Redescendons sur Terre !

Analyse et modélisation du réchauffement climatique

réalisées par

Jean-Baptiste CASSIN,
Salim ABDELOUAHAB
et Thomas CHAULET

Introduction

Problématique:

Le réchauffement climatique est-il réel?

Selon l'Organisation des Nations Unies, la période 2011-2020 a été la décennie la plus chaude jamais enregistrée. En 2019, la température moyenne de la planète se situait 1,1 °C au-dessus des niveaux de l'ère préindustrielle. Le réchauffement climatique dû aux humains augmente actuellement à un rythme de 0,2 °C par décennie.
L'utilisation de combustibles fossiles, la déforestation et l'élevage de bétail influent de plus en plus sur le climat et la température de la terre.

Ces activités libèrent d'énormes quantités de gaz à effet de serre, qui viennent s'ajouter à celles naturellement présentes dans l'atmosphère, renforçant ainsi l'effet de serre et le réchauffement de la planète.

Dans cette étude, nous allons nous poser les questions suivantes:

- *Grâce aux données dont nous disposons, pouvons-nous confirmer le réchauffement climatique?*
- *Quel est l'évolution des températures sur différents points de la Terre sur une période de plusieurs centaines d'années?*
- *Les anomalies de températures sont-elles corrélées aux émissions de CO2?*
- *Pouvons-nous prédire l'évolution des températures dans les prochaines décennies?*

Objectifs de notre étude:

- *Constater le dérèglement climatique global à l'échelle de la planète sur les derniers siècles et dernières décennies.*
- *Analyser par zone géographique pour voir les évolutions différentes.*
- *Comparer avec des phases d'évolution de température antérieure à notre époque.*
- *Comparer avec d'autres facteurs comme le taux de CO2 dans l'air.*
- *Modéliser et prédire des variations de températures sur plusieurs dizaines d'années*

Sommaire

c) Analyse du CO2 dans l'atmosphère (données du Climate & Energy College)

d) Comparaison avec les anomalies de température et les émissions de C02 (données Gistemp et Climate & Energy College)

e) Analyse des émissions de C02 par pays (Our World In Data)

B - Modélisation et prédiction à l'aide de techniques de Machine Learning

1. Machine Learning avec régression polynomiale

a) Préparation des données

b) Prédictions

c) Conclusion

2. Machine Learning avec Random Forest

a) Préparation des données

b) Prédictions

c) Conclusion

3. Machine Learning avec Facebook Prophet

a) Préparation des données

b) Prédictions

c) Conclusion

4. Conclusion

C - Bilan

D - Suite du projet

E - Ressources

A - Exploration, Data Visualisation et pre-processing des données

1. Exploration des données

a) Etude des jeux de données des températures terrestres

Pour cette étude, nous avons fait des recherches sur Internet pour déterminer quelles étaient les sources les plus appropriées pour notre analyse

Pour l'analyse de la température terrestre, les jeux de données principalement utilisés proviennent du *projet Berkeley Earth Project* disponibles sur le site de l'Université de Berkeley https://berkeleyearth.org/data/

Les données proviennent des enregistrements des stations météo et sont supposées être représentatives d'une zone entourant chacune des stations météo formant l'ensemble de tous les points plus proches de cette station que de tout autre (algorithme de Voronoï). Comme les stations ne sont pas réparties de manière homogène et que leur nombre a considérablement évolué dans le temps, des « erreurs algorithmiques » sont associées à cette méthode de moyennage spatial.

La température moyenne (en fait son anomalie) est calculée en faisant la somme des données individuelles provenant des différentes stations et en attribuant à chaque point un poids proportionnel à la cellule correspondante (moyenne pondérée). Comme la taille des cellules a changé au fil du temps, le poids des points d'origine (les stations météo) a également changé, ce qui induit un biais dans le calcul de la valeur moyenne globale.

Les anomalies de température présentées dans les différents jeux de données sont une présentation chiffrée des écarts de températures par rapport aux moyennes sur la période 1951-1980.

Dans la liste des données ajustées, l'effet des influences non climatiques , tel que l'effet de l'îlot de chaleur urbain, est éliminé autant que possible. Initialement, seuls les cas documentés étaient ajustés. Cependant, la procédure actuelle utilisée par la NOAA / NCEI applique un système automatisé basé sur des comparaisons systématiques avec les stations voisines pour traiter les fluctuations documentées et non documentées qui ne sont pas directement liées au changement climatique.

Nous avons plusieurs jeux de données à notre disposition:

- Liste des variations de température globale (Terre uniquement) en degré

Celsius de Janvier 1753 à Septembre 2022
- Liste des variations de température dans l'hémisphère nord (Terre uniquement) en degré Celsius de Janvier 1880 à Septembre 2022
- Liste des variations de température dans l'hémisphère Sud (Terre uniquement) en degré Celsius de Janvier 1880 à Septembre 2022
- Liste des variations de température dans chaque pays en degré Celsius (les périodes varient de pays en pays)

En complément du jeu de données précédent, nous avons utilisé celui de GISS Surface Temperature Analysis que l'on peut trouver sur le site de la NASA:
https://data.giss.nasa.gov/gistemp/
Les informations présentes dans les jeux de données sont similaires à celles que l'on a dans le jeu de Berkeley mais présentées sous une autre forme permettant une utilisation plus aisée pour certaines visualisations.

Nous utilisons pour notre étude 2 datasets:
GLB.Ts+dSST (anomalies globales)
ZonAnn.Ts+dSST (anomalies par zones hémisphériques et par tranche de latitudes)
Les anomalies de température présentées dans ce jeu de données sont une présentation chiffrée des écarts de températures par rapport aux moyennes sur la période 1951-1980, les anomalies ont été standardisées et les outliers supprimés.

Pour compléter notre analyse, nous avons décidé d'utiliser les données scientifiques du PAGES2k Consortium 2019 qui fournit un support pour la collecte et la synthèse des observations, des reconstitutions et la modélisation des dynamiques climatiques, écosystémiques, environnementales et sociétales dans le passé.
https://pastglobalchanges.org/science/wg/2k-network/intro

le dataset se trouve sur cette page:
https://web.archive.org/web/20200229093647/https://figshare.com/articles/Reconstruction_ensemble_median_and_95_range/8143094

Nous avons une nouvelle reconstitution de la température globale remontant à l'an 1AD grâce au travail de l'équipe PAGES2k. Cette reconstruction comprend des données provenant d'une grande variété d'enregistrements proxy tels que les cernes d'arbres, les dépôts de grottes, les coraux, etc.

b) Etude des jeux de données des émissions de CO2

Pour cette analyse, nous avons utilisé les données scientifiques du Climate & Energy College qui est une équipe internationale de chercheurs en début de

carrière. Le Collège mène des recherches sur les systèmes climatiques et énergétiques dans un environnement interdisciplinaire, faisant progresser les connaissances et éclairant les réponses aux défis complexes du changement climatique.

Ce centre de recherche de classe mondiale est situé à l'Université de Melbourne, en collaboration avec des instituts de recherche australiens et allemands de premier plan. Leurs recherches sont centrées sur le changement climatique et les transitions énergétiques.
https://www.climatecollege.unimelb.edu.au/

le dataset se trouve sur cette page:
https://www.climatecollege.unimelb.edu.au/cmip6

dataset:
mole_fraction_of_carbon_dioxide_in_air_input4MIPs_GHGConcentrations_CMIP_UoM-CMIP-1-1-0_gr3-GMNHSH_0000-2014.csv

Il recense les taux de C02 dans l'atmosphère depuis l'année 0 jusqu'à l'année 2014. Il comprend 3 variables: le taux global, le taux dans l'hémisphère Nord et celui dans l'hémisphère Sud.

Pour aller plus loin dans l'analyse des émissions de C02, nous avons utilisé les données issues du projet Our World in Data du Global Change Data Lab. Cette organisation à but non lucratif basée au Royaume-Uni, partenaire du l'université d'Oxford à pour mission de publier « la recherche et les données de milliers de chercheurs et spécialistes pour progresser contre les plus grands problèmes du monde ».
https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

le dataset se trouve sur cette page:
https://github.com/owid/co2-data

2. Etude de la volumétrie et Preprocessing

a) Données de Berkeley Earth

Chaque jeux de données est une série temporelle comprenant l'année, le mois, l'anomalie de température et l'incertitude liée à la mesure.

Les données ont déjà été nettoyées et homogénéisées par les chercheurs du projet Berkeley Earth. Les valeurs aberrantes ont été supprimées.
Les jeux de données sont très propres. Les valeurs manquantes apparaissent la

plupart du temps au 19ème siècle, due au fait que les capteurs de températures étaient plus rares et souvent moins fiables qu'actuellement.

Nous décidons de ne pas remplacer les valeurs manquantes car l'absence de valeurs n'indique pas que la valeur est nulle et nous ne voulons pas ajouter de fausses informations à nos données.

Nous avons agrégé les températures globales et hémisphériques dans un seul et même fichier csv sous le nom:
températures_globales_hémisphériques.csv

Pour les données des anomalies par pays, nous avons environ 240 jeux de données, correspondant à l'ensemble des pays et territoires de la terre. Nous avons décidé d'agréger l'ensemble de ces données dans un seul et même Dataset afin de faciliter le traitement et l'analyse, sous le nom:
températures_par_pays.csv
Cette tâche n'était pas simple car les fichiers n'étaient pas tous semblables et a demandé plusieurs jours pour arriver à un résultat convenable.

b) Données Gistemp

Le dataset comprend les anomalies de températures Terre-Océan de l'année 1880 à 2023. Ces anomalies sont présentées par mois et par trimestre.
Il n'y a pas de valeurs manquantes

c) Données de PAGES2k

Le fichier de données est le suivant:
Full_ensemble_median_and_95pct_range_edit.txt

Le fichier contient la médiane et la plage de 95 % (2,5e et 97,5e centiles) de l'ensemble complet de 7 000 membres pour toutes les méthodes des reconstructions de la température moyenne mondiale du PAGES2k Consortium 2019.

Les colonnes de ce tableau sont : année , données brutes sur les cibles instrumentales, ensemble de reconstruction 50e, 2,5e et 97,5e centiles, données cibles instrumentales filtrées par Butterworth sur 31 ans, reconstruction filtrée par Butterworth sur 31 ans, 50e, 2,5e et 97,5e centiles
Toutes les valeurs sont en °C par rapport à 1961-1990 sur la moyenne annuelle de mai à avril.

Il n'y a pas de données manquantes.

d) Données du Climate & Energy College

Ce jeu de données recense les émissions de CO2 depuis l'année 0 jusqu'à 2014. Nous avons 3 colonnes: l'année, les émissions de C02 globales, les émissions de C02 dans l'Hémisphère Nord et les émissions de C02 dans l'Hémisphère Sud

C'est un jeu de données très propre car il n'y a aucune donnée manquante et pas de outliers.

e) Données Our World in Data
Ce jeu de données comprend de nombreuses catégories et types d'émissions de C02 en millions de tonnes, classées par pays et par années.

Il y a beaucoup de données manquantes due principalement au fait que les pays n'étaient pas égaux dans la captation des données de CO2.

Nous décidons de ne pas remplacer les valeurs manquantes car l'absence de valeurs n'indique pas que la valeur est nulle et nous ne voulons pas ajouter de fausses informations à nos données.

Pour notre analyse, nous décidons de garder uniquement la colonne C02, qui est la valeur globale d'émission de C02 pour le pays, à la date donnée.

