DataScientest • com

Rapport Technique d'évaluation

UnhapPy Earth

Étude de cas : Analyse du réchauffement climatique

Participants :

- Olga Fedorova
- Boris Baldassari
- Nicolas Cristiano

Mentor projet : Sébastien Sime

Chef de cohorte : Antoine Tardivon

Promotion : Data Analyse / formation continue mars 22

Introduction

Contexte

En quelques décennies à peine, la question du réchauffement climatique est devenue un sujet majeur, inquiétant pour l'avenir de la planète et de sa biodiversité, y compris l'espèce humaine. Pourtant il fait débat, oppose experts et climato-sceptiques, à l'heure même où les manifestations et les conséquences de ce changement sont chaque jour plus flagrantes.

Ce projet nous a amenés à utiliser la plupart des compétences acquises au cours de notre formation :

- Acquisition et manipulation de jeux de données ;
- Production de visualisations graphiques afin de "faire parler" ces données, et éventuellement, susciter des intuitions ;
- Construction d'algorithmes de régression et de prédiction.

Le réchauffement climatique est incroyablement coûteux : humainement, socialement, géopolitiquement, financièrement et économiquement. Sur la dernière décennie, le coût du réchauffement climatique estimé par un rapport des Nations Unies datant de Avril 2022 est de 140 Milliards par an, et devrait atteindre 2000 Milliards d'ici 2030 – voir l'article. Anticiper et mieux comprendre ses conséquences, donc le coût économique, devient donc une priorité pour beaucoup d'entreprises et d'organisations.

Nous voulons savoir si, grâce à la Data Analyse, nous étions en mesure d'apporter des conclusions se rapprochant de celles données par les experts en climatologie. Nous voulons mettre à disposition du public un moyen simple et vérifiable, au travers de sources factuelles et ouvertes, de constater le réchauffement climatique.

Objectifs

Notre objectif est de proposer, à l'aide de données fiables et librement accessibles, une analyse du réchauffement climatique, en répondant aux problématiques suivantes :

1. Pouvons-nous confirmer grâce aux données le phénomène de changement climatique ? Le réchauffement est-il réellement observable ?
2. Quand le phénomène apparaît-il ? De manière soudaine ou graduelle ? Au même moment sur l'ensemble du globe ?
3. Son évolution est-elle uniforme à travers le monde ou certaines zones sont-elles plus impactées que d'autres ?
4. D'après la grande majorité des experts en climatologie, les émissions de CO2 dans l'atmosphère seraient, parmi l'ensemble des gaz à effet de serre, la principale cause du réchauffement de la planète. Quel degré de corrélation existe-t-il entre les émissions de CO2 et l'évolution des températures ?
5. Des événements historiques ont-ils eu un impact sur l'évolution de la température ?
6. Quelles sont nos prédictions de températures sur les prochaines années ?

7. Par cette Data Analyse, parvenons nous à des conclusions similaires à celles des scientifiques - climatologues ?

Nous avons tous entendu parler du réchauffement climatique, de ses détracteurs et des débats sur son existence. A l'heure où ses conséquences (sécheresses, pluies torrentielles à répétition) se font de plus en plus sentir, nous avons souhaité identifier les éléments factuels disponibles et faire notre propre analyse, sans avoir pour aucun d'entre nous de connaissance préalable en climatologie. Nous espérons qu'en produisant un travail reproductible et librement accessible ces informations pourront être réutilisées par d'autres personnes curieuses et non spécialistes.

Nous avons lu nombre d'articles et de publications disponibles sur internet. Ceux-ci nous ont permis de mieux comprendre pourquoi et comment sont ainsi construits les jeux de données étudiés, ainsi que les enjeux relatifs aux problématiques citées.

Classification du problème

Grâce aux outils classiques de Data Analyse et de Dataviz, nous allons explorer, nettoyer, fusionner, visualiser et analyser nos données, et ainsi pouvoir répondre à nos problématiques **1**, **2**, **3**.

Deux algorithmes de machine learning nous permettront de répondre aux problématiques **4** et **6** :

- Une régression linéaire afin d'étudier la relation entre les quantités de CO_2 émises et l'évolution des températures. Celle-ci permet d'estimer si deux variables évoluent ensemble, et sera ensuite évaluée par une métrique de performance.
- Un modèle prédictif afin de proposer une prévision de l'évolution des températures.

