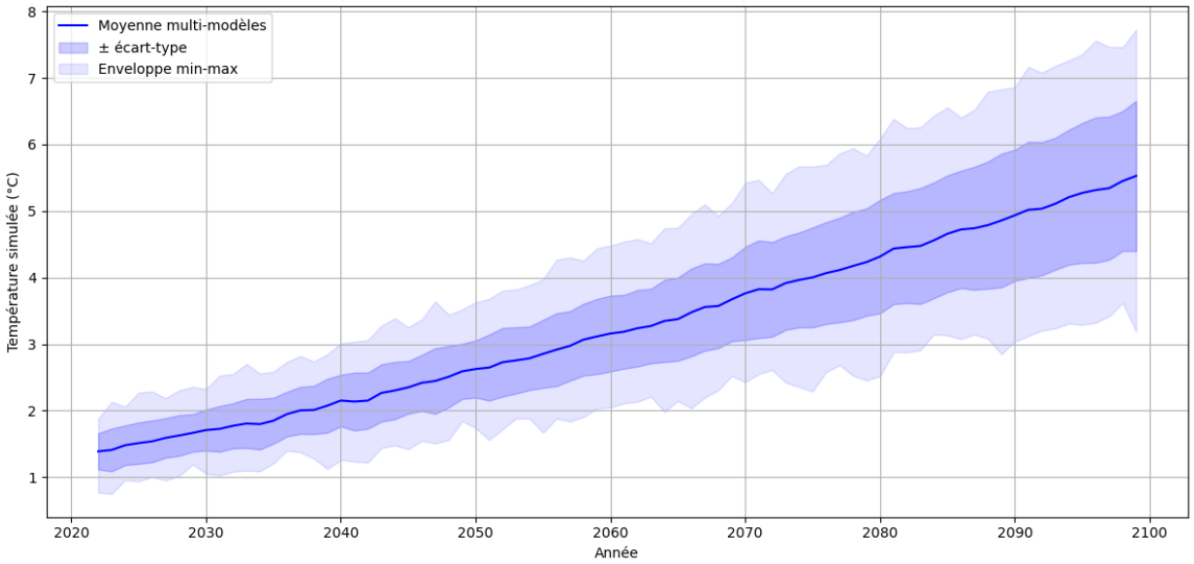

FIGURE 2 – Projections climatiques futures (Méthode multi-modèles)

La méthode 0 met en évidence une hausse continue de la température moyenne simulée jusqu'à la fin du XXI^{e} siècle, atteignant environ 5,5 °C d'anomalie en 2099. Cette trajectoire reflète le forçage climatique imposé dans les modèles, en lien avec l'accumulation des gaz à effet de serre (CO_2, CH_4, vapeur d'eau) et des rétroactions positives telles que la réduction de l'albédo ou l'augmentation de l'humidité atmosphérique.

L'incertitude intermodèle augmente également avec le temps : l'écart-type passe de moins de 0,5 °C au début de la période à plus de 1 °C en 2100, et l'enveloppe min–max dépasse 2 °C d'écart entre modèles extrêmes. Cette dispersion reflète les divergences structurelles entre modèles climatiques (paramétrisation des nuages, sensibilité climatique, résolution), ce qui justifie le recours à des méthodes plus robustes de pondération ou de contrainte.

Sur la période passée (1850–2021), on observe une cohérence générale entre les simulations et les observations, avec une tendance nette au réchauffement dès le milieu du XX^{e} siècle. Toutefois, certaines

différences apparaissent : les observations présentent parfois une variabilité plus marquée que la moyenne des modèles, en particulier autour des années 1940 et 1970, ce qui peut s'expliquer par la superposition de plusieurs facteurs exogènes abordés dans le cours.

Les variations interannuelles résultent à la fois de la variabilité interne du climat (oscillations naturelles comme El Niño), et de forçages externes naturels, tels que l'activité solaire ou les éruptions volcaniques comme Krakatoa en 1883, ou Pinatubo en 1991. Ces événements ponctuels peuvent temporairement refroidir ou réchauffer l'atmosphère et expliquer certaines divergences ponctuelles entre simulations et observations.

Enfin, la dispersion entre modèles s'accroît avec le temps, illustrant l'incertitude croissante dans les projections en raison des différences dans les formulations physiques et les paramétrisations propres à chaque modèle. Cette variabilité intermodèle justifie l'usage de méthodes statistiques pour mieux contraindre les projections futures, ce qui constitue l'objectif principal du projet.

4 Méthode 1 : Moyenne pondérée

Cette méthode propose d'améliorer la moyenne muti-modèles (méthode 0) en pondérant les modèles selon leur capacité à reproduire fidèlement le climat observé sur la période passée. L'idée centrale est que les modèles les plus proches des observations historiques sont jugés plus fiables pour projeter l'évolution future du climat. Il s'agit d'une approche simple mais plus informative que la moyenne multi-modèles, car elle introduit une notion de performance passée dans l'estimation.

À partir de cette section, toutes les méthodes suivantes sont restreintes à la période récente 1992–2021, soit les 30 dernières années disponibles.

4.1 Implémentation de la moyenne pondérée

On définit un poids w_i pour chaque modèle $i \in \{1, \dots, M\}$, qui dépend de deux critères :
— la performance du modèle, mesurée par sa proximité à l'observation réelle X_0,
— son indépendance vis-à-vis des autres modèles de la base.

La pondération est donnée par :

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

où :
— $D_i = \|X_i - X_0\|$ est la distance entre les données simulées passées du modèle i et les observations réelles X_0,
— $S_{i,j} = \|X_i - X_j\|$ est la distance entre les modèles i et j,
— σ_D écart-type des distances entre chaque modèle et l'observation, et σ_S écart-type des distances entre tous les couples de modèles (hyperparamètres).

La projection pondérée est alors calculée pour chaque année t selon :

$$\hat{Y}(t) = \sum_{i=1}^{M} w_i Y_i(t)$$

et l'incertitude associée est donnée par la variance pondérée :

$$\sigma_Y^2(t) = \sum_{i=1}^{M} w_i \left(Y_i(t) - \hat{Y}(t)\right)^2$$

4.2 Etude de l'impact des paramètres sur la projection et son incertitue.

Les poids w_i attribués aux modèles dans la méthode de Brunner dépendent de deux hyperparamètres : σ_D, qui régule l'importance de la proximité aux observations passées (critère de performance), et σ_S, qui contrôle l'effet d'indépendance entre modèles.

Un σ_D faible favorise fortement les modèles proches des observations (X_0), tandis qu'un σ_D élevé atténue les différences de performance et tend vers une pondération uniforme. De même, un σ_S faible

pénalise les modèles redondants en privilégiant ceux situés dans des zones isolées de l'espace des simulations, alors qu'un σ_S grand neutralise ce critère d'indépendance.

Ainsi :
— Fixer $\sigma_S \to \infty$ revient à ne considérer que la performance historique (pondération selon D_i uniquement).
— Fixer $\sigma_D \to \infty$ revient à pondérer uniquement selon l'indépendance structurelle.

Nous commençons par une configuration de référence avec $\sigma_D = 1.0$ et $\sigma_S = 1.0$, avant d'optimiser ces paramètres par validation croisée.

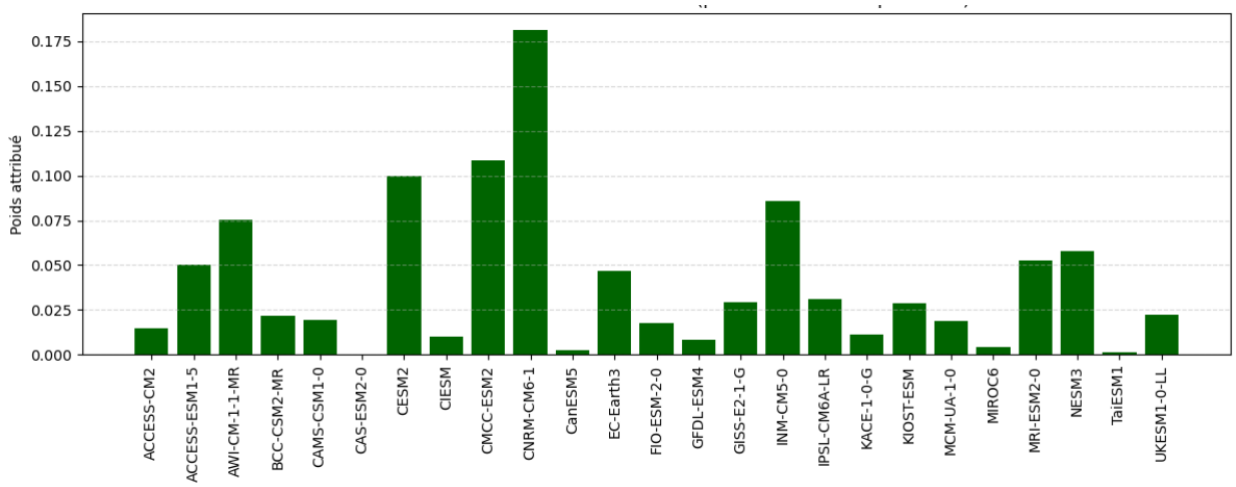

FIGURE 3 – Poids attribués aux modèles climatiques

Cette pondération met en évidence une forte sélection sur quelques modèles, suggérant qu'ils sont à la fois proches des observations et relativement indépendants des autres.