Nous exportons un fichier sous le nom:
co2_par_pays.csv

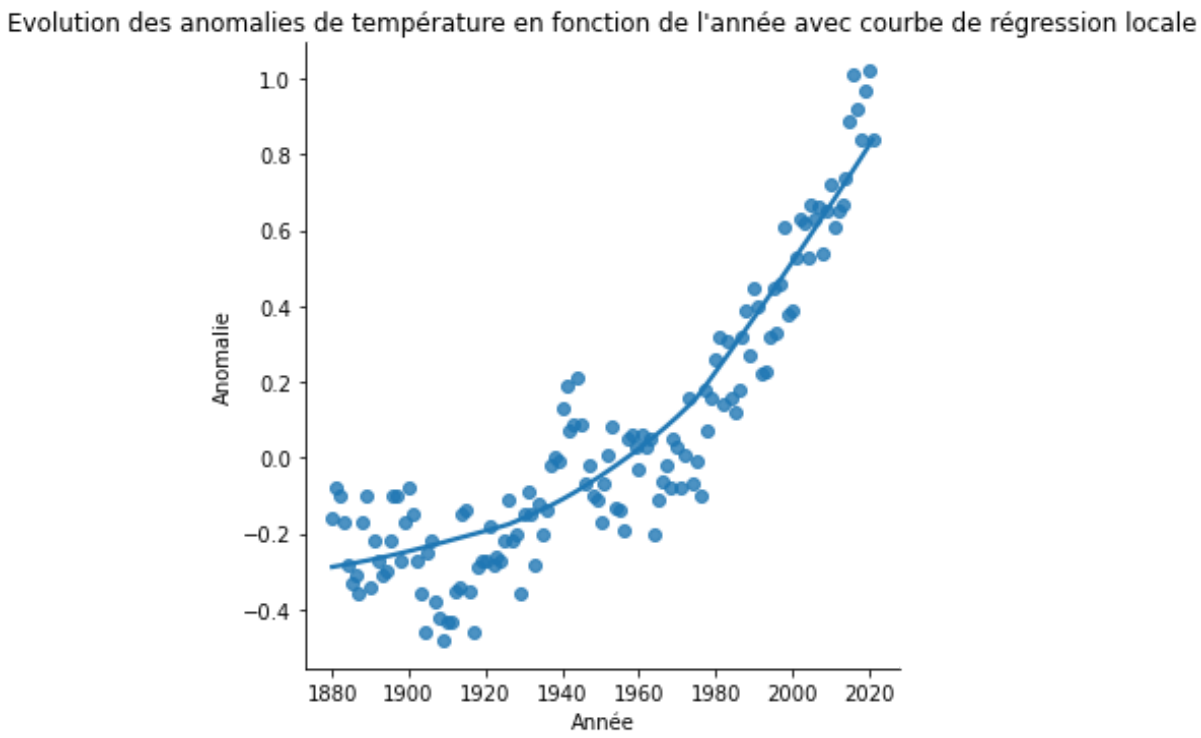
3. Visualisation et outils statistiques

a) Analyse des anomalies de températures

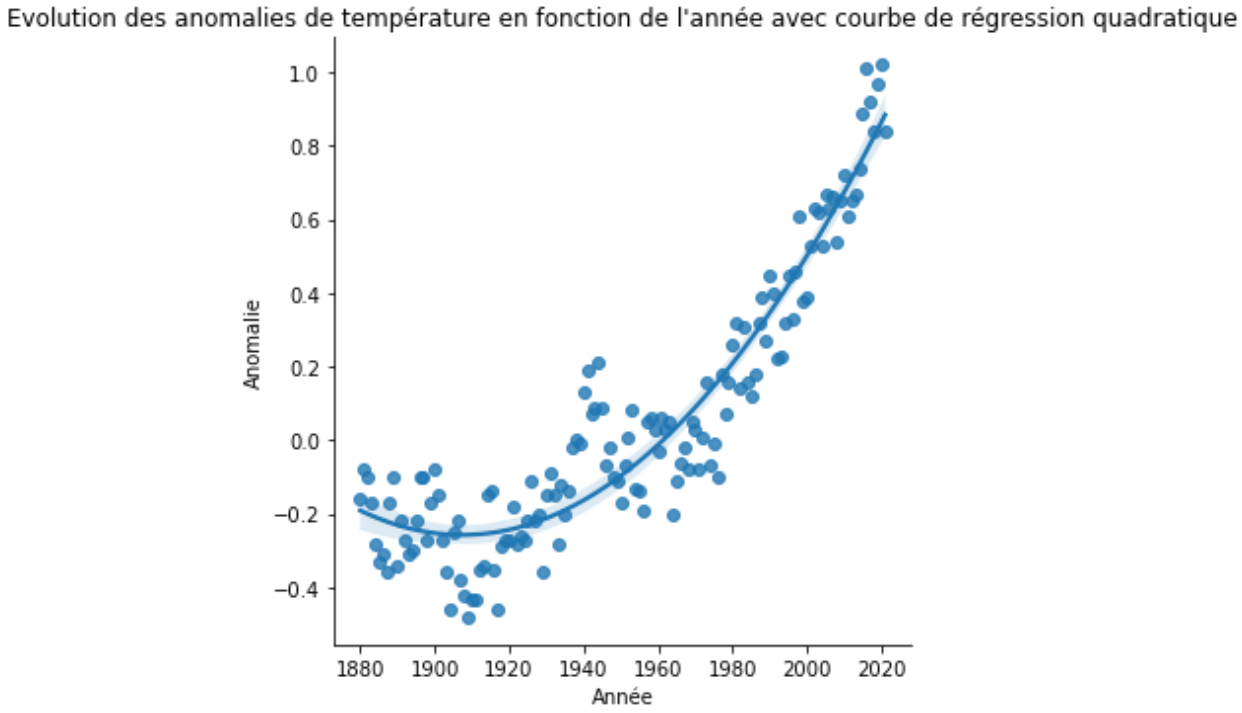
Hypothèse: nous faisons l'hypothèse que les variations de températures enregistrées ont tendance à être positives, ce qui signifie que la température du globe à tendance à augmenter de manière générale.
Nous essaierons de déterminer les zones où les températures ont tendance à le plus augmenter et à savoir si cette tendance haussière est homogène au fil du temps.

Premières visualisations avec évolution des moyennes globales, par hémisphère
Pour cette étude, nous utilisons les données Gistemp.

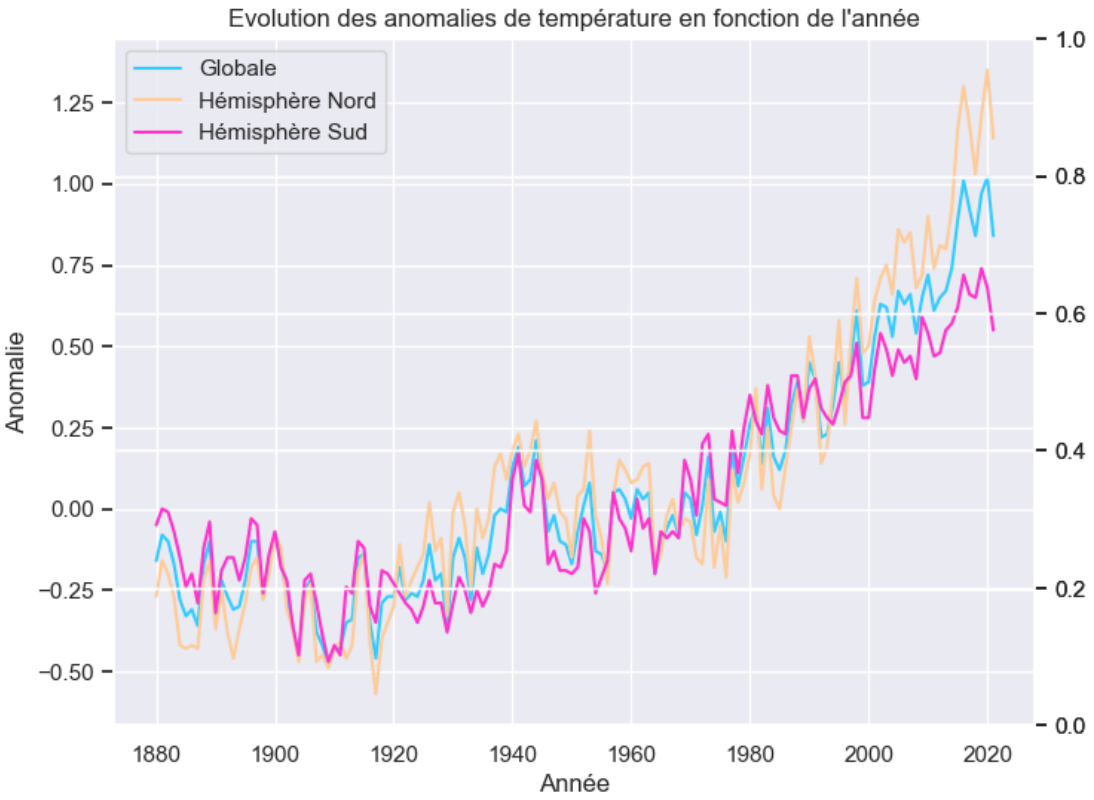
Températeur globale, nuage de points et courbe lisse par régression polynomiale d'ordre 2 (LOWESS)



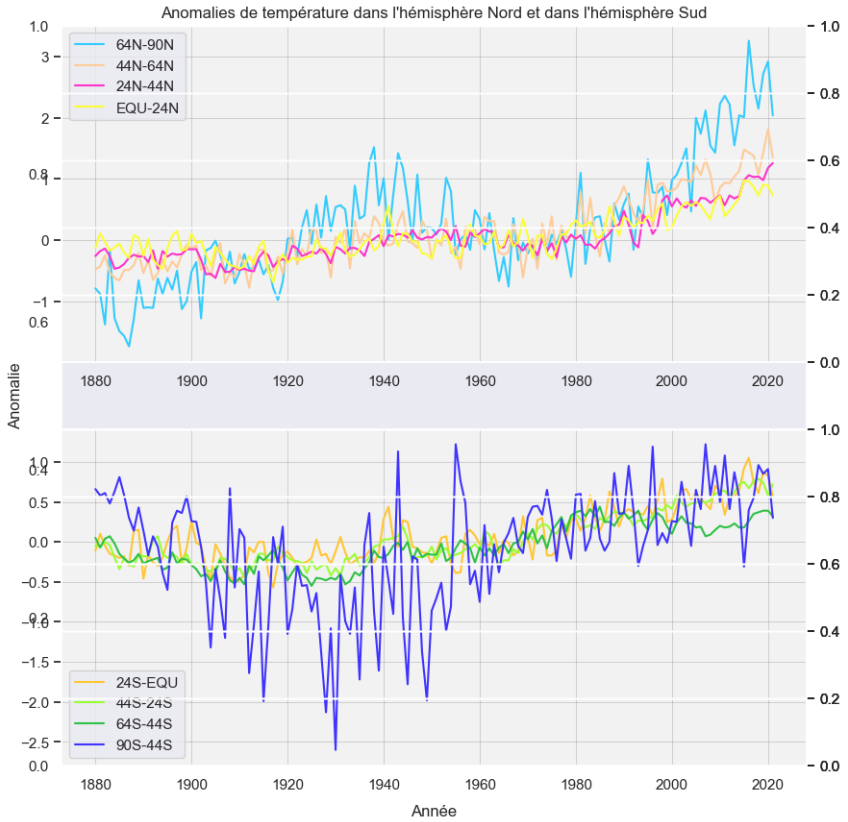
Température dans l'hémisphère nord, nuage de points et courbe lisse par régression polynomiale d'ordre 2 (LOWESS)



Superposition des températures globale, dans l'hémisphère nord et dans l'hémisphère sud.