Enfin, c'est plutôt un travail de recherche qui nous aidera à répondre à la problématique n°**5**, et plus globalement, à nous assurer de l'adéquation de nos résultats avec les recherches courantes (point n°**7**).

Librairies utilisées

La totalité du code est en langage Python, utilisant les packages (librairies) et sous-packages suivants :

Package	Sous-package et Fonction spécifiques	Utilisation
pandas		Création et manipulation de DataFrames
numpy		Fonctions mathématiques
matplotlib		Visualisation (graphiques, colorisation)
seaborn		Visualisation (heatmap)
re		Expressions régulières pour extraction de données
GeoPandas		Données géospatiales et représentations cartographiques
pycountry		Base de données ISO (codes pays)
mapclassify		Classification cartographique
SciPy	stats, pearsonr	Tests de corrélation
Scikit-learn	linear_model, LinearRegression	Modèle de régression lineaire
Scikit-learn	metrics, r2_score	Evaluation du modèle de régression linéaire
FB prophet	Prophet	Analyse et prédictions de séries temporelles
warnings		Non-affichage des messages d'avertissement

Identification des sources

Nous avons identifié deux sources de données pour nos recherches : Berkeley Earth et Our world in data. Toutes nos données sont publiquement téléchargeables, sous une licence spécifique pour le premier et CC-BY pour le second.

Berkeley Earth est une organisation américaine indépendante qui fournit des données de température historiques nettoyées et cross-vérifiées. Recoupant elle-même plusieurs sources, notamment l'analyse GISTEMP réalisée sous l'égide de la NASA, elle fournit un ensemble de jeux de données actualisés, les mesures vont jusqu'à 2021, et complets, avec plus de 19 millions d'observations depuis 46 000 stations météo.

Our world in data est un organisme reconnu pour la qualité de leurs publications de données sur un ensemble de sujets d'actualité tels que la pollution, la santé ou la population. Leurs données sont librement téléchargeables sur GitHub, et sont notamment utilisées pour l'enseignement et la recherche à travers le monde. Nous nous appuyons sur leurs données d'émission de CO_2 disponibles dans leur référentiel GitHub.

Données de températures

Les jeux de données de Berkeley Earth sont fournis sous forme de fichiers .txt, chacun d'entre eux correspondant à une région (totalité du globe, par hémisphère ou par pays). La fréquence d'échantillonnage est mensuelle.

Ils sont tous construits de la même manière, avec en en-tête une description complète, incluant un élément crucial : une liste de 12 températures de **référence absolues mensuelles** et leur incertitude sur une période fixe de 30 ans, de janvier 1951 à décembre 1980. Cette période a été retenue d'une part pour la fiabilité et la complétude des observations effectuées, mais également car elle représente une sorte de médiane sur l'ensemble du dataset. Ensuite, un tableau donne le détail des résultats moyens sous forme d'**anomalies**, ou **températures relatives**, assorties de leur incertitude, observées par mois pour la région donnée, pour une période allant de 1750 au plus tôt, et jusqu'en 2021. L'**anomalie de température** est une valeur relative, exprimée en degrés Celsius correspondant à l'écart, positif ou négatif, entre la température mesurée et la température moyenne de référence correspondante. Plus d'informations concernant ce mesurement et l'intérêt de raisonner en termes d'anomalies peut être trouvé sur le site de la NASA. L'**incertitude** est la dispersion liée à différents facteurs, notamment de sous-échantillonnage statistique et spatial, influant *in fine* sur la qualité de la mesure. Elle représente l'intervalle de confiance à 95 %. **Ces références nous permettent de calculer les températures en valeurs absolues.**

Les données sont réparties sur 12 colonnes :

- Les quatre premières colonnes fournissent les informations suivantes : l'année, le mois de l'année, l'anomalie de température moyenne estimée pour ce mois et son incertitude.

- Les huit dernières colonnes rapportent les anomalies et incertitudes sous la forme de moyennes glissantes annuelles, quinquennales, décennales et vicennales, centrées sur le mois considéré. Par exemple, la moyenne annuelle de janvier à décembre 1950 est rapportée à juin 1950, ce qui explique la présence de « NaNs » en début et fin de ces colonnes dans les Data Sets.

Les fichiers que nous utilisons pour nos analyses ont tous été récupérés sur le site de Berkeley Earth, et sont les suivants :

- Températures globales :
  - Complete_TAVG_complete.txt : liste des températures globales moyennes sur terre.
- Températures par hémisphères :
  - northern-hemisphere-TAVG-Trend.txt : liste des températures moyennes par hémisphère (nord)
  - southern-hemisphere-TAVG-Trend.txt : liste des températures moyennes par hémisphère (sud)
- Températures par pays :
  - Sets_by_country : liste des températures moyennes par pays.