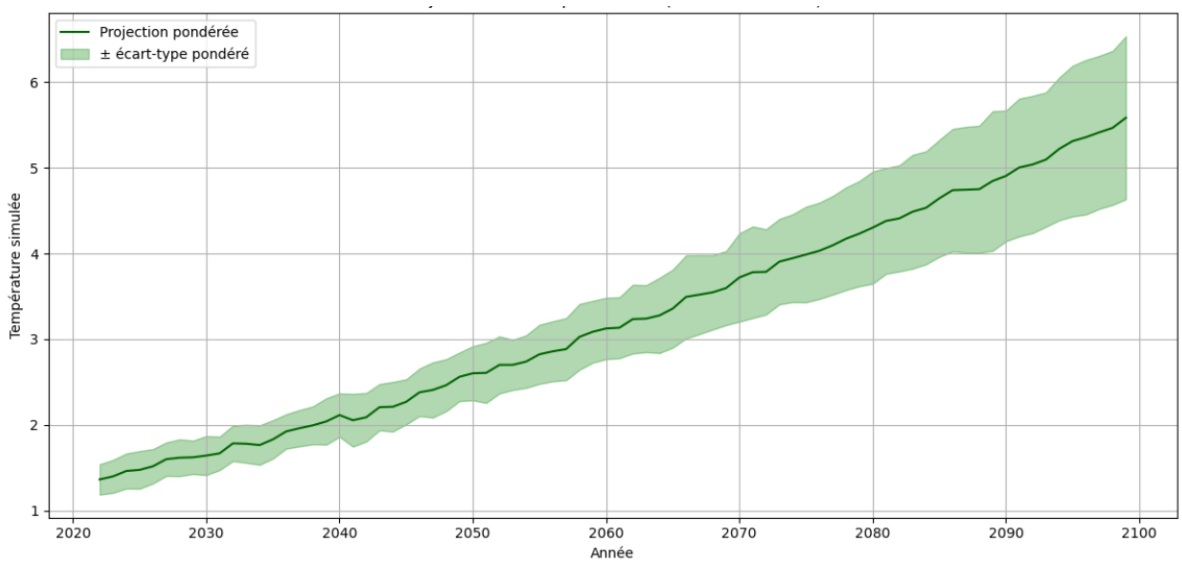

FIGURE 4 – Projections climatiques futures avec méthode de pondération naïve

La méthode de pondération naïve aboutit à une trajectoire future cohérente avec celle obtenue par la méthode 0, avec une température atteignant environ 5.5 °C d'anomalie en 2099.

4.3 Comparaison de l'incertitude avec celle obtenue avec la moyenne multi-modèles

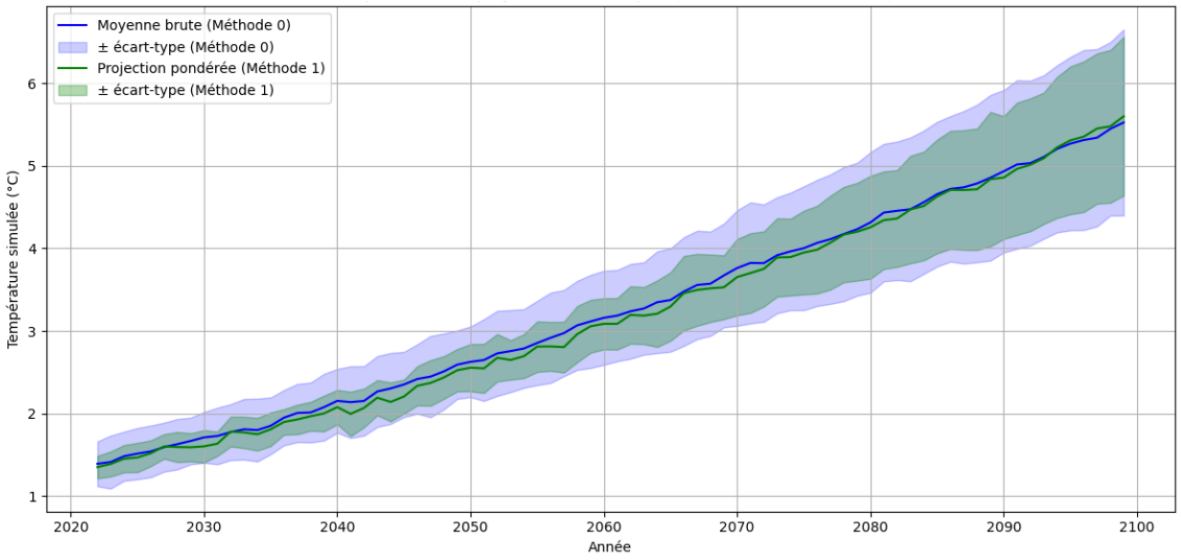

FIGURE 5 – Comparaison des projections climatiques (méthode multi-modèles vs Méthode pondérée)

L'incertitude est légèrement réduite, comme le montre l'enveloppe d'écart-type plus resserrée.

Cependant, les poids attribués aux différents modèles sont tous très proches (autour de 4 %), ce qui suggère que leur performance sur la période passée est globalement similaire. La distance quadratique utilisée pour évaluer la proximité aux observations ne parvient pas à fortement différencier les modèles dans ce cas.

Le modèle initial, utilisant une pondération non optimisée ($\sigma_D = \sigma_S = 1.0$), présente une erreur de projection de RMSE = 0.6091. L'objectif est désormais d'améliorer cette performance en optimisant les hyperparamètres via validation croisée.

4.4 Réglage des paramètres par validation croisée

Afin d'améliorer la précision des projections fournies par la méthode de pondération proposée par Brunner, nous avons exploré l'influence des deux hyperparamètres σ_D et σ_S.

Une validation croisée Leave-One-Out (LOO) a été utilisée pour calculer l'erreur de prédiction (RMSE) associée à chaque couple de valeurs testées, et ainsi sélectionner les combinaisons optimales. Deux visualisations séparées ont permis d'analyser l'impact de chaque paramètre individuellement sur la qualité de la prédiction.

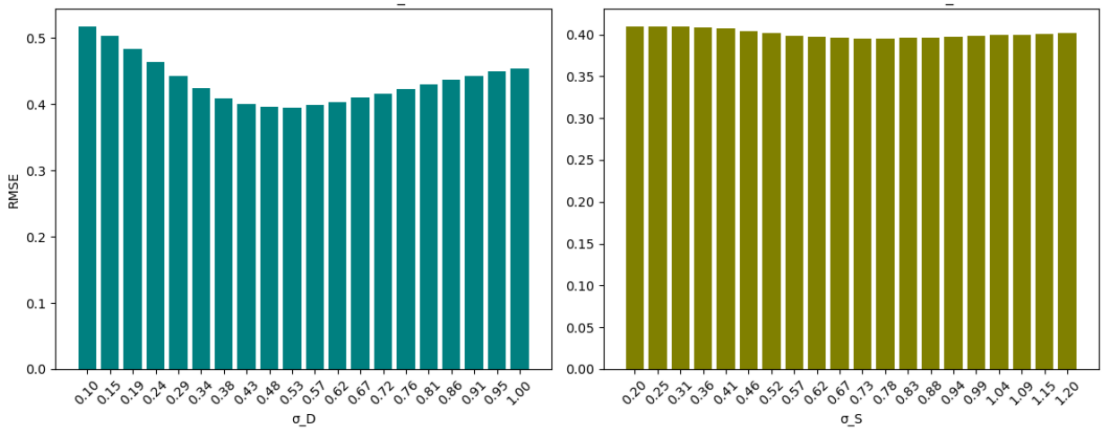

FIGURE 6 – Evolution du RMSE selon σ_D et σ_S

Après ajustement progressif des intervalles de recherche, les résultats montrent que la valeur optimale des hyperparamètres est atteinte pour $\sigma_D = 0.5263$ et $\sigma_S = 0.7263$, avec un RMSE minimal de 0.3955.

L'évolution du RMSE selon σ_D met en évidence une diminution marquée de l'erreur jusqu'à un plateau autour de cette valeur, indiquant que la prise en compte fine de la performance historique des modèles est essentielle pour améliorer la projection. En revanche, la courbe du RMSE selon σ_S reste relativement plate, suggérant que le critère d'indépendance structurelle entre modèles a un impact limité dans notre configuration. Cela peut s'expliquer par une faible diversité structurelle des modèles CMIP6 ou par une forte redondance dans les simulations passées.

L'optimisation conjointe de ces paramètres permet donc une amélioration notable par rapport à la version naïve.