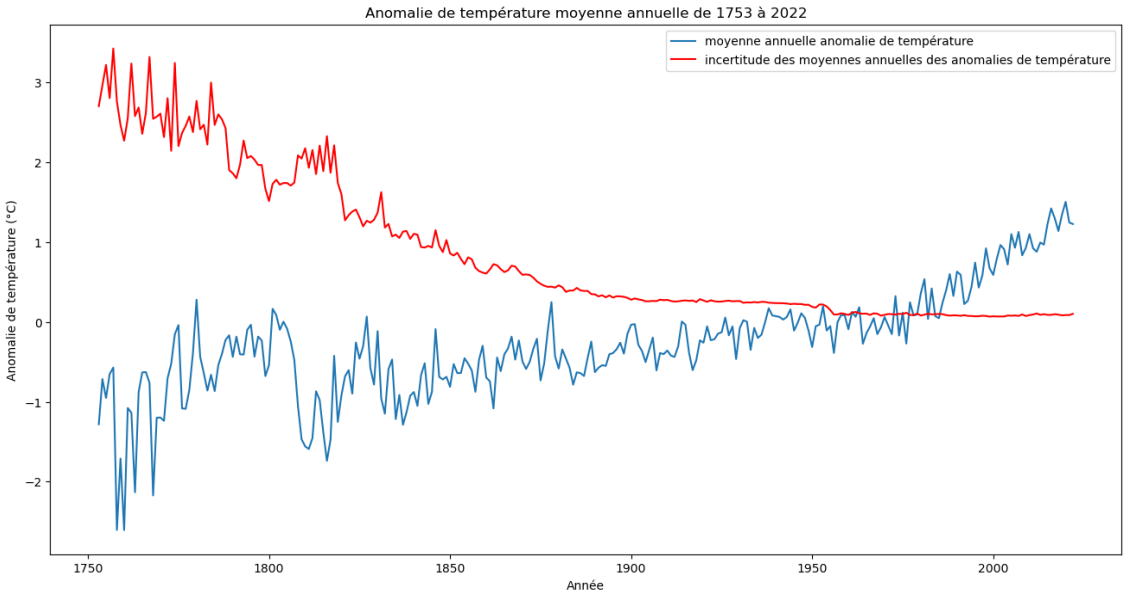


Variation des températures en fonction de la zone (plage de latitudes)

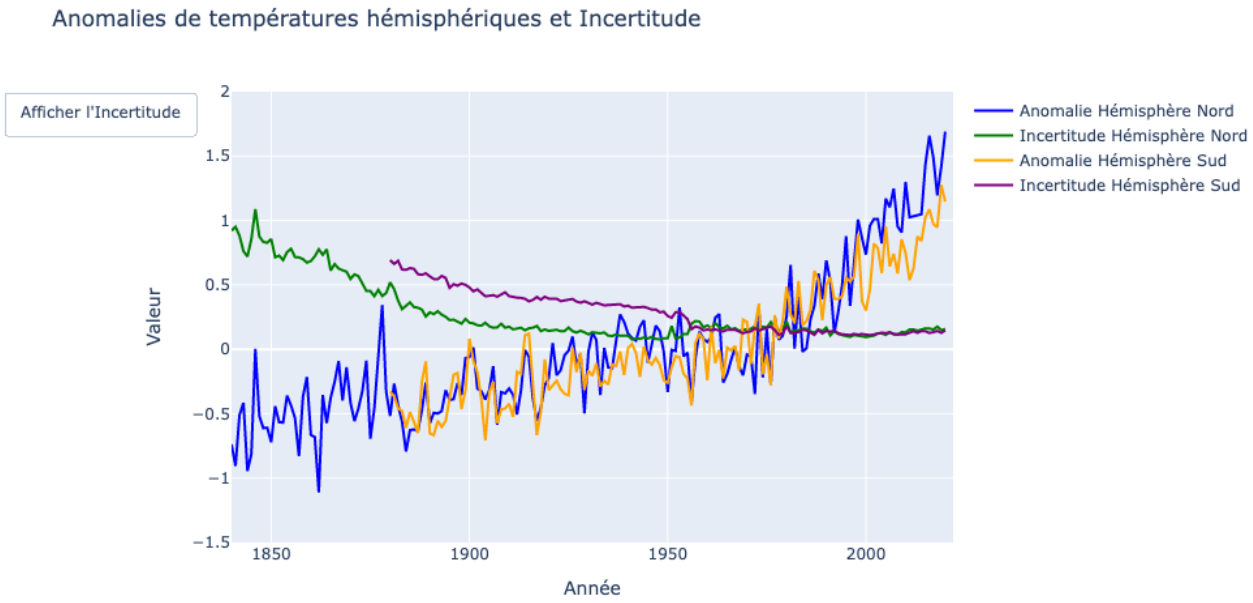


En utilisant les données issues de Berkeley, nous pouvons faire apparaître d'autres éléments de compréhension.

Ce premier graphe montre la courbe des anomalies globales ainsi que l'incertitude des précisions. Nous constatons que l'incertitude était très élevée lors des premières mesures, ici 1753 et qu'elle a grandement diminué pour être pratiquement nulle depuis les années 1960. Cela est dû au fait que les équipements de mesures sont de plus en plus précis et que leur fiabilité est croissante.



Le deuxième graphe nous montre les anomalies et les incertitudes pour chaque hémisphère. Comme pour le graphe précédent, l'incertitude était très importante au moment des premières mesures mais elle a beaucoup diminué au fur et à mesure du temps. Les données ont été collectées plus tôt dans l'hémisphère Nord et nous constatons, comme avec les données de la NASA que l'hémisphère Nord est plus impacté par le réchauffement climatique que l'hémisphère Sud

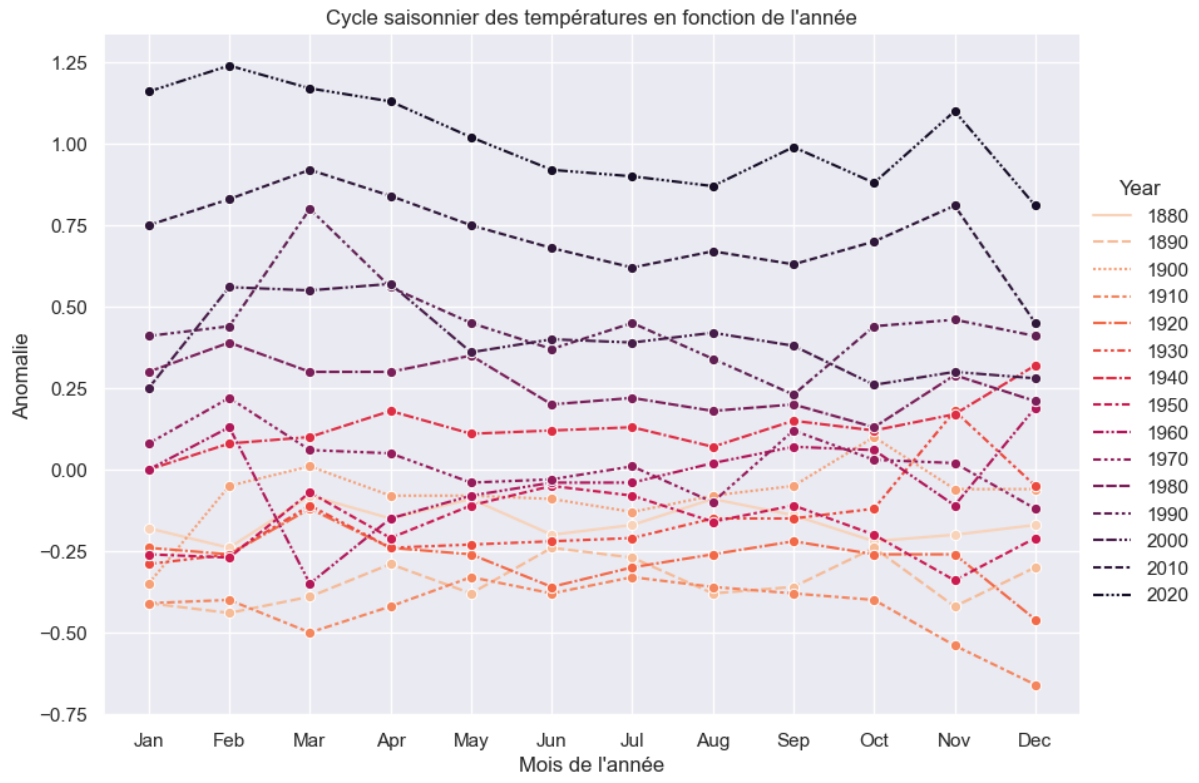


Nous constatons que la variation des températures est à la hausse depuis 1880, c'est-à-dire la fin de la Révolution Industrielle.. Après la fin de cette période et la fin de l'utilisation massive du charbon, la température a légèrement diminué;

mais à partir de 1910 la température a suivi une courbe légèrement exponentielle à la hausse.
Cette tendance haussière s'est accentuée à partir des années 1970. La crise COVID a légèrement influé sur les températures à court terme mais sur le long terme, la tendance reste la même.
Cette augmentation est homogène et suit la même évolution quelque soit la zone et le mois de l'année.
Nous constatons également que cette hausse est plus élevée dans l'hémisphère nord que dans l'hémisphère sud.
Et enfin nous constatons que c'est aux pôles que les variations de températures sont les plus importantes et c'est au pôle Nord que les températures ont le plus augmenté.

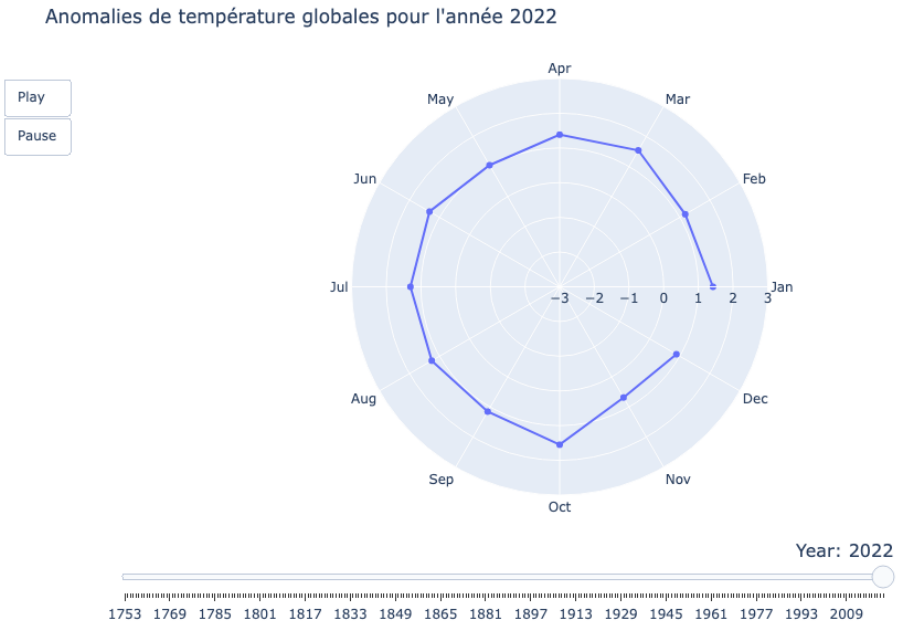
Évolution des températures mensuelles:

Nous avons étudié les températures mensuelles présentes dans le dataset GLB.Ts+dSST.csv. En filtrant le dataset pour ne garder que les années se terminant par 0, nous obtenons ce graphique:



Avec cette représentation, nous constatons que les températures ont sensiblement augmenté sur tous les mois de l'année depuis 1880 et que cette tendance s'est renforcée ces 20 dernières années.