Pour le data set contenant les températures pour l'ensemble du globe, les informations de colonnes sont les suivantes :

Nom	Type	NaNs	Description	Exemple	Moyenne	Min	Max
date	datetime	0	Date au format yyyy-mm-dd.	1950-12-24	X	X	X
year	int	0	L'année au format xxxx.	1950	X	X	X
month	int	0	Le mois au format numérique.	120	X	X	X
ano	float	1	Anomalie constatée (°C).	0.122	-0.265	-6.018	5.531
uncert	float	3	Incertitude (°C).	0.041	0.903	0.022	6.521
abs	float	1	Température absolue (°C).	13.124	8.329	-2.423	17.780

Données de dioxyde de carbone

Nous identifions deux jeux de données pour le CO_2 : le premier est la quantité totale produite par pays et par an, l'autre détaille les émissions dues à la production industrielle et à l'utilisation des sols, tous pays confondus et par an.

Les données sont téléchargées depuis le référentiel GitHub de Our World in Data. Le fichier contient les colonnes suivantes :

Nom	Type	NaNs	Description	Exemple	Moyenne	Min	Max
year	int	0	L'année au format xxxx.	1950	X	X	X
iso_code	object	4029	Code ISO du pays.	AFG	X	X	X
country	object	0	Dénomination usuelle du pays.	Afghanistan	X	X	X
co2	float	1319	Quantité de CO_2 émise (millions de tonnes).	5.333	326.658	0	36702.503
gdp	float	12520	Produit Intérieur Brut ($).	3.045e+10	2.89e+11	5.543e+07	1.136e+14
population	float	3097	Population du pays.	4.87e+06	7.068e+07	1490	7.795e+09
total_ghg	float	20338	Quantité totale de gaz à effet de serre émise, en équivalent de CO_2 (millions de tonnes).	33.9	420.52	-178.71	48939.71

Les pays sont identifiés par leur trigramme ISO 3166 - alpha-3, et comme pour les températures, les données peuvent manquer en fonction des années et des pays.

Les **données de CO_2 globales** nous fournissent une autre information : la quantité de CO_2 émise par l'utilisation des terres, en plus de la génération de CO_2 liée à la production d'énergie et industrielle. Les terres, en fonction de leur utilisation, produisent une certaine quantité de CO_2 qui vient s'ajouter à la production industrielle ; dans certains cas elles peuvent aussi en consommer, et nous pouvons avoir des valeurs négatives sur cette mesure. Les colonnes du jeu de données sont les suivantes :

Nom	Type	NaNs	Description	Exemple	Moyenne	Min	Max
Year	int	0	L'année au format xxxx.	1950	X	X	X
Land use emissions (GtCO2)	float	0	Quantité de CO_2 émise par l'utilisation des terres (millions de tonnes).	4.39e+09	2.499e+09	0	6.99e+09

Nom	Type	NaNs	Description	Exemple	Moyenne	Min	Max
Fossil fuel and industry emissions (GtCO2)	float	0	Quantité de CO_2 émise par l'industrie et la production d'énergie (millions de tonnes).	4.17e+09	9.894e+09	1.969e+08	3.67e+10

Lecture et pré-traitement des données

Les données de température sont fournies sous forme de fichiers txt, et doivent être pré-traitées afin d'en extraire les mesures de références et calculer les valeurs de température absolues. Nous avons écrit une fonction Python qui prend soin de lire les fichiers et retourne des datasets Pandas formatés de la même manière, incluant les valeurs d'anomalie et de température absolue.

Nettoyage des données

La plupart des jeux de données lus incluent des NaNs ; nous les identifions et décidons de leur maintien ou de leur remplacement au cas par cas.