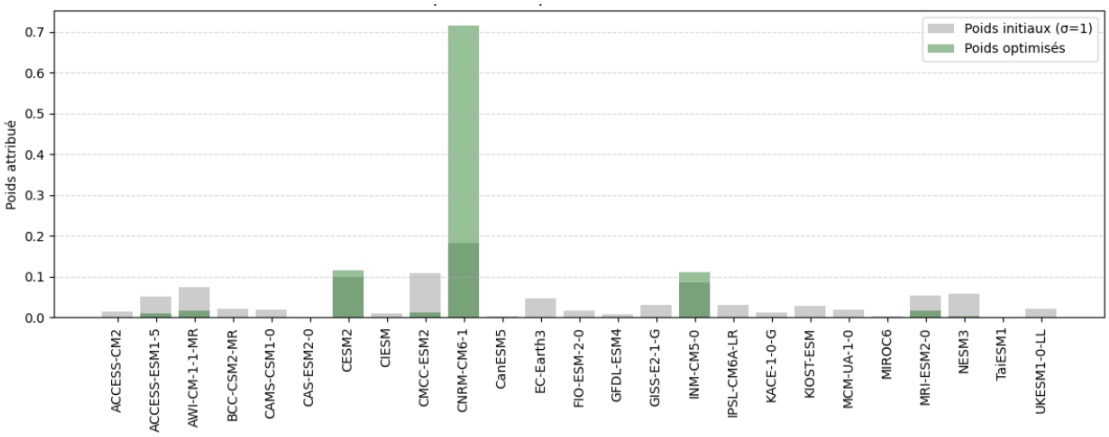

FIGURE 7 – Comparaison des poids attribués aux modèles

Avec $\sigma_D = 0.5263$ et $\sigma_S = 0.7263$, les modèles CNRM-CM6-1, CESM2 et INM-CM5-0 reçoivent une plus forte pondération.

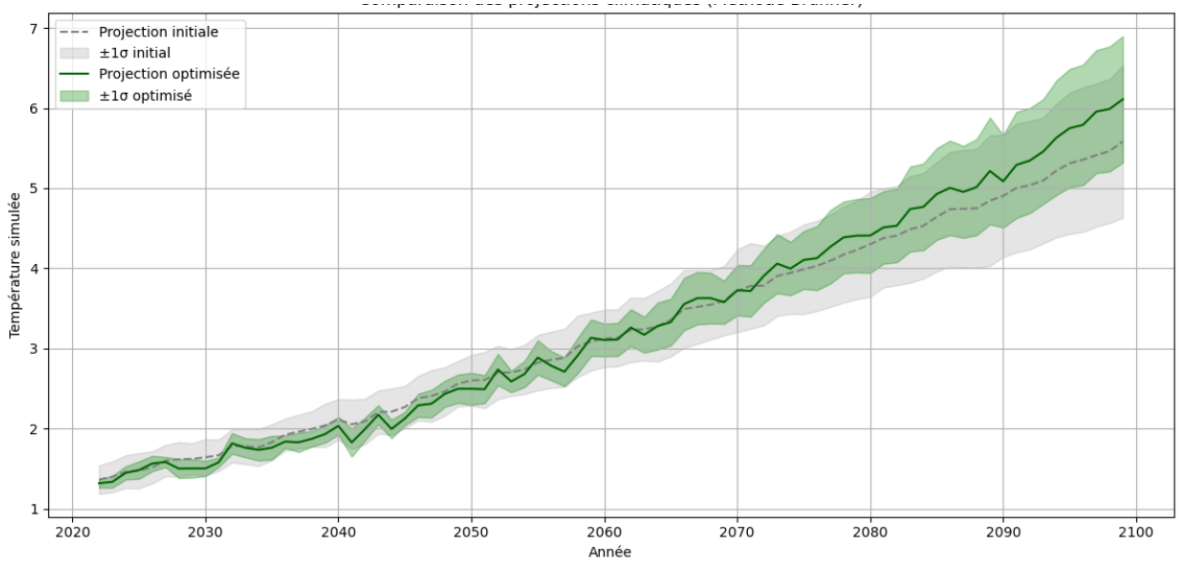

FIGURE 8 – Comparaison des projections climatiques (méthode pondérée initiale vs méthode pondérée optimisée)

La méthode de pondération optimisée améliore significativement la qualité des projections climatiques en réduisant l'incertitude autour des prévisions futures, comme en témoigne la bande $\pm\sigma$ plus étroite. La trajectoire obtenue est également plus réaliste, traduisant un réchauffement cohérent avec les observations passées.

Enfin, cette approche favorise une meilleure cohérence inter-modèles, en atténuant l'influence des simulations redondantes et en valorisant les modèles fiables et distinctifs.

4.5 Comparaison des performances selon la période du prédicteur X

Nous avons testé la méthode de pondération optimisée, en maintenant les valeurs de $\sigma_D = 0.5263$ et $\sigma_S = 0.7263$ constantes, tout en faisant varier la période de calcul du prédicteur X. Pour chaque période, nous avons calculé le RMSE entre la moyenne pondérée des modèles projetés et la moyenne correspondante des données simulées de référence.

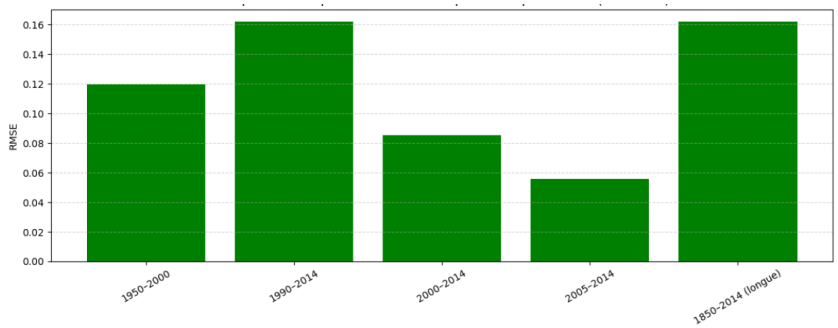

FIGURE 9 – Comparaison des performances selon la période du prédicteur X

Les résultats montrent que les périodes récentes (1990–2014 ou 2005–2014) donnent de meilleures performances, alors que les périodes anciennes (1850–1950) ou trop longues (1850–2014) dégradent la qualité de la prédiction.

4.6 Extension en multivarié

L'approche multivariée repose sur l'utilisation de toutes les années passées (ici les 30 dernières années : 1992–2021) comme prédicteurs. Chaque vecteur $X_i \in \mathbb{R}^{30}$ représente la série temporelle des températures simulées passées d'un modèle.

Afin de réduire la dimension et d'éviter les problèmes liés à la redondance ou à la colinéarité entre années (forte corrélation temporelle), nous appliquons une Analyse en Composantes Principales (ACP) sur la matrice $X \in \mathbb{R}^{25\times 30}$.
Cette technique permet de projeter les données dans un sous-espace de plus faible dimension tout en conservant l'essentiel de l'information. Le résultat montre que les cinq premières composantes suffisent à expliquer plus de 90% de la variance, ce qui justifie leur sélection.

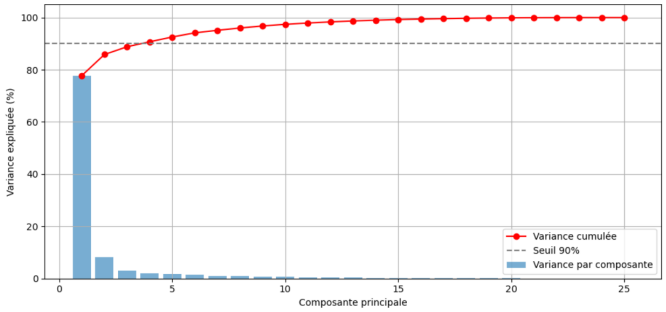

FIGURE 10 – Variance expliquée par l'ACP

Les vecteurs simulés X_i sont alors projetés dans ce nouvel espace réduit, tout comme l'observation réelle X_0. Les distances de performance D_i et d'indépendance $S_{i,j}$ sont recalculées dans cet espace, puis injectées dans la formule de pondération optimisée déjà calibrée avec les meilleurs σ_D et σ_S trouvés précédemment.