En utilisant les données de Berkeley, nous obtenons le graphe suivant:



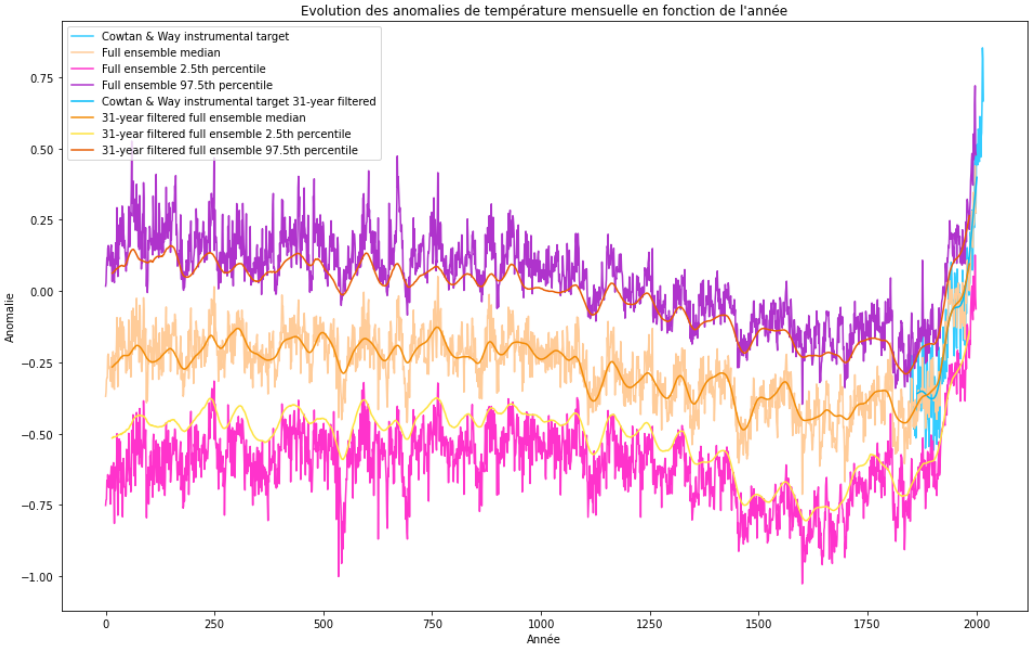
Nous constatons que en 2022, les anomalies de températures sont sensiblement identiques sur tous les mois de l'année et que ce phénomène est donc globale dans le temps.

Conclusion:
A partir de l'analyse et de la visualisation des données, nous démontrons que la hausse des températures est apparue à la fin de la Révolution Industrielle, qu'elle s'est accentuée à partir des années 1970, quel que soit le mois de l'année.
Les pays de l'hémisphère Nord sont les plus impactés car on trouve dans cette zone les pays les plus industrialisés. Le pôle Nord est l'endroit le plus impacté par le changement climatique.

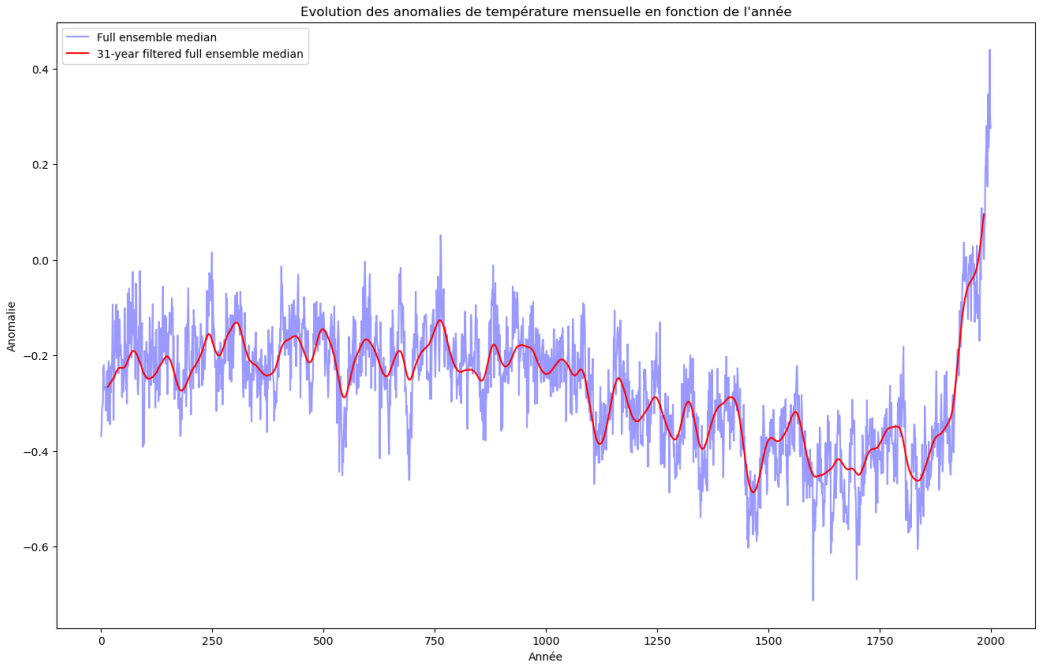
b) Analyse des anomalies de température depuis 2000 ans

Pour cette étude, nous utilisons les données de PAGES2k.

Visualisation avec toutes les données disponibles:



Visualisation avec uniquement les valeurs de la Full Ensemble Median ainsi que la moyenne filtrée sur une période de 31 ans (31-year filtered full ensemble median):



Nous constatons que la courbe est assez plate pendant le 1er millénaire, avec une légère hausse vers la période 1000-1200 puis une baisse allant d'environ 1500 à 1800, puis une montée très prononcée à partir du 20 ème siècle.

Les données montrent que la période moderne est très différente de ce qui s'est passé dans le passé. La période chaude médiévale et le petit âge glaciaire souvent cités sont des phénomènes réels, mais petits par rapport aux changements récents.

Le réchauffement au cours des 50 dernières années est brutal par rapport aux variations qui se sont produites naturellement au cours des 2000 dernières années.

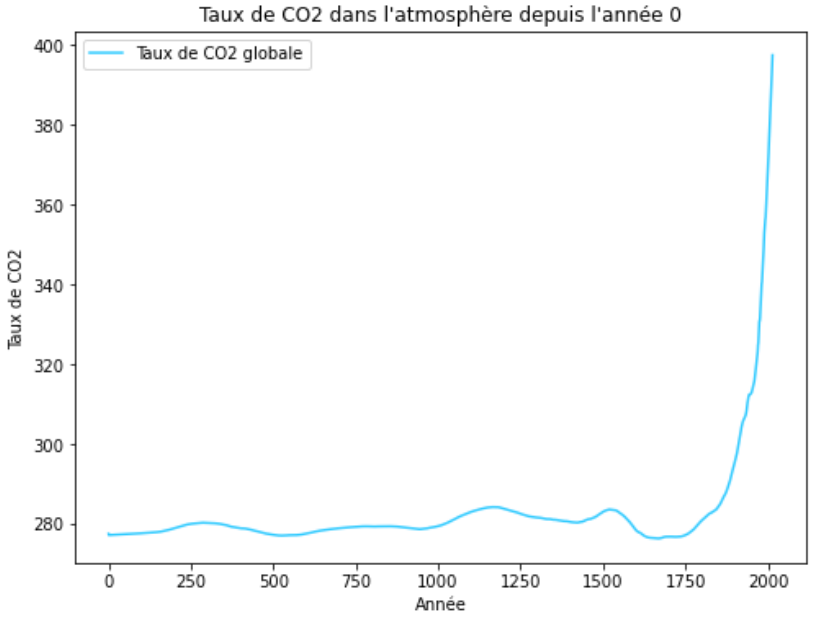
Conclusion:
L'analyse de cette courbe par rapport aux données de Gistemp et Berkeley Earth nous conforte dans l'idée que cette hausse des températures est apparue uniquement à l'époque moderne et n'est pas liée à un cycle qui se serait déjà produit dans le passé.

c) Analyse du CO2 dans l'atmosphère

Pour cette analyse, nous utilisons les données du Climate & Energy College

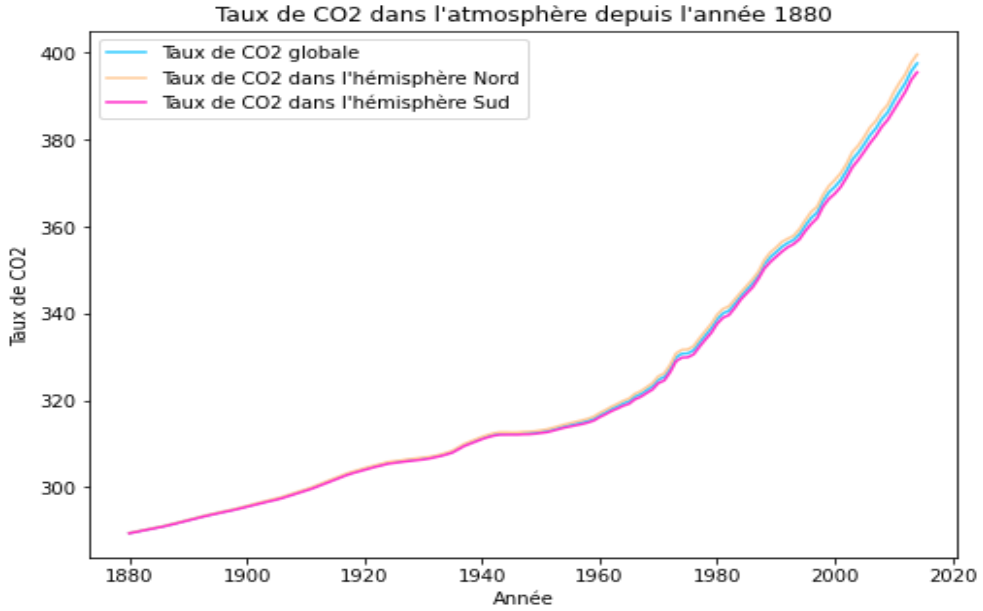
Après traitement des données nous faisons apparaître 2 graphiques:

Le 1er graphique montre le taux de C02 globale dans l'atmosphère depuis l'année 0. Ce taux était assez stationnaire mais nous constatons que depuis le début du 19ème siècle et l'avènement de la révolution industrielle le taux a très fortement augmenté.



Le 2ème graphique est un zoom sur la période 1880 à 2014, période similaire à notre jeu de données sur les anomalies de températures.
Ce graphique nous montre que le taux à très largement augmenté depuis 1880

avec une tendance haussière encore plus prononcée depuis 1960. Nous constatons également que le taux de CO2 est légèrement plus élevé dans le l'hémisphère Nord que dans le l'hémisphère Sud et que cet écart à tendance à augmenter au fil du temps.



Conclusion:

Le taux de C02 dans l'atmosphère suit les mêmes tendances que les anomalies de C02. Ce n'est qu'à partir de la fin du XIXème siècle que le taux de CO2 à commencer à fortement augmenté, porté par la Révolution Industrielle et le développement des pays industrialisés.
Ces émissions de CO2 n'ont rien d'égal avec ce que la terre a connu depuis 2 millénaires.