- Pour les **températures globales**, il est rare qu'aucun capteur ne soit disponible au niveau planétaire et les quelques NaNs sont présents en tout début de jeu, dans les années 1751/1752. Nous en identifions seulement 3 lignes, non-consécutives, et décidons de les remplacer par interpolation linéaire.
- Pour les **températures par hémisphère**, il y a une seule valeur manquante sur l'hémisphère nord et 2 sur l'hémisphère sud. Nous utilisons encore une fois une interpolation linéaire pour remplacer ces valeurs en s'appuyant sur leur continuité. Il faut noter également que la date de début de collecte est différente entre les deux jeux de données, dès 1840 pour le Nord et à partir de 1880 pour le Sud.
- Pour les températures par hémisphère toujours, nous calculons également une moyenne glissante sur 12 mois, qui permet de lisser les variations saisonnières et d'obtenir des courbes plus lisibles.
- Pour les **températures par pays**, les mesures sont souvent manquantes - lorsque les capteurs ont été installés tardivement, ou lors de périodes d'instabilité géopolitique. Pour ces raisons les NaNs se répartissent sur de grandes plages de temps – correspondant par exemple à la durée d'une guerre ou la mise en autarcie d'un pays. Nous ne pouvons donc utiliser d'interpolation pour les remplacer et décidons de les conserver tels quels.

Les **informations de CO_2 par pays** sont disponibles en format CSV ; comme pour le dataset précédent des températures par pays, les valeurs manquantes sont liées à des périodes où le pays, pour quelque raison que ce soit, n'a pas pu collecter les données. Ces

trous représentent des plages de temps importantes, et nous les laisserons tels quels pour ne pas compromettre l'intégrité statistique des données.

Agrégation des données

Les informations par hémisphère et par pays sont fournies dans des fichiers distincts (par hémisphère dans le premier cas, et par pays dans le second). Nous souhaitons les rassembler afin d'obtenir les données semblables (par hémisphère, par pays) en un jeu de données unique, plus facile à manipuler, analyser et visualiser.

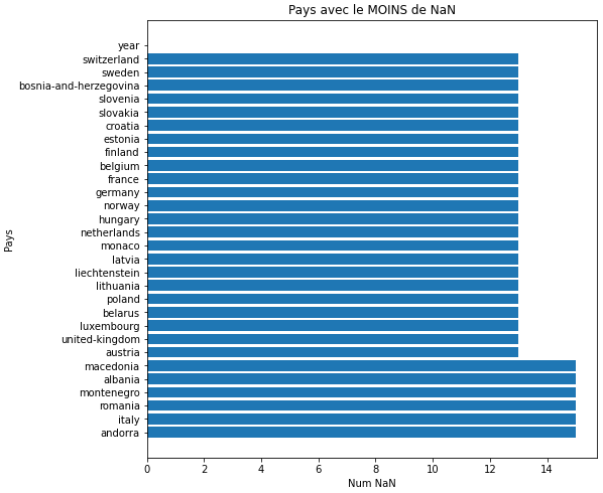
Les **températures par hémisphère** contiennent à la fois les données sud et nord sur la plus grande période possible (i.e. à partir de 1840). Cela introduit un grand nombre de NaNs pour les 40 premières années, les mesures ayant commencé plus tard dans l'hémisphère sud. Comme cela représente une grande plage de temps, nous ne pouvons les remplacer sans introduire de biais et décidons donc de les conserver telles quelles.

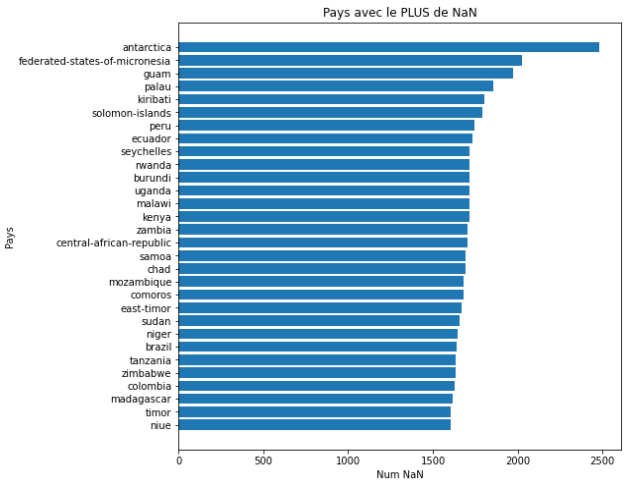
Le jeu de températures produit pour toutes hémisphères confondues est décrit ci-après, et on peut y noter les 487 NaNs présents en début de période :