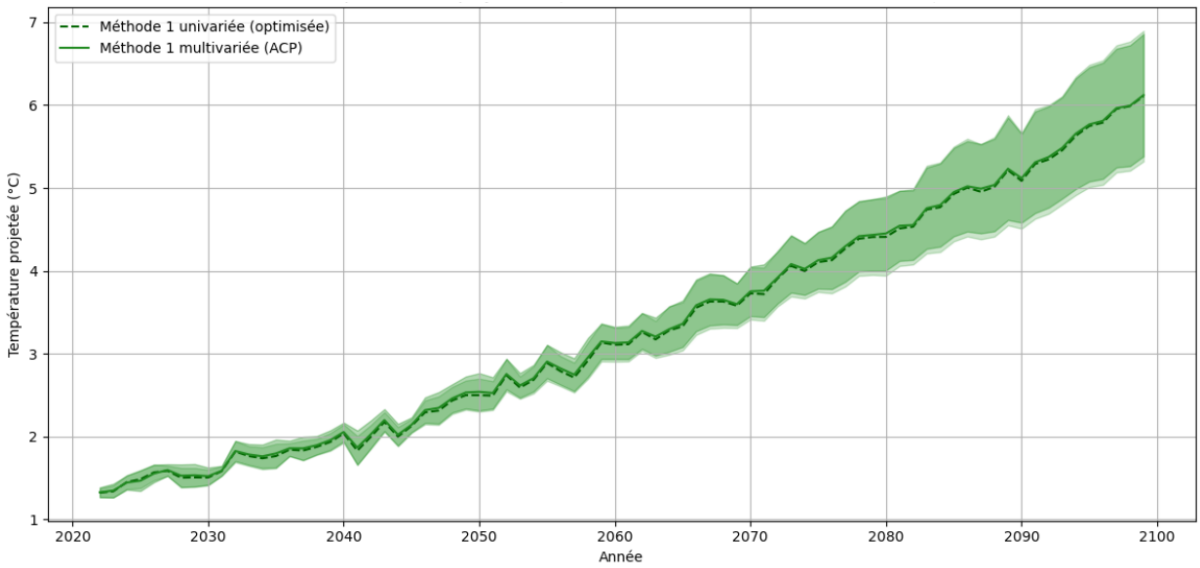

FIGURE 11 – Comparaison des projections (Méthode 1 : univariée vs multivariée)

Le RMSE obtenu avec la version multivariée est de 0.2121, très proche de celui de la version univariée (0.2089), ce qui montre que la réduction de dimension n'a pas significativement dégradé les performances. La projection multivariée apparaît légèrement plus lissée, avec une trajectoire temporelle plus régulière. L'incertitude est aussi légèrement élargie, reflétant une pondération plus diffuse dans l'espace réduit. Cette approche permet de capturer certaines structures globales entre modèles sans compromettre la qualité des prévisions, tout en montrant que l'information utile est concentrée dans un petit nombre de composantes principales ce qui est cohérent avec la forte redondance temporelle des données climatiques.

5 Méthode 2 : Régression linéaire

Cette méthode exploite une relation empirique observée entre une variable passée X et une variable future Y simulée par les modèles climatiques. L'objectif est d'estimer la température future Y_0 à partir de l'observation X_0, en s'appuyant sur une régression linéaire entre les paires (X_i, Y_i) issues des modèles CMIP6.

5.1 Implémentation de la regression linéaire

On suppose que la relation entre la variable simulée passée X_i et la projection future Y_i suit un modèle linéaire :

$$Y_i = aX_i + b + \varepsilon_i$$

où (a, b) sont estimés par moindres carrés à partir des paires (X_i, Y_i). On utilise ensuite la valeur observée X_0 pour prédire :

$$\hat{Y_0} = aX_0 + b$$

L'incertitude associée est estimée à partir de la variance résiduelle du modèle.

Nous entraînons ici un modèle de régression linéaire simple reliant la température moyenne passée simulée par les modèles climatiques (X_i) à leur projection moyenne future (Y_i). L'objectif est de prédire la température future Y_0 à partir de l'observation réelle du climat passé X_0.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.348
Model:	OLS	Adj. R-squared:	0.320
Method:	Least Squares	F-statistic:	12.30
Date:	Thu, 24 Apr 2025	Prob (F-statistic):	0.00190
Time:	02:15:25	Log-Likelihood:	-16.839
No. Observations:	25	AIC:	37.68
Df Residuals:	23	BIC:	40.12
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.9981	0.375	5.335	0.000	1.223	2.773
x1	1.4252	0.406	3.507	0.002	0.584	2.266

Omnibus:	0.428	Durbin-Watson:	1.087
Prob(Omnibus):	0.808	Jarque-Bera (JB):	0.427
Skew:	0.269	Prob(JB):	0.808
Kurtosis:	2.653	Cond. No.	7.46

FIGURE 12 –

Les résultats montrent que la température passée explique environ 35 % de la variabilité des projections futures ($R^2 = 0.348$), avec une relation linéaire positive significative entre les deux variables ($\hat{Y_i} = 1.4252 \cdot X_i + 1.9981$, $p < 0.01$).

La figure suivante illustre la droite de régression ajustée sur les modèles CMIP6, ainsi que la prédiction obtenue à partir de l'observation réelle X_0.

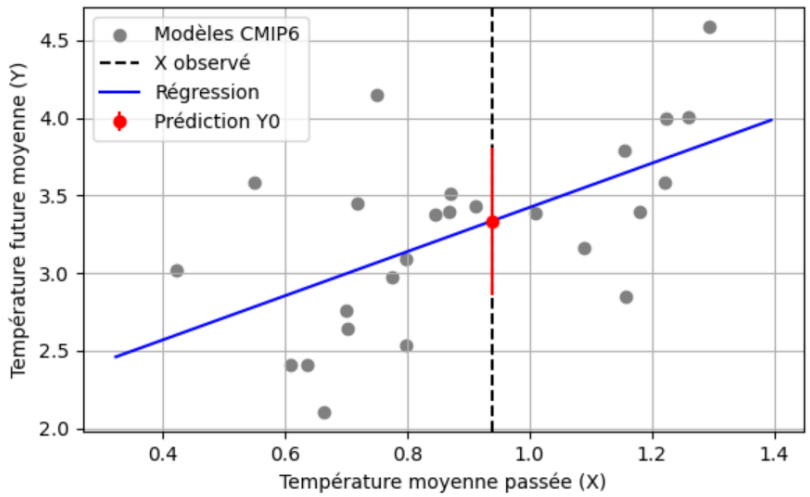

FIGURE 13 – Projections futures par régression émergente

La régression linéaire entre les températures moyennes passées simulées X et les températures moyennes futures Y montre une relation modérément croissante. La pente estimée est positive, traduisant une corrélation cohérente avec l'intuition physique : un modèle qui simule un réchauffement plus important dans le passé a tendance à projeter un réchauffement plus élevé dans le futur.

La prédiction obtenue pour Y_0 à partir de la valeur observée X_0 (ligne verticale noire) est représentée en rouge avec une barre d'erreur liée à la dispersion des résidus. La projection obtenue est proche de celle des méthodes précédentes, mais l'incertitude est ici fondée uniquement sur l'ajustement linéaire, sans tenir compte explicitement de la variabilité intermodèle. Ce point distingue cette méthode des approches 0 et 1.

La régression appliquée année par année permet de reconstituer une trajectoire temporelle contrainte par l'observation passée.