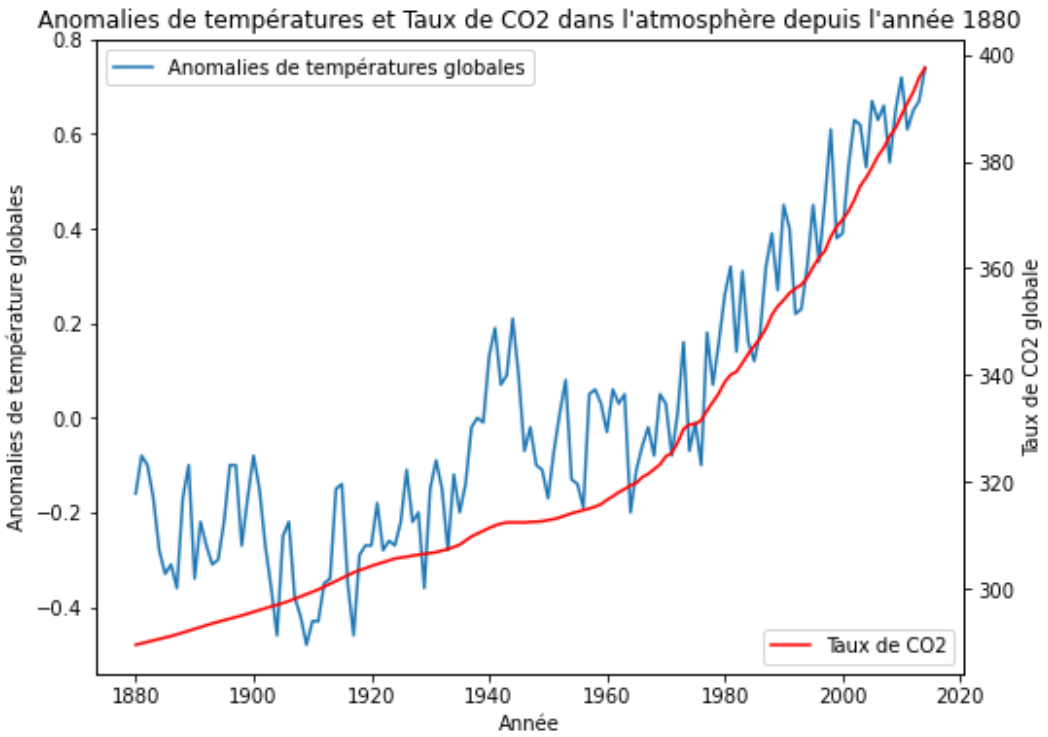
Nous allons donc nous pencher un peu plus sur la relation entre anomalies de températures et émissions de C02.

d) Comparaison avec les anomalies de température et les émissions de C02

Dans ce graphique, nous superposons le taux de CO2 et les anomalies de température sur la période de 1880 à nos jours en mergeant ce dataset avec celui des températures globales de la Nasa (GLB.Ts+dSST)

Nous constatons que les 2 courbes suivent la même tendance haussière avec la même inflexion vers le haut depuis le milieu du XXème siècle.
Il y a donc de très fortes probabilités que ces 2 paramètres soient liés et que la hausse du taux de gaz à effet de serre dans l'atmosphère contribue à faire augmenter la température à la surface du globe.

Nous décidons d'utiliser une méthode statistique pour montrer la corrélation entre la hausse des températures et l'augmentation des émissions de CO2
Nous avons donc effectué un test de corrélation de Pearson entre le taux de CO2 globale et les anomalies globales de température.



Nous obtenons les résultats suivants:

- Coefficient de corrélation = 0.9276260482581129
- pvalue = 9.717225418853804e-59

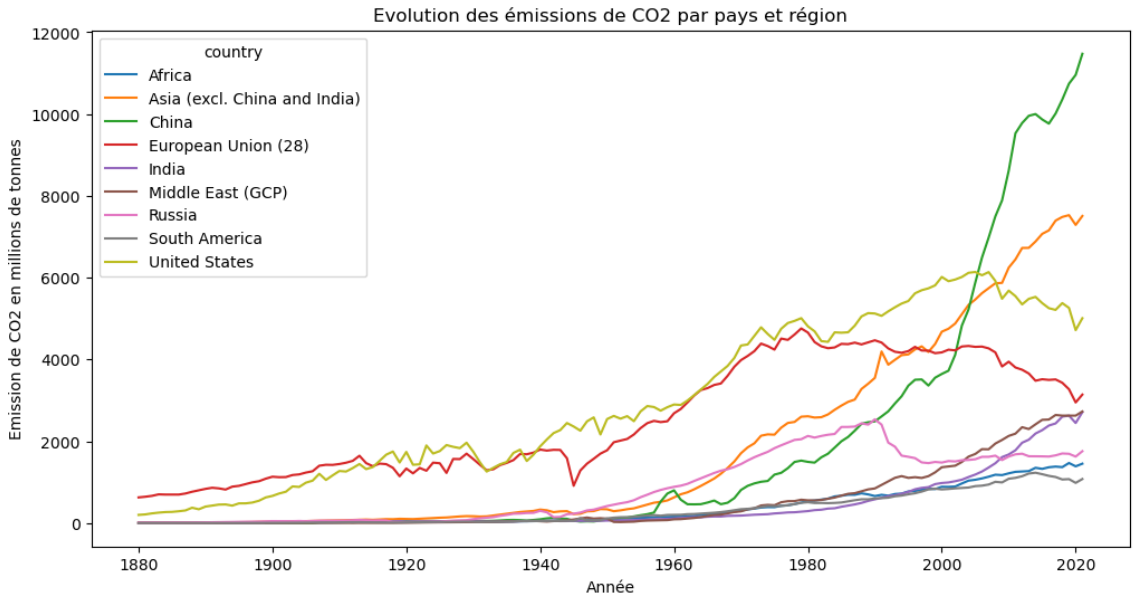
Conclusion:
Les résultats prouvent qu'il existe une forte corrélation positive et statistiquement significative entre le taux de CO2 et les anomalies de température.

e) Analyse des émissions de C02 par pays

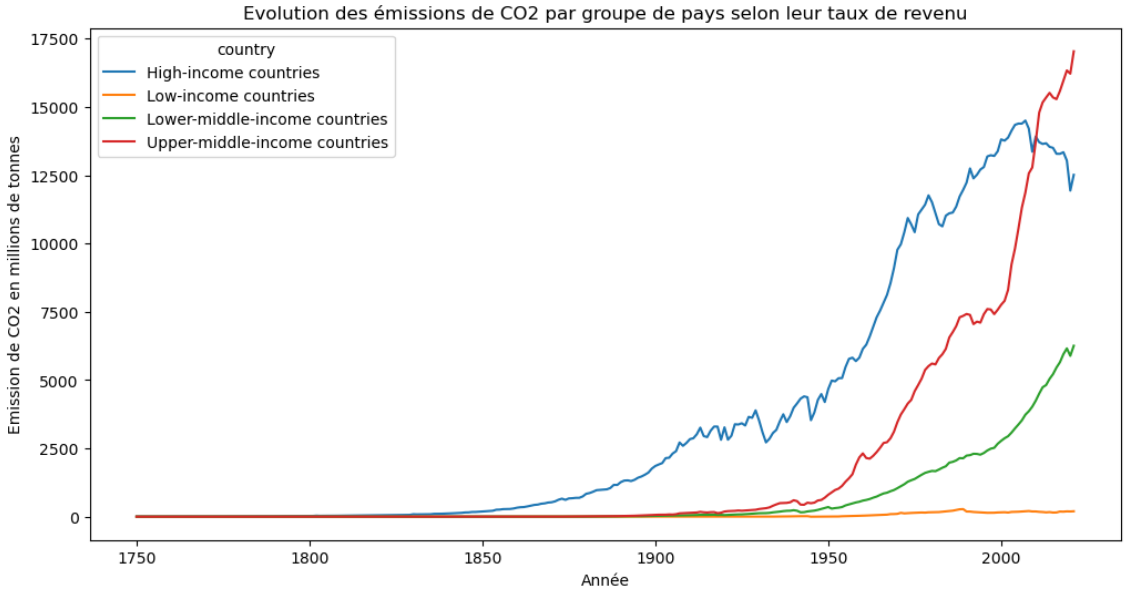
Pour cette analyse, nous utilisons les données Our World in Data avec notre dataset co2_par_pays.csv

Ce premier graphique montre que l'évolution des émissions de CO2 dans les grands pays industrialisés et les grands pôles d'activités n'est pas la même. Dans les pays d'Europe et les Etats-Unis, les émissions de CO2 ont globalement reculées, contrairement à la Chine ou l'Asie.

Si on prend l'exemple de l'Union Européenne, on voit dans la courbe que les émissions de CO2 ont baissé en 1945, après la 2ème Guerre Mondiale, puis de nouveau suite aux choc pétroliers de 1973 et 1979. Depuis, la courbe nous montre que les émissions ont tendance à baisser, et particulièrement pendant la période de la pandémie de COVID-19.
Depuis 2020, l'ensemble des émissions à tendance à repartir à la hausse.

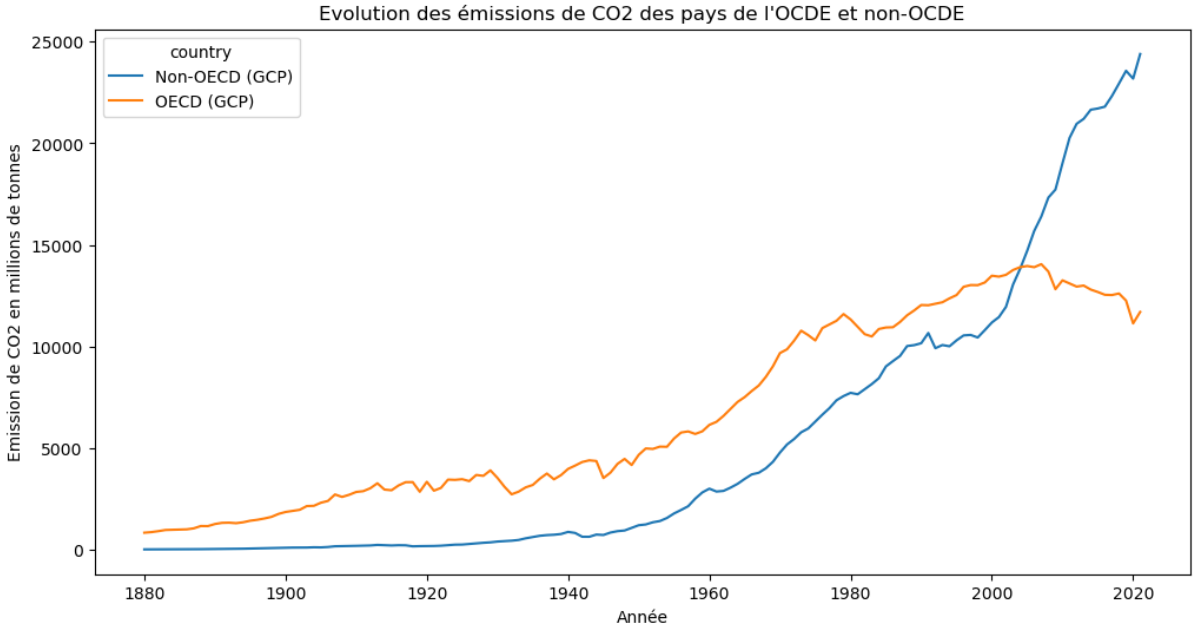


Le deuxième graphique montre l'évolution des émissions en fonction du revenu moyen de chaque pays.
4 groupes ont été créés afin de différencier les pays avec un taux de revenu élevé et ceux avec un taux de revenu faible.
Nous constatons que les 4 courbes ont une tendance haussière mais qu'elles sont différentes..
Pour les pays avec un taux de revenu élevé (High-income countries), les émissions sont plus importantes, mais depuis 2008 et le choc économique les émissions ont baissé.
Pour les pays avec un plus faible taux de revenu (Upper-middle-income et lower-middle-income countries), les émissions étaient très faibles avant 1945 mais elles ont très fortement augmenté depuis, sans pour autant atteindre celle de l'autre groupe de pays.
On constate la même inflexion sur les 4 courbes lors de la crise du COVID-19.