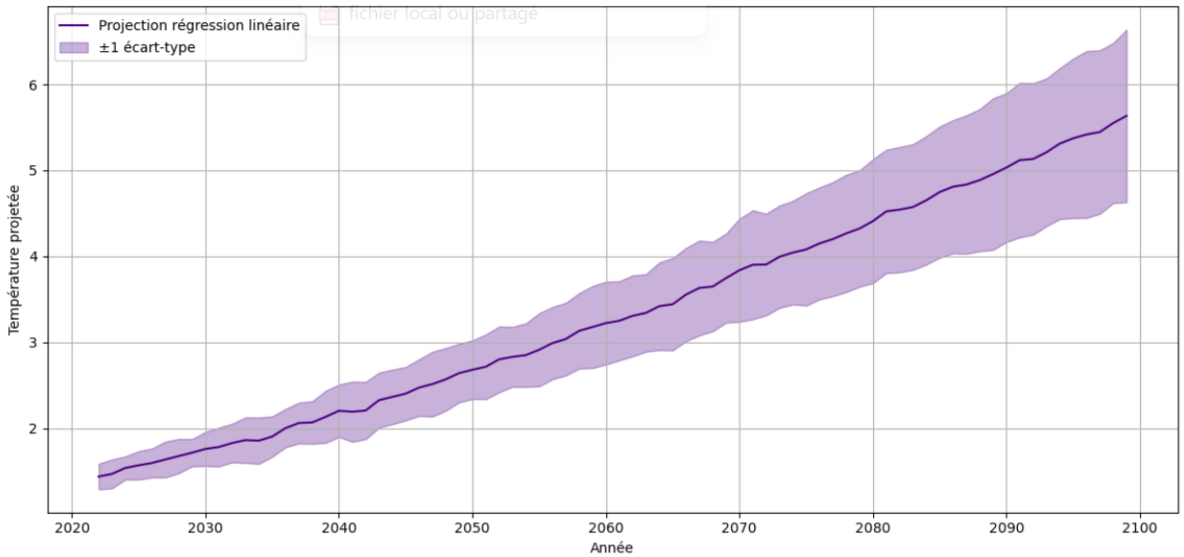

FIGURE 14 – Projections futures par régression linéaire

5.2 Comparaison de l'incertitude avec celle obtenue avec la moyenne multi-modèles

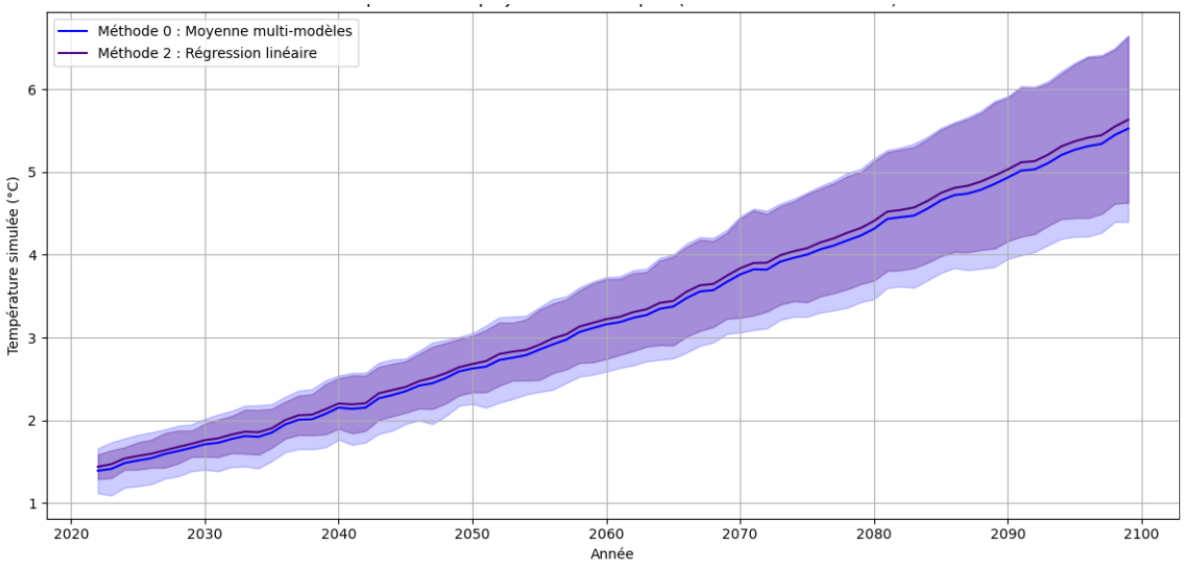

FIGURE 15 – comparaison des projections climatiques (Méthodes multi-modèles vs Régression linéaire)

Le RMSE de la régression linéaire est de 0.0746 ce qui est relativement faible et indique donc une faible erreur de projection.

Le graphique illustre clairement que la régression linéaire permet une réduction significative de l'incertitude par rapport à la moyenne multi-modèles. En effet, la bande d'incertitude associée à la méthode linéaire est nettement plus étroite tout au long de la période étudiée, ce qui indique une meilleure stabilité des prédictions. Alors que la moyenne multi-modèles intègre la variabilité entre différents modèles ce qui augmente l'incertitude globale. La régression linéaire fournit une estimation plus précise et plus cohérente de l'évolution des températures

5.3 Qualité et interprétation de la régression

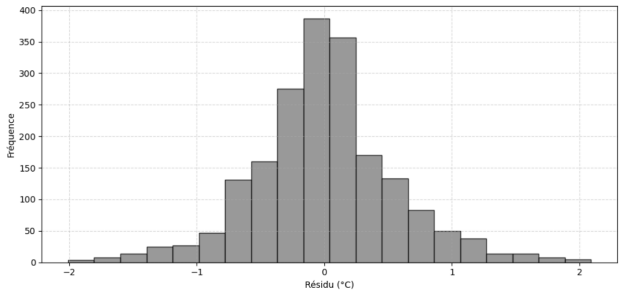

FIGURE 16 – Histogramme des résidus

L'histogramme des résidus montre une distribution globalement centrée autour de zéro, avec une forme quasi-gaussienne. Cela confirme la validité du modèle linéaire comme approximation, même si une asymétrie légère subsiste. La majorité des résidus restent dans l'intervalle $[-1, 1]$, ce qui indique une performance relativement homogène de la régression sur l'ensemble des années simulées.

En comparaison avec les méthodes précédentes, cette approche présente un compromis intéressant : elle exploite les observations réelles pour contraindre chaque année future, sans dépendre d'une pondération explicite ni supposer une distribution a priori sur les modèles.

5.4 Validation croisée Leave-One-Out appliquée à la régression linéaire

Dans notre contexte climatique, la régression linéaire ne cherche pas à ajuster des hyperparamètres, mais à identifier la définition optimale du prédicteur X.

Nous appliquons une approche validation croisée, notamment de type Leave-One-Out (LOO) en considérant uniquement les 30 dernières années (1992–2021) comme prédicteur. Pour chaque année future entre 2022 et 2099, un modèle est entraîné sur $M - 1$ modèles climatiques et testé sur le modèle restant, ce qui permet d'obtenir une projection par modèle ainsi qu'une mesure d'incertitude basée sur la dispersion des résidus.

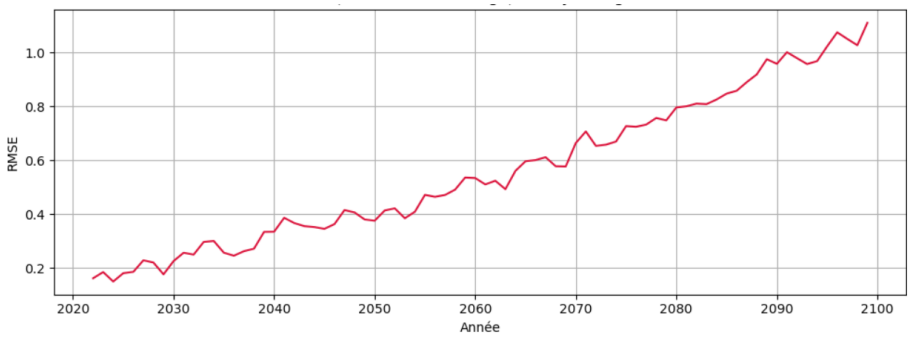

FIGURE 17 – RMSE annuel (Leave-One-Out-Ridge)

La courbe montre l'évolution du RMSE annuel entre les prédictions et la vérité simulée multi-modèle. L'erreur augmente graduellement dans le temps, atteignant une moyenne globale de 0.559. Cette tendance traduit la difficulté croissante à estimer les températures futures lointaines à partir d'un historique restreint. L'absence de rupture brutale suggère toutefois une stabilité globale du modèle. Cette évaluation met en évidence l'intérêt de la validation croisée pour estimer la fiabilité des projections en conditions de données limitées.

5.5 Comparaison des performances selon la période du predicteur X

Nous analysons ici l'impact du choix de la période temporelle du prédicteur X sur les performances de la régression linéaire optimisée par validation croisée Leave-One-Out (LOO). Pour chaque plage temporelle, la température moyenne passée est utilisée pour prédire les températures futures simulées entre 2022 et 2099. Le RMSE entre les prédictions et la moyenne des simulations CMIP6 permet de comparer l'efficacité des différents prédicteurs.

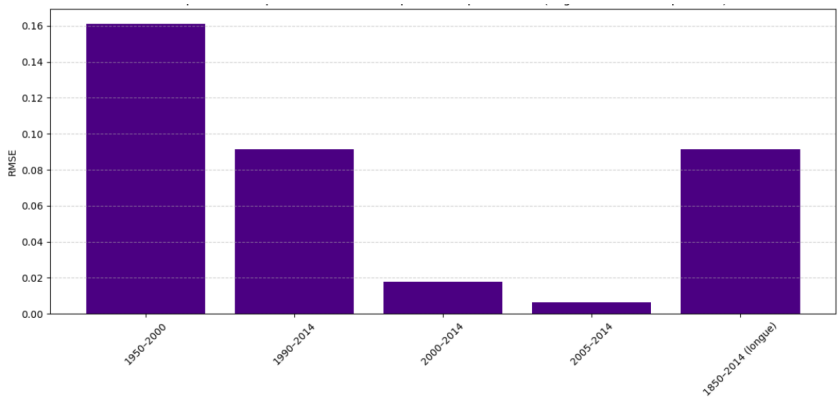

FIGURE 18 – Comparaison des performances selon la période du prédicteur X

Valeurs numériques obtenues :

$X_0 = 0.940, \quad \mu_X = 0.889, \quad \sigma_X^2 = 0.059, \quad \rho = 0.590, \quad \hat{Y}_0 = 3.333 \pm 0.481, \quad \text{SNR} = 18.392, \quad R = 33.04\,\%$

Les résultats montrent que les périodes les plus récentes (ex. 2005–2014) offrent les meilleures performances, du fait de leur proximité temporelle avec la période projetée, permettant de mieux capturer les dynamiques climatiques actuelles. À l'inverse, les périodes plus anciennes ou très longues (comme 1850–2014) dégradent la précision des projections en intégrant des informations moins pertinentes ou diluées.

5.6 Extension en multivariée

Nous étendons ici la régression univariée en considérant comme prédicteurs non plus une seule moyenne temporelle, mais l'ensemble des températures simulées pour une période passée de 30 ans (1992–2021). Ce choix permet d'exploiter davantage d'informations temporelles, capturant ainsi à la fois la moyenne, la tendance et la variabilité interannuelle. Cette approche multivariée ouvre la voie à des méthodes de modélisation plus flexibles.

Régression multivariée par forêt aléatoire Dans cette approche, nous utilisons une forêt aléatoire pour modéliser la relation entre les températures passées simulées (matrice $X \in \mathbb{R}^{M \times p}$) et les températures futures ($Y \in \mathbb{R}^{M \times T}$), où $M = 25$ est le nombre de modèles climatiques, $p = 30$ le nombre d'années passées, et $T = 78$ le nombre d'années futures projetées (2022–2099). Pour chaque année future t, un modèle est entraîné à partir des p-uplets de température passée de tous les modèles, et utilisé pour prédire la température associée à l'observation réelle X_0.

L'incertitude associée à la prédiction est estimée par l'écart-type des erreurs résiduelles sur l'échantillon d'entraînement. Contrairement aux méthodes précédentes (moyenne simple ou pondérée par distance), la forêt aléatoire permet de modéliser des relations non linéaires complexes, souvent présentes dans les dynamiques climatiques multivariées.

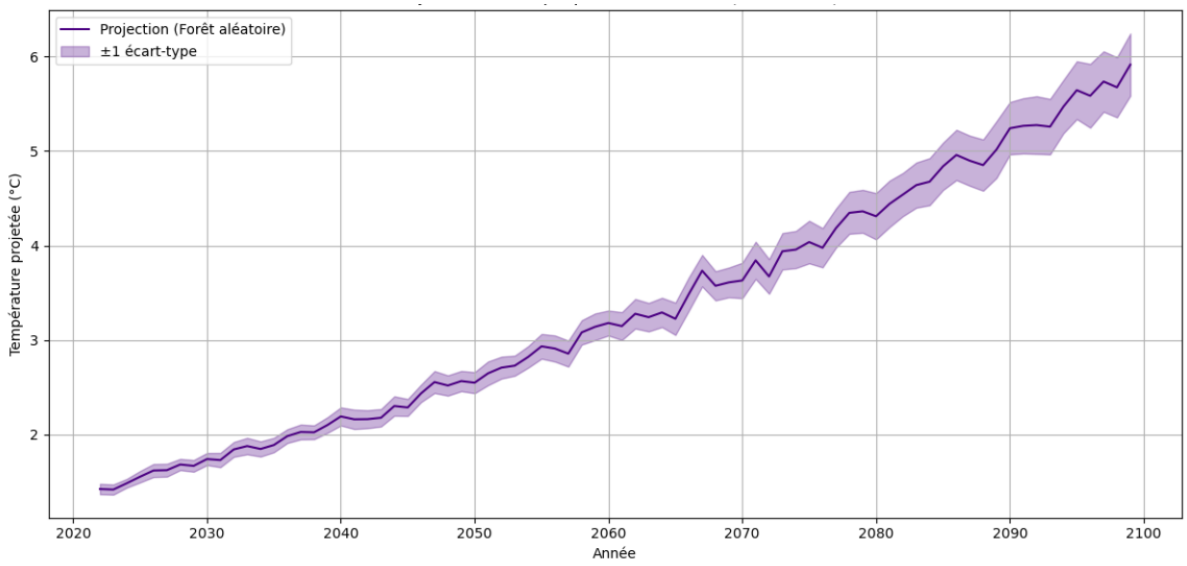

FIGURE 19 – Projection climatique multivariée par forêt aléatoire

La projection obtenue présente une hausse continue de la température jusqu'à la fin du siècle, avec une forme lissée et une bande d'incertitude fine. Comparée à la méthode 0 (moyenne multi-modèles) ou à la méthode 1 (pondération), cette approche se distingue par sa capacité à intégrer des patterns

temporels complexes sans supposer de relation linéaire. La forêt aléatoire s'avère plus robuste, grâce à son mécanisme d'agrégation d'arbres décisionnels. Cette méthode offre donc une alternative crédible et performante.

6 Méthode 3 : Filtre de Kalman

Le filtre de Kalman permet d'estimer la température future moyenne Y à partir d'une observation passée bruitée X_0, en supposant que les variables simulées X et Y suivent une loi jointe gaussienne. L'observation réelle est alors considérée comme une réalisation bruitée de X, et l'estimateur de Kalman fournit une solution optimale au sens de l'erreur quadratique moyenne.

6.1 Filtre de Kalman analytique (One-step)

Dans ce cadre, nous utilisons le filtre de Kalman comme une méthode d'estimation conditionnelle basée sur une modélisation jointe gaussienne des variables climatiques passées et futures. L'objectif est de calculer l'espérance conditionnelle $\mathbb{E}[Y \mid X_0]$, où X_0 est l'observation bruitée du climat passé, et Y la température moyenne future.

On suppose que :

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \quad X_0 = X + B, \quad B \sim \mathcal{N}(0, \sigma_0^2), \quad B \perp X$$

avec $\rho = \mathrm{corr}(X, Y)$ estimée sur les 25 modèles CMIP6, et σ_0^2 la variance d'observation extraite des données.

La loi conditionnelle $Y \mid X_0$ est également normale :

$$Y_0 \mid X_0 \sim \mathcal{N}(\hat{Y}_0, \sigma^2_{Y|X_0})$$

avec :

$$\hat{Y}_0 = \mu_Y + \frac{\rho \sigma_Y \sigma_X}{\sigma_X^2 + \sigma_0^2}(X_0 - \mu_X), \quad \sigma^2_{Y|X_0} = \sigma_Y^2 \left(1 - \frac{\rho^2 \sigma_X^2}{\sigma_X^2 + \sigma_0^2}\right)$$

Le terme de correction est appelé *gain de Kalman*. On définit également :

$$\mathrm{SNR} = \frac{\sigma_X^2}{\sigma_0^2}, \qquad R = \frac{\rho^2}{1 + \frac{1}{\mathrm{SNR}}}$$

où SNR est le rapport signal/bruit et R la réduction relative d'incertitude sur Y due à l'introduction de X_0.

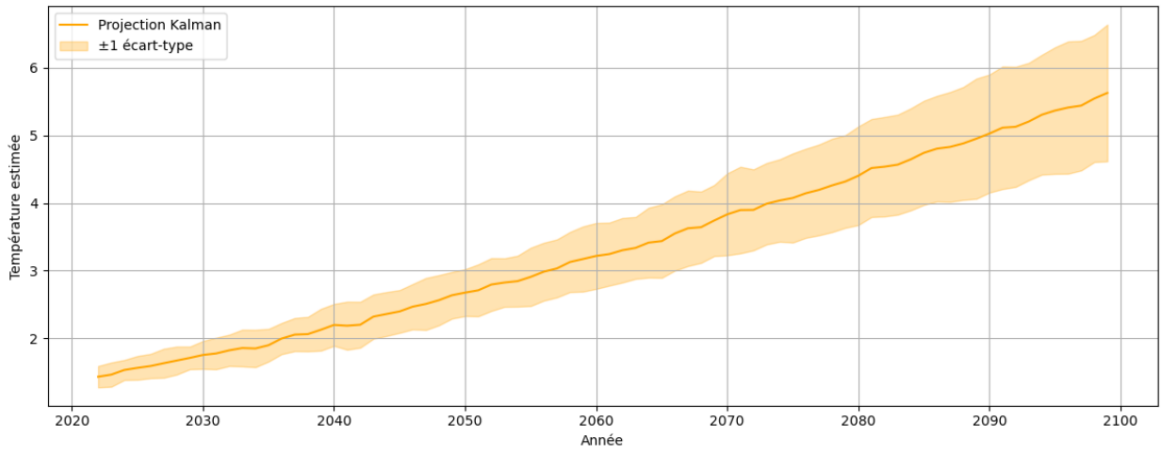

FIGURE 20 – Projections futures (2022–2099) obtenues par filtre de Kalman analytique.

Valeurs numériques obtenues :

$X_0 = 0.940, \quad \mu_X = 0.889, \quad \sigma_X^2 = 0.059, \quad \rho = 0.590, \quad \hat{Y}_0 = 3.333 \pm 0.481, \quad \text{SNR} = 18.392, \quad R = 33.04\,\%$

La projection temporelle obtenue via le filtre de Kalman montre une tendance croissante similaire aux autres méthodes. Le gain de température projetée atteint environ 5.7 °C en 2099, avec une incertitude modérée (±1 °C). L'estimation globale de la température moyenne sur la période 2022–2099 est de $\hat{Y}_0 = 3.33$ °C, avec un écart-type de 0.48 °C. Le gain d'information sur l'incertitude, quantifié par la réduction R, atteint ici 33 %, grâce à la bonne corrélation entre X et Y et à une faible erreur d'observation.