Le 3ème graphique montre l'évolution des émissions divisés en 2 groupes : Le 1er groupe comprend les pays appartenants à l'OCDE (Organisation de coopération et de développement économiques), organisation qui regroupe 38 pays industrialisés, principalement de l'hémisphère Nord (sauf la Chine).
(https://www.oecd.org/fr/apropos/document/ratification-convention-ocde.htm)
Le 2ème groupe comprend tous les autres pays.

L'évolution des émissions des membres de l'OCDE est très similaire à celle des pays ayant le plus de revenus, vu dans le graphe précédent.
Elle suit une tendance haussière jusqu'en 2008 puis à tendance à baisser.
Alors que la courbe des pays non-membres suit une tendance haussière exponentielle, portée principalement par les émissions de la Chine.
On constate que les 2 courbes se croisent au début des années 2000 et ce sont désormais les pays non-membres de l'OCDE qui sont les plus émetteurs de CO2.



Conclusion:

Cette étude des émissions de CO2 nous montre que les anomalies de température sont très fortement liées à la hausse des émissions de CO2 au niveau mondial.
On constate que des pays ont fait des efforts pour réduire leurs émissions de CO2 mais la montée en puissance de la Chine et des pays en voie de développement a contrebalancé ce phénomène.
La hausse des émissions est croissante et ce phénomène va surement impacter sur les anomalies de température et va provoquer une hausse de la température globale de la Terre.

Nous avons également travaillé sur des visualisations interactives que vous pourrez retrouver dans notre projet Streamilt lors de la présentation officielle.
Pour ces visualisations, nous avons utilisé les librairies Plotly et Dash.

B - Modélisation et prédiction à l'aide de techniques de Machine Learning

Pour aller plus loin dans notre étude, nous décidons d'utiliser des techniques de Machine Learning avec de prédire les évolutions de températures sur plusieurs dizaines d'années.

Afin de trouver les meilleures prédictions possibles, nous avons testé 3 modèles différents, pour ainsi voir quel était le plus performant.

Pour chacun de modèles nous avons utilisé les données de Berkeley avec notre jeu de données: températures_globales.csv

Nous avons vu plus haut que l'incertitude des anomalies de températures était pratiquement faible à partir des années 1960 et nous avons décidé d'utiliser les données seulement à partir du 1er Janvier 1960.

1. Machine Learning avec régression polynomiale

a) Préparation des données

Nous avons procédé à quelques transformations sur les données.
Tout d'abord, nous avons transformé au format DateTime les variables temporelles pour faciliter leurs utilisations.

Puis nous avons séparé les données Test et Train en fonction de la date.
Données Train : du 01/01/1960 au 31/12/2009
Données Test : du 01/01/2010 au 30/11/2022

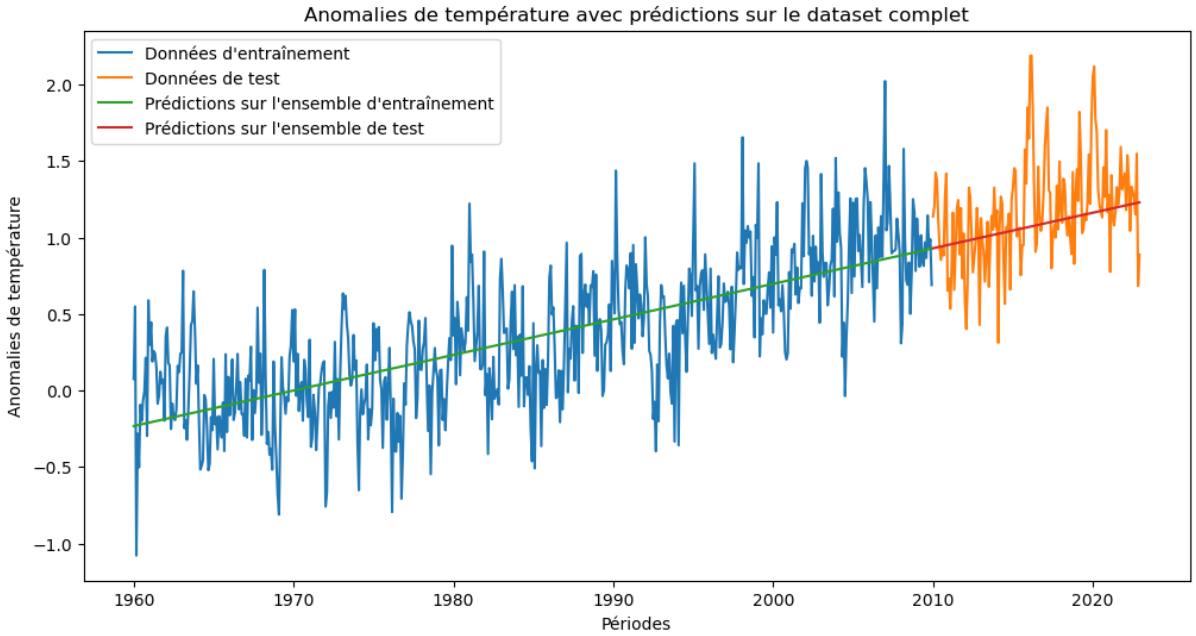
Nous avons utilisé la fonction GridSearchCV pour trouver les meilleurs paramètres à notre régression. Le meilleur hyperparamètre que nous avons obtenu est le Polynomial Features Degree d'ordre 1.

Nous avons utilisé ces paramètres pour le calcul de notre prédiction.

b) Prédiction

Nous avons procédé à 2 tests:
- Prédiction sur les données d'entraînement
- Prédiction sur les données de test

Nous obtenons le graphique suivant:



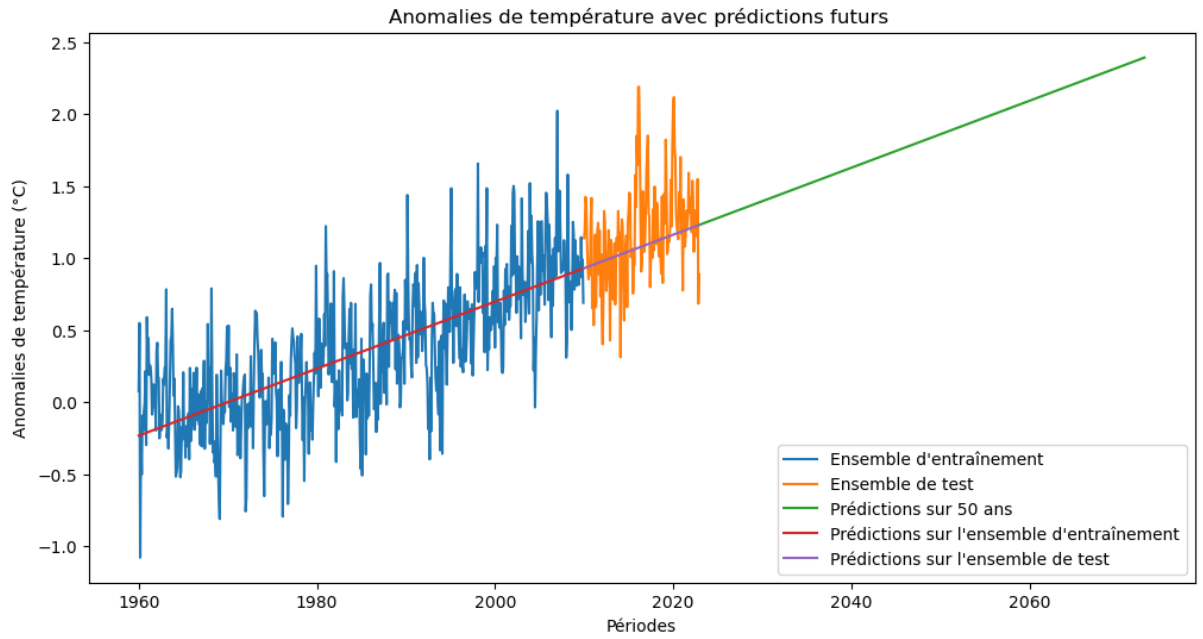
Nous avons procédé au calcul des métriques de performance pour les 2 modèles:

Performances du modèle sur l'ensemble d'entraînement :
MAE : 0.26096916098541445
MSE : 0.11189778207153495
RMSE : 0.33451125851237795
R^2 : 0.5014505205687894

Performances du modèle sur l'ensemble de test :
MAE : 0.2394088342016738
MSE : 0.10693830147168185
RMSE : 0.3270142221244847
R^2 : 0.5014505205687894

Nous avons ensuite testé notre modèle afin de prédire les anomalies de températures sur une période de prédictions sur 50 ans, c'est-à-dire du 01/12/2022 au 02/12/2072.

Nous obtenons le graphique suivant:



Pour l'ensemble d'entraînement :
MAE (Mean Absolute Error) : 0.26096916098541445
MSE (Mean Squared Error) : 0.11189778207153495
RMSE (Root Mean Squared Error) : 0.33451125851237795
R^2 (Coefficient de détermination) : 0.5014505205687894

Pour l'ensemble de test :
MAE (Mean Absolute Error) : 0.2394088342016738
MSE (Mean Squared Error) : 0.10693830147168185
RMSE (Root Mean Squared Error) : 0.3270142221244847
R^2 (Coefficient de détermination) : 0.5014505205687894

La MAE représente l'erreur moyenne absolue entre les prédictions et les valeurs réelles. Dans les deux cas, la MAE est relativement faible, ce qui indique que le modèle est capable de prédire avec précision les anomalies de température pour les données d'entraînement et de test.

La MSE représente la moyenne des carrés des erreurs entre les prédictions et les valeurs réelles. La MSE est également relativement faible pour les deux ensembles de données, ce qui indique que les erreurs sont faibles.

Le RMSE est simplement la racine carrée de la MSE. Il mesure également l'erreur moyenne entre les prédictions et les valeurs réelles, mais donne plus de poids aux

erreurs les plus importantes. Le RMSE est relativement faible dans les deux cas, ce qui suggère que le modèle est capable de prédire avec précision les anomalies de température.

Le coefficient de détermination (R^2) mesure la proportion de la variance totale de la variable de réponse (anomalies de température) qui est expliquée par le modèle. Dans ce cas, le R^2 est de 0,5014 pour l'ensemble d'entraînement et de test. Cela indique que le modèle explique environ 50 % de la variance totale, ce qui n'est pas très élevé mais peut être considéré comme acceptable.

c) Conclusion

Dans l'ensemble, les métriques de performance suggèrent que le modèle de régression polynomiale est capable de prédire avec précision les anomalies de température pour les 50 prochaines années. Cependant, le faible coefficient de détermination suggère qu'il pourrait y avoir des facteurs non linéaires qui influencent les anomalies de température et qui ne sont pas pris en compte par le modèle. Il est donc important de prendre en compte ces résultats avec précaution et de continuer à améliorer le modèle en utilisant des approches plus sophistiquées.
Avec ce graphique, nous constatons visuellement que ce modèle suggère une hausse d'environ 2,4°C des températures dans plus de 50 ans.