6.2 Comparaison de l'incertitude avec celle obtenue avec la moyenne multi-modèles

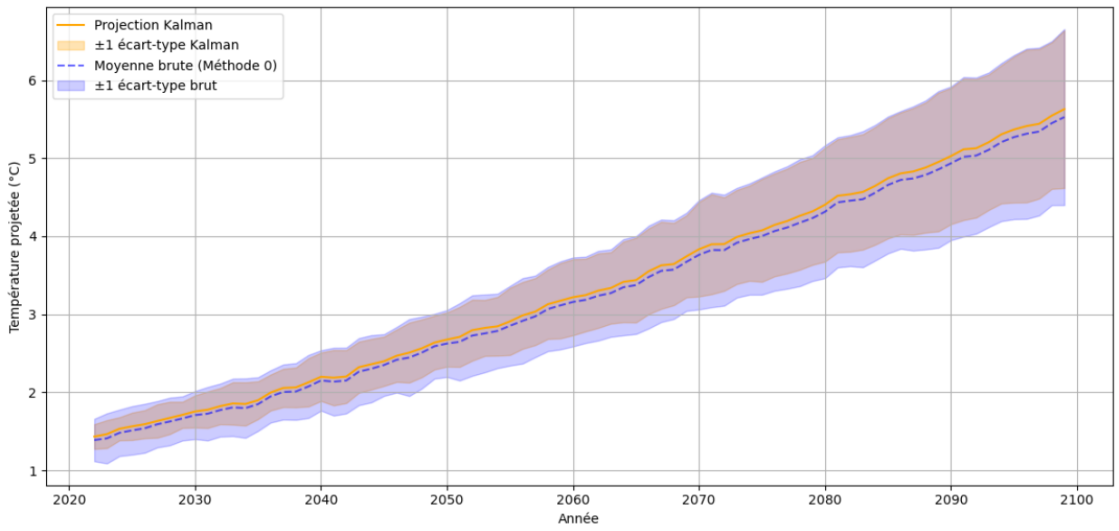

FIGURE 21 – Comparaison des projections climatiques (Kalman vs Moyenne multi-modèles)

Le graphique montre que la projection obtenue par la méthode de Kalman suit une trajectoire similaire à celle de la moyenne multi-modèles, tout en présentant une incertitude plus faible tout au long de la période. Cette réduction de l'intervalle d'incertitude suggère une amélioration de la précision des projections, probablement liée à une meilleure exploitation des données passées et à une pondération plus adaptée des modèles.

6.3 Influence de la corrélation et du rapport signal/bruit sur l'incertitude des projections

Le rapport signal sur bruit global (SNR) est estimé à environ 4.1, indiquant que la variabilité inter-modèle sur le climat passé est significativement plus importante que l'incertitude sur les observations. Toutefois, ce ratio reste modéré, ce qui limite la capacité du filtre de Kalman à réduire fortement l'incertitude sur les projections futures.

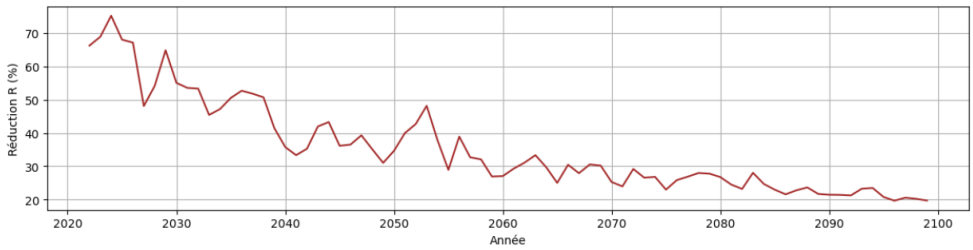

FIGURE 22 – Taux de réduction d'incertitude $R(t)$ par la méthode du filtre de Kalman, exprimé en pourcentage, pour chaque année entre 2022 et 2099.

La réduction d'incertitude moyenne sur l'ensemble de la période atteint environ 6.3 %, avec une nette décroissance au fil du temps. Cette diminution s'explique principalement par la perte progressive de corrélation ρ entre les températures passées et futures, qui devient très faible après 2050. Cela se traduit par des valeurs de $R(t)$ souvent inférieures à 30 %, et même à 20 % en fin de siècle.

En comparaison, la méthode de Kalman apporte une estimation légèrement plus contrainte que la moyenne multi-modèles (méthode 0) ou la pondération simple (méthode 1), tout en reposant sur un formalisme probabiliste rigoureux. Toutefois, son efficacité reste limitée dans notre cas, en raison d'une corrélation faible entre passé et futur, et d'un SNR modérément élevé. À performances comparables, la régression linéaire (méthode 2) s'avère plus simple à mettre en œuvre, mais n'intègre pas explicitement l'incertitude d'observation.

6.4 Validation croisée du filtre de Kalman

Dans le cadre du filtre de Kalman statique (régression bayésienne linéaire), le paramètre central est la variance d'observation σ_0^2, qui détermine le poids accordé à l'observation X_0 par rapport aux modèles. Sa calibration a été effectuée par validation croisée Leave-One-Out sur les 25 modèles CMIP6, afin de minimiser l'erreur de projection.

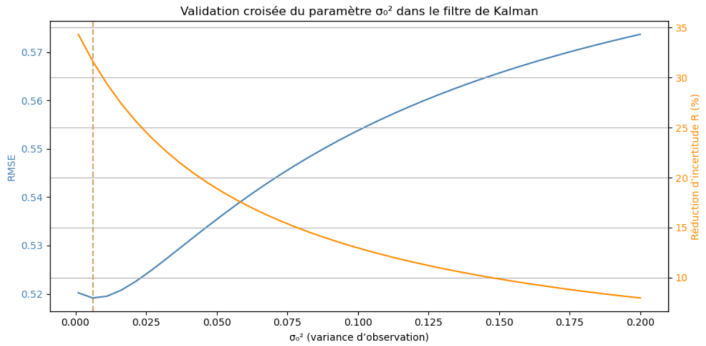

FIGURE 23 – Validation croisée du paramètre σ_0^2 dans le filtre de Kalman

Le balayage de σ_0^2 entre 0.001 et 0.2 met en évidence un minimum du RMSE pour $\sigma_0^2 = 0.0061$, avec une erreur moyenne de 0.5193. Cette configuration conduit également à une réduction moyenne de l'incertitude de 31.63 %

Les résultats révèlent une forte sensibilité aux faibles valeurs de σ_0^2, traduisant un compromis important : des valeurs trop petites mènent à une surexploitation des observations (risque de sur-ajustement), tandis que des valeurs trop grandes réduisent leur influence dans l'estimation. Le choix optimal de σ_0^2 permet ainsi de tirer profit des observations sans les surpondérer.

6.5 Comparaison des performances selon la période du prédicteur X

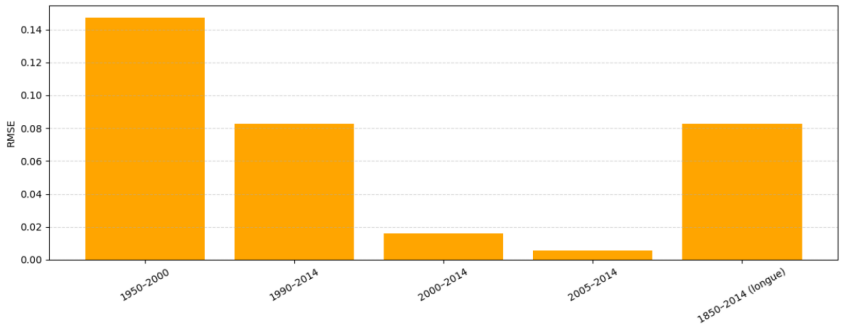

FIGURE 24 – Comparaison des performances selon la période du prédicteur X (Méthode 3 : Kalman)

Nous avons testé le filtre de Kalman sur différentes périodes de calcul du prédicteur X, en maintenant $\sigma_0^2 = 0.0061$. Les résultats confirment que les périodes récentes (2000–2014, 2005–2014) minimisent le RMSE, tandis que les périodes anciennes ou trop longues dégradent la précision.

Comme pour la méthode 1, un prédicteur temporellement proche de la période projetée renforce la corrélation ρ, et donc la qualité de l'estimation. Le filtre de Kalman partage cette sensibilité au choix du prédicteur, bien qu'il intègre explicitement l'incertitude.

6.6 Extension en multivarié

Le filtre de Kalman peut également être adapté à une approche multivariée. Dans ce cas, l'équation d'estimation devient une régression linéaire bayésienne exploitant la covariance entre les vecteurs de températures simulées X_i et les sorties Y_i. Cette méthode repose sur une estimation des matrices Σ_{XX}, Σ_{YX} et de l'incertitude associée aux observations. Elle peut être combinée avec une réduction de dimension pour limiter le sur-apprentissage en grande dimension.

Contrairement à la version univariée ou à l'approche ACP, cette méthode exploite l'ensemble des covariances entre les températures passées X et futures Y, simulées par les modèles climatiques.

Le filtre de Kalman repose ici sur une modélisation jointe gaussienne du couple (X, Y). L'estimation de la projection $\hat{Y}$ s'obtient par :

$$\hat{Y} = \mu_Y + K(X_0 - \mu_X), \quad \text{où} \quad K = \Sigma_{YX}(\Sigma_{XX} + \Sigma_0)^{-1}$$

où :

— μ_X, μ_Y : moyennes des données passées et futures simulées,
— Σ_{XX}, Σ_{YX} : matrices de covariance intra et croisée entre X et Y,
— Σ_0 : matrice diagonale représentant l'incertitude sur l'observation X_0.

L'incertitude sur la projection est donnée par :

$$\text{Var}(\hat{Y}) = \text{diag}\left(\Sigma_{YX}(\Sigma_{XX} + \Sigma_0)^{-1}\Sigma_{XY}\right)$$

Cette version analytique du filtre de Kalman est particulièrement bien adaptée à notre contexte : les observations X_0 sont fixes et connues en bloc, sans dynamique temporelle à modéliser. Un filtre récursif classique (avec mise à jour pas à pas) ne présenterait donc pas d'avantage ici, sauf dans un contexte de données séquentielles ou de modélisation stochastique du climat.

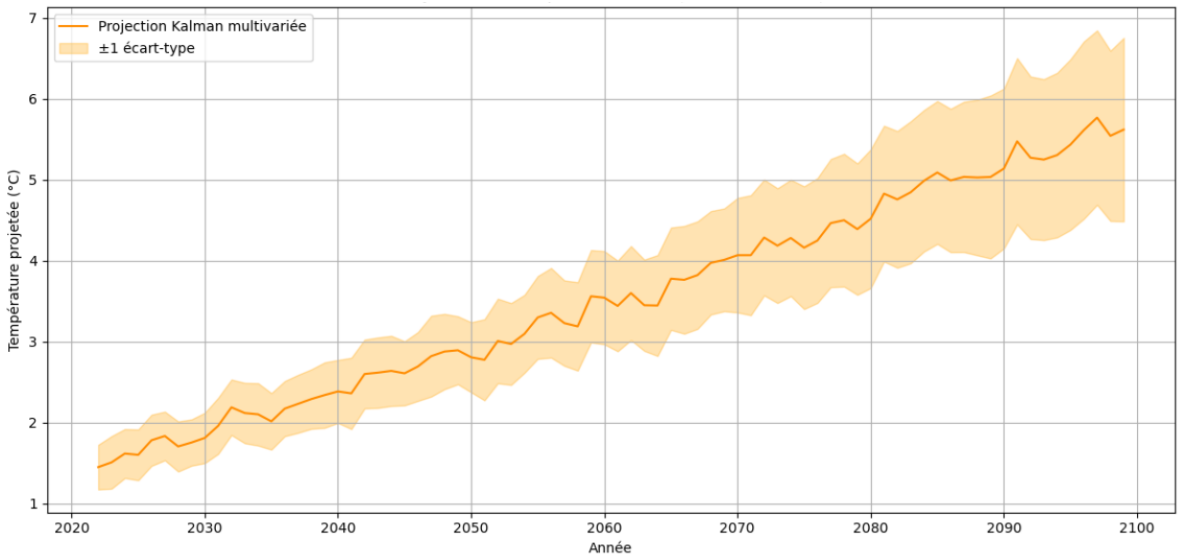

FIGURE 25 – Projection climatique obtenue par le filtre de Kalman multivarié.

La projection obtenue suit une dynamique lissée et physiquement cohérente, avec une température projetée autour de 5.7 °C à l'horizon 2100. L'incertitude reste modérée tout au long de la période, traduisant la forte structure corrélée exploitée par le filtre.

Par rapport aux méthodes précédentes, cette approche ne nécessite pas de choisir un prédicteur unique ou réduit, et exploite directement la structure multivariée complète. Cela la rend plus rigoureuse et plus robuste, bien qu'elle soit plus coûteuse en calcul (inversion de matrices) et sensible aux données bruitées.

En résumé, le filtre de Kalman multivarié constitue une alternative probabiliste solide pour contraindre les projections climatiques, en maximisant l'usage de l'information passée disponible.

7 Comparaison des méthodes

7.1 Synthèse des performances et hypothèses des méthodes

Méthode	Hypothèses	Paramètres principaux	Forces	Limites
Moyenne multi-modèles (0)	Aucune (non-paramétrique)	Aucun	Simple, robuste, neutre	Incertitude forte, pas de contrainte observationnelle
Pondération naïve (1)	Proximité = performance, indépendance utile	σ_D, σ_S	Intègre les observations, améliore l'incertitude	Sensible au choix des hyperparamètres, redondance des modèles
Régression linéaire (2)	Relation linéaire $Y = aX + b$	Période de X, coefficients a, b	Simple, interprétable, contrainte forte sur X_0	Relation faible ou bruitée $\Rightarrow$ erreur résiduelle élevée
Filtre de Kalman (3)	Gaussianité, linéarité, bruit additif	μ_X, μ_Y, ρ, σ_X^2, σ_Y^2, σ_0^2	Réduction d'incertitude explicite, optimalité quadratique	Hypothèses fortes, sensibilité à ρ, calcul matriciel complexe

7.2 Un bon modèle pour le passé est-il forcément fiable pour le futur ?

Il peut être tentant de penser qu'un modèle climatique qui reproduit bien les températures passées sera également performant pour les projections futures. Pourtant, cette idée mérite d'être nuancée. Voici pourquoi :

- **Corrélation ne signifie pas causalité** : un modèle peut bien s'ajuster aux observations historiques simplement parce qu'il a été calibré dans ce but, sans pour autant modéliser correctement les mécanismes physiques sous-jacents. Ce bon ajustement ne garantit donc pas qu'il saura prédire les dynamiques futures.
- **Biais structurels communs** : plusieurs modèles partagent des blocs de code ou des paramétrisations similaires (par exemple sur les nuages ou les océans). Cela peut créer une illusion de consensus ou de performance, alors qu'ils souffrent en réalité de biais collectifs non détectés.
- **Changement de régime climatique** : les prochaines décennies pourraient être marquées par des phénomènes encore absents des données historiques, comme des rétroactions amplificatrices (fonte du permafrost, effondrement d'un courant océanique, etc.). Dans ce cas, les modèles basés uniquement sur le passé risquent de sous-estimer les risques.

Autrement dit, toutes les méthodes qui utilisent les observations passées pour contraindre les projections futures reposent implicitement sur une hypothèse de continuité ou de stationnarité du système

climatique. Or, cette hypothèse devient plus fragile à mesure qu'on s'éloigne dans le temps, surtout dans un contexte de changement climatique rapide.

7.3 Paramétrisation des méthodes et risque de surapprentissage

Chaque méthode présentée repose sur un ensemble de paramètres, qui influencent directement la qualité des projections. Ces paramètres peuvent être de plusieurs natures :

- **Fixés à priori** : comme les hyperparamètres σ_D et σ_S dans la méthode de pondération, ou la variance d'observation σ_0^2 dans le filtre de Kalman.
- **Estimés sur les données** : tels que les moyennes μ_X, μ_Y, la corrélation ρ, ou les coefficients de régression linéaire (a, b).
- **Ajustés par validation croisée** : comme dans la recherche des meilleurs hyperparamètres pour minimiser l'erreur de projection (RMSE).

Cette dépendance aux paramètres pose plusieurs enjeux :

- **Trop de paramètres** peut rapidement conduire à un risque de surapprentissage, en particulier dans les approches multivariées (comme la forêt aléatoire ou le Kalman multivarié), où les modèles peuvent s'ajuster aux moindres variations du petit échantillon disponible ($M = 25$ modèles).
- **La sensibilité aux réglages** est parfois très forte : une valeur trop faible de σ_0^2 ou σ_D peut par exemple faire basculer toute l'estimation, en donnant trop de poids à l'observation ou à un seul modèle.
- **La validation croisée** (notamment de type Leave-One-Out) est utile pour éviter ce surajustement, mais elle reste fragile sur des échantillons réduits : un seul modèle atypique peut fortement influencer le résultat.

En résumé, la paramétrisation est un maillon clé de la robustesse des méthodes. Un bon réglage permet d'exploiter au mieux l'information disponible sans tomber dans la surconfiance, tandis qu'un mauvais choix peut au contraire produire des projections faussées ou trop optimistes. C'est pourquoi il est crucial de documenter ces choix et d'évaluer leur sensibilité à travers des tests complémentaires.

8 Conclusion

Chacune des méthodes explorées apporte un éclairage complémentaire sur la réduction d'incertitude dans les projections climatiques :

- La **moyenne multi-modèles** est un point de départ neutre mais peu informatif.
- La **pondération par performance passée** améliore légèrement l'incertitude mais dépend fortement des paramètres.
- La **régression linéaire** est simple, robuste, et assez performante dans notre cas.
- Le **filtre de Kalman**, en version univariée ou multivariée, offre une approche probabiliste rigoureuse avec une réduction d'incertitude formelle, mais au prix de fortes hypothèses de distribution et de corrélation.

En pratique, une combinaison prudente de ces méthodes (ex. moyenne pondérée régularisée ou Kalman multivarié après ACP) pourrait offrir un bon compromis entre réduction d'incertitude, robustesse, et cohérence physique. Néanmoins, il convient de garder à l'esprit que toutes ces approches sont sensibles à la qualité des données d'entrée, au choix du prédicteur, et aux hypothèses structurelles implicites — d'où la nécessité d'une évaluation critique continue et d'une transparence sur les incertitudes restantes.

Références

[1] Gouvernement du Canada. (2023). *Liste des modèles CMIP6 utilisés dans le Portail des scénarios climatiques du Canada*. https://scenarios-climatiques.canada.ca/?page=cmip6-model-list

[2] Bowman, K. W., Cressie, N., Qu, X., and 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

[3] Brunner, L., Lorenz, R., Zumwald, M., and 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] L.Dambrine, F.Dossou, M.Kane (2025). *Filtre de Kalman*, Projet d'expertise Master 1, supervisé par A.Genadot.

[5] ZEMMARI Akka (2025). *Cours Analyse, Classification, Indexation des Données*