2. Machine Learning avec Random Forest

a) Préparation des données

Nous avons procédé à quelques transformations sur les données.
Tout d'abord, nous avons transformé au format DateTime les variables temporelles pour faciliter leurs utilisations.

Puis nous avons séparé les données Test et Train en fonction de la date.
Données Train: du 01/01/1960 au 31/12/2011
Données Test: du 01/01/2012 au 30/11/2022

Nous avons utilisé la fonction GridSearchCV pour trouver les meilleurs paramètres à notre régression. Les meilleurs hyperparamètres que nous avons obtenu sont:
best_max_depth : 2
best_n_estimators : 10
best_min_samples_split : 2
best_min_samples_leaf : 2

Nous avons utilisé ces paramètres pour le calcul de notre prédiction.

b) Prédiction

Nous avons procédé à 2 tests:

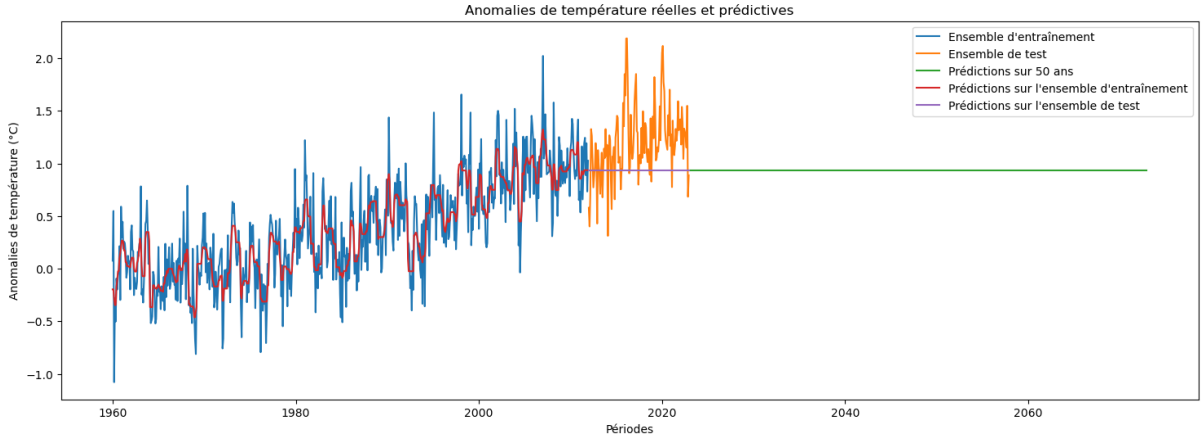
- Prédiction sur les données d'entraînement
- Prédiction sur les données de test

Performances du modèle sur le jeu d'entraînement :
MAE : 0.24692529955849643
MSE : 0.10031142438976218
RMSE : 0.31671978844044807
R^2 : 0.5722886743969964

Performances du modèle sur le jeu de test :
MAE : 0.37694403321137093
MSE : 0.21890709453607013
RMSE : 0.4678750843292151
R^2 : -0.8937430596524423

Nous avons ensuite testé notre modèle afin de prédire les anomalies de températures sur une période de prédictions sur 50 ans, c'est-à-dire du 01/12/2022 au 02/12/2072.

Nous obtenons le graphique suivant:



c) Conclusion

Malgré de bonnes prédictions sur l'ensemble d'entraînement, le modèle de prédiction en Random Forest n'a pas été capable d'apprendre sur de nouvelles données et de prédire sur des données qu'il ne connaît pas, nous émettons l'hypothèse que le modèle a besoin de plus de variables explicatives en entrée.

Ce modèle n'est donc pas adapté à notre besoin en l'état.

3. Machine Learning avec Facebook Prophet

a)Préparation des données

Nous avons procédé à quelques transformations sur les données.
Tout d'abord, nous avons transformé au format DateTime les variables temporelles pour faciliter leurs utilisations.

Pour l'utilisation du Facebook Prophet, nous devons suivre un cahier des charges précis concernant la préparation des données et les étapes de modélisation pour arriver à un résultat probant.

La variable de temps doit être sous le nom **ds** et l'anomalie sous le nom **y**.
Nous faisons volontairement le choix de diviser les données en 90% pour l'ensemble d'entraînement et de 10% pour l'ensemble de test .

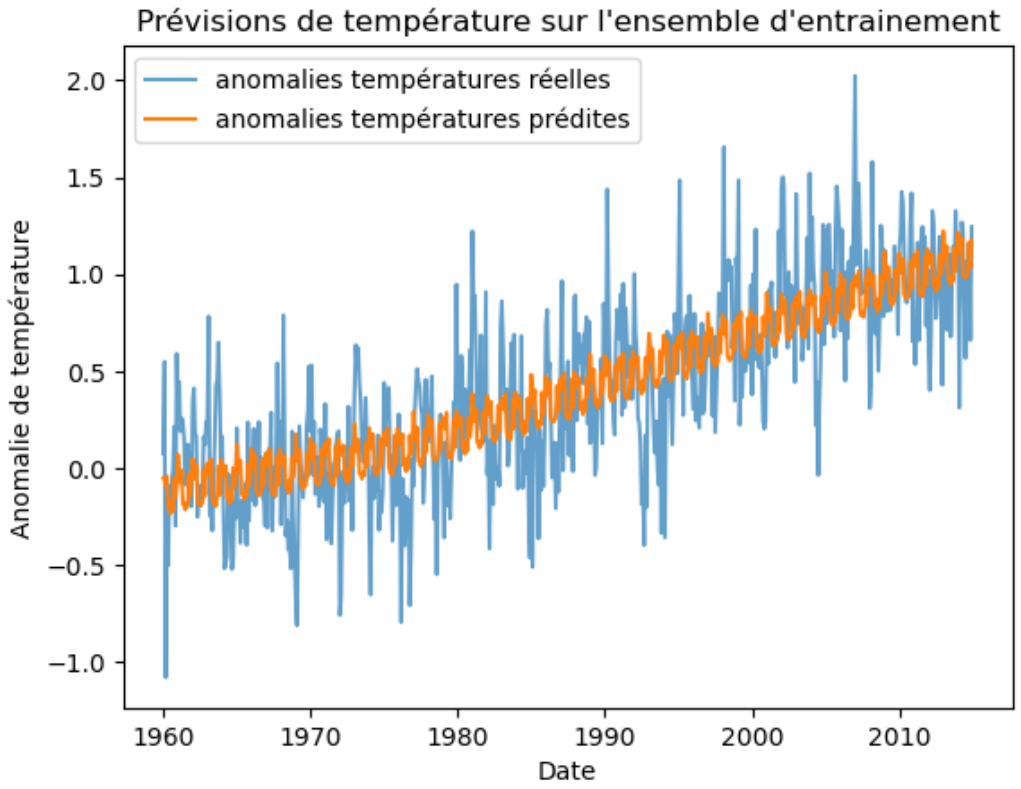
b) Prédiction

Puis nous avons séparé les données Test et Train en fonction de la date.
Données Test: du 01/01/1960 au 31/01/2014
Données Train: du 01/01/2015 au 30/11/2022

Nous créons un modèle Prophet avec des hyperparamètres par défaut : modèle additif
model = Prophet(seasonality_mode='additive')

Puis nous faisons une prédiction sur 50 ans:
future = model.make_future_dataframe(periods=50*12, freq='M')
forecast = model.predict(future)

Nous affichons les résultats suivants:



Notre modèle de prédiction suit fortement la courbe des températures réelles.

Nous obtenons les métriques de performances suivantes:
Métriques de performance du modèle d'entraînement :
MAE : 0.24032665570363626
MSE : 0.09662258273358046
RMSE : 0.31084173261256354
R^2 : 0.601853008358856

Ces métriques peuvent être interprétées comme suit :

MAE (Mean Absolute Error) : il s'agit de la moyenne des écarts absolus entre les valeurs prédites et les valeurs réelles. Plus le MAE est faible, meilleure est la performance du modèle. Dans ce cas, le MAE est de 0.24, ce qui signifie que les prédictions du modèle ont en moyenne une erreur absolue de 0.24 par rapport aux valeurs réelles.

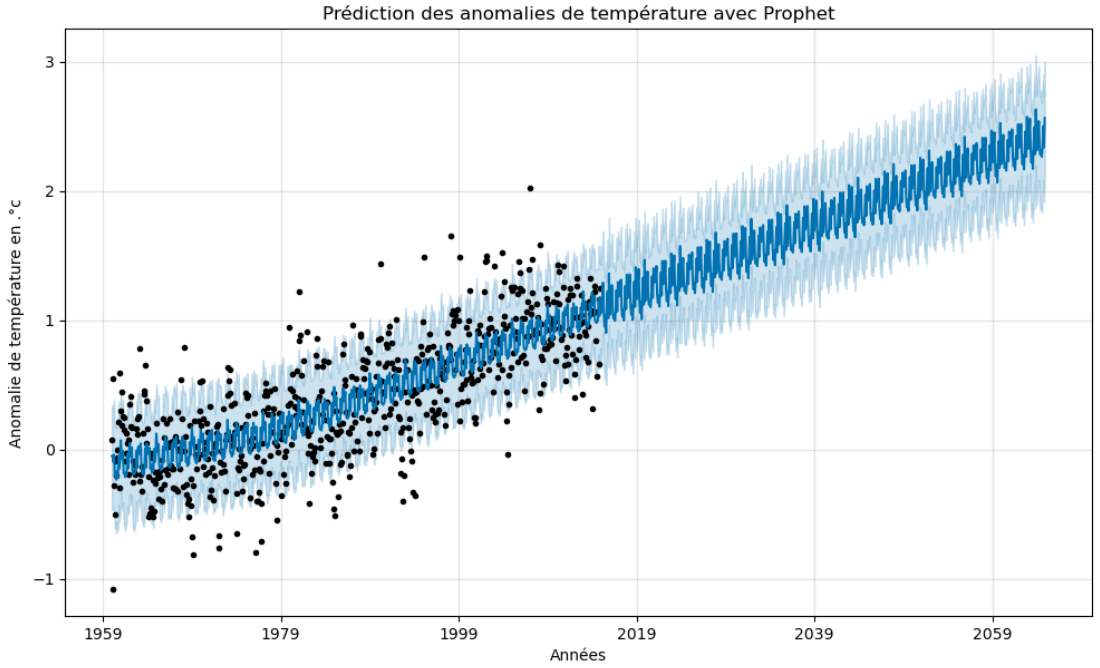
MSE (Mean Squared Error) : il s'agit de la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réelles. Plus le MSE est faible, meilleure est la performance du modèle. Dans ce cas, le MSE est de 0.096, ce qui est relativement faible et suggère que le modèle fait des prédictions précises.

RMSE (Root Mean Squared Error) : il s'agit de la racine carrée du MSE. Le RMSE est une mesure de la dispersion des erreurs. Plus le RMSE est faible, meilleure est la performance du modèle. Dans ce cas, le RMSE est de 0.31, ce qui est relativement faible et suggère que le modèle est capable de faire des prédictions précises.

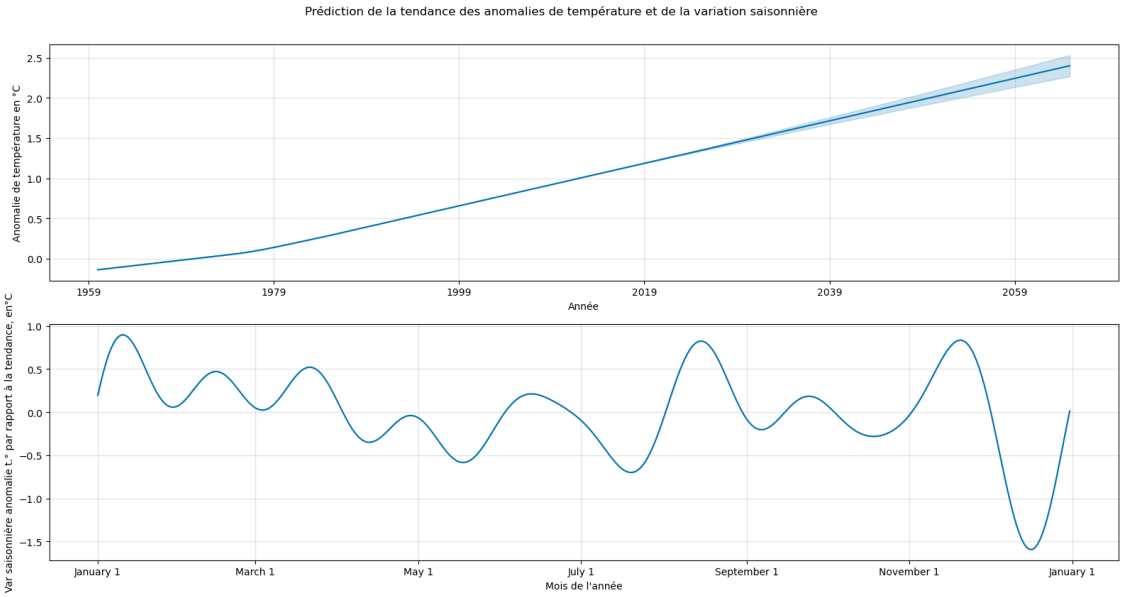
R^2 (R-squared) : il s'agit d'une mesure de l'ajustement du modèle aux données. R^2 varie entre 0 et 1, où 1 indique un ajustement parfait. Dans ce cas, le R^2 est de 0.6, ce qui indique que le modèle explique environ 60% de la variance des données. Bien que ce ne soit pas un ajustement parfait, un R^2 de 0.6 est considéré comme un bon ajustement pour de nombreux problèmes de prédiction.

Les métriques de performance du modèle Prophet suggèrent que le modèle est capable de faire des prédictions précises pour les anomalies de température dans les 50 années à venir. Cependant, il est important de noter que ces métriques ne fournissent qu'une indication de la performance du modèle et qu'il est toujours important de vérifier la qualité des prédictions à l'aide d'autres outils et méthodes.

En calculant les prédictions sur 50 ans nous obtenons le graphique suivant:



En affichant les composantes de la prédiction, c'est à dire la tendance et la saisonnalité, nous obtenons les graphiques suivants:



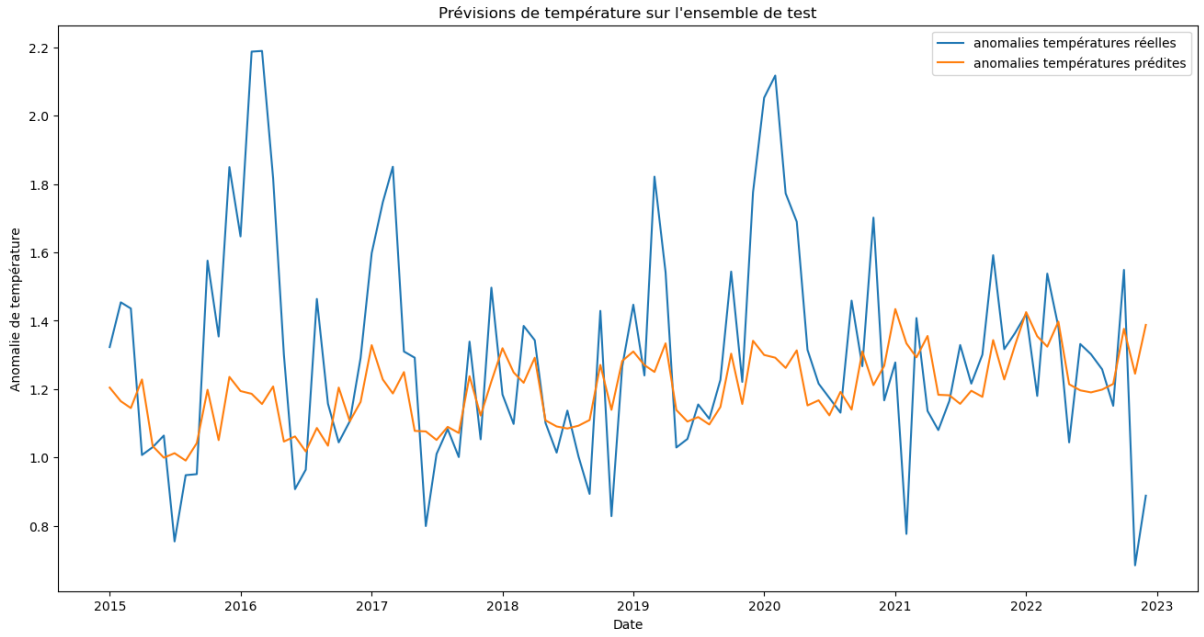
On observe une tendance à la hausse mais aussi une saisonnalité avec des pics sur mi-août et le mois de décembre où l' on constate qu'il fait plus chaud que la moyenne des autres mois dans l'année.

Nous mesurerons la performance du modèle en comparant les prévisions aux valeurs réelles de l'ensemble de test et nous obtenons les métriques suivantes:

Métriques de performance du modèle prophet :
MAE : 0.21657790185881462
MSE : 0.0947528066184659
RMSE : 0.30781943833758435
R^2 : 0.03478913247595228

Ces métriques indiquent que le modèle Prophet n'a pas bien été généralisé sur les données de test, car les valeurs des métriques de performance sont relativement élevées. En particulier, le MAE et le RMSE indiquent que le modèle fait des prédictions en moyenne à environ 0.22 de l'écart absolu et 0.31 de l'écart type de la véritable valeur, respectivement. De plus, le R^2 est très faible, ce qui signifie que le modèle n'explique qu'une très petite proportion de la variance dans les données de test.

Nous affichons les prévisions pour l'ensemble de test:



d) Conclusion

'''Selon l'Organisation météorologique mondiale (OMM), l'année 2016 restera dans les annales: température moyenne record, banquise exceptionnellement réduite et poursuite inexorable de la hausse du niveau de la mer et du réchauffement des océans... Ce compte rendu confirme que l'année 2016 est la plus chaude qui ait jamais été enregistrée: la hausse de la température par rapport à l'époque préindustrielle atteint, chose remarquable, 1,1 °C, soit 0,06 °C de plus que le record précédent établi en 2015. Cette augmentation de la température moyenne s'inscrit dans la logique des autres changements intervenant dans le système climatique», a souligné le Secrétaire général de l'OMM, Petteri Taalas.''' Ces commentaires données par l'OMM explique cet écart flagrant entre la prédiction pour 2016 et l'anomalie de température mesurée .

2017 est la 2e année la plus chaude sur Terre, d'après NCEP-NCAR

2020 est en passe de devenir l'une des trois années les plus chaudes jamais enregistrées

Les températures extrêmes que l' on a eu en 2016-2017-2020 montrent les limites de notre modèle. Le modèle ne peut pas les anticiper. Sur une période de 9 ans , on a 3 années exceptionnelles en termes de record de température qui impact donc la performance de n.

On peut prévoir une tendance sur l'avenir mais il est difficile de prévoir des années qui battent des records !

Néanmoins, le modèle Facebook Prophet nous a permis de déterminer que la

hausse des températures allait se poursuivre, et la prévision est d'environ +2,6°C dans 50 ans.

4. Conclusion

Après avoir testé 3 modèles de Machine Learning, nous arrivons à la conclusion que les températures vont continuer à augmenter.
Chaque modèle nous a prédit une hausse, avec plus où moins de fiabilité.

D'après notre étude, c'est le modèle Facebook Prophet qui est le plus précis en termes de prédiction, principalement dû au fait que ce modèle est très adapté au Time Series.

Nos conclusions concordent avec les résultats du dernier rapport du GIEC et leurs prévisions.

C - Bilan

Grâce à l'analyse de données, nous avons pu répondre aux questions de notre problématique et atteindre nos objectifs.

Nous avons confirmé que le réchauffement climatique est bien réel.
Nous avons montré que l'augmentation des températures est apparue avec la révolution industrielle de la fin du XIXème siècle, phénomène qui s'est accéléré depuis les années 1970.
Tous les pays sont touchés par cette hausse des températures et ce sont principalement les pays de l'hémisphère Nord qui sont impactés.
L'augmentation des émissions de CO_2 sont très fortement corrélées avec l'augmentation des températures à la surface de la Terre.
Les prédictions réalisées grâce au Machine Learning nous montre que les températures vont continuer à augmenter dans les prochaines années et que nous aurons plus de 2°C d'augmentation dans 50 ans.

Le réchauffement climatique est un phénomène global de transformation du climat qui a de nombreuses conséquences:

Une augmentation des températures à cause du réchauffement climatique affecte l'ensemble de l'écosystème mondial et pas seulement la chaleur ressentie. La météo s'en trouve perturbée, avec une augmentation des phénomènes météorologiques extrêmes, des changements des modèles météorologiques habituels. Cela veut dire plus de tempêtes, plus d'inondations, plus de cyclones et de sécheresses.

La capacité de régulation des océans est aussi affectée par une augmentation des températures. Si les températures globales augmentent de façon très importante, il y aura donc augmentation des niveaux des océans, mais aussi une acidification et une désoxygénation des zones océaniques. En outre, une acidification des océans trop prononcée pourrait limiter la capacité des mers de la planète à produire de l'oxygène et à stocker le CO2, et donc augmenter encore le réchauffement climatique. Mais cela peut aussi affecter des zones de forêts et les écosystèmes fragiles (barrière de corail, forêt amazonienne) ainsi que la biodiversité (les coraux, certains insectes et même des mammifères pourraient ne pas survivre).

Sur la société et l'économie, le réchauffement climatique peut avoir potentiellement plusieurs conséquences : la capacité des sociétés à s'adapter à un nouveau climat, à adapter leurs infrastructures, notamment médicales, mais aussi leurs bâtiments. Le réchauffement climatique aura aussi des conséquences sur la santé publique, la capacité alimentaire des pays...
Il est temps de redescendre sur Terre et de faire les efforts nécessaires pour endiguer le réchauffement climatique.

D - Suite du projet

Dans notre étude, nous avons axé nos analyses sur les anomalies de températures et leurs corrélations avec les émissions de C02. Nous aurions pu tenter de faire apparaître d'autres corrélations avec les activités humaines (Transport, agriculture...).
Par ailleurs, nous pourrions montrer l'influence du changement climatique sur de nombreux axes tels que la biodiversité, l'espérance de vie, l'augmentation du niveau des eaux, l'augmentation du nombre d'évènements climatiques...
Afin d'améliorer nos prédictions et d'affiner les résultats des modélisations, d'autres modèles de Machine Learning peuvent être utilisés.

E - Ressources

https://www.ipcc.ch/languages-2/francais/

https://berkeleyearth.org/data/

https://data.giss.nasa.gov/gistemp/

https://ourworldindata.org/

https://www.climato-realistes.fr/climat-biais-et-erreurs-mesure-temperatures-globales/