Nom	Type	NaNs	Description	Exemple	Moyenne	Min	Max
date	datetime	0	Date au format yyyy-mm-dd.	1950-12-24	X	X	X
year	int	0	L'année au format xxxx.	1950	X	X	X
month	int	0	Le mois au format numérique.	120	X	X	X
north_ano	float	0	Anomalie constatée (°C).	0.122	-0.0089	-2.322	2.883
north_abs	float	0	Température absolue (°C).	0.041	10.117	-4.382	22.587
north_uncert	float	0	Incertitude (°C).	13.124	0.282	0.042	2.518
south_ano	float	487	Anomalie constatée (°C).	0.122	0.037	-1.38	1.709
south_abs	float	487	Température absolue (°C).	0.041	16.882	11.522	22.005
south_uncert	float	487	Incertitude (°C).	13.124	0.287	0.064	0.83

Les **températures par pays** rassemblent les données de température de 188 pays, sur la même période de temps (de 1740 pour les premières mesures à 2020). Comme on peut s'y attendre, cela introduit un grand nombre de NaNs pour les pays qui ont commencé la collecte après la date de début du jeu. Les pays les plus complets sont ceux d'où est partie

l'initiative de mesure et sont principalement européens. A l'inverse, les pays en voie de développement, non-habités ou non-vivables (e.g. l'antartique) ont beaucoup moins de mesures disponibles :





Nous utilisons la méthode de pivot de pandas pour les placer horizontalement et les écrire sur le disque :

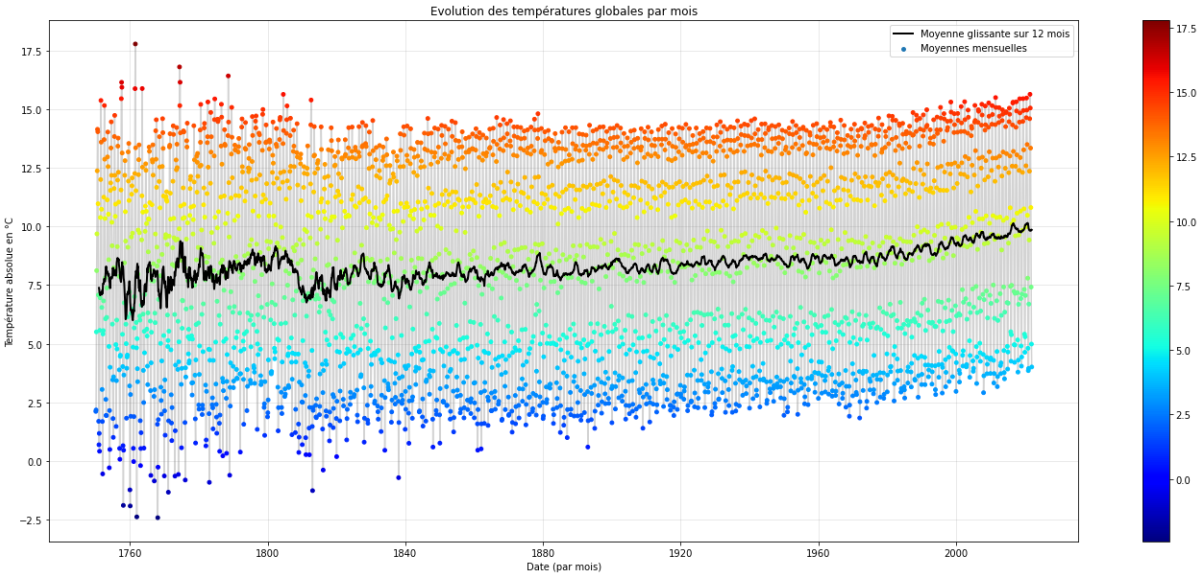
	date	year	afghanistan	albania	algeria	andorra	angola	anguilla	antarctica	argentina	armenia	aruba	australia	austria	azerbaijan	bahamas	bahrain	bangladesh
0	1743-11-15	1743	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1744-04-15	1744	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	1744-05-15	1744	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1744-06-15	1744	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	1744-07-15	1744	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3259	2020-08-15	2020	26.295	24.017	34.687	21.719	22.385	NaN	-32.168	10.782	18.640	NaN	17.309	17.877	21.636	NaN	36.614	29.313
3260	2020-09-15	2020	20.147	21.339	31.515	17.295	24.295	NaN	-29.611	13.377	18.195	NaN	21.780	13.577	20.292	NaN	34.803	29.152
3261	2020-10-15	2020	12.916	14.932	25.514	11.254	25.356	NaN	-26.018	16.502	11.396	NaN	23.555	7.342	14.475	NaN	29.182	28.697
3262	2020-11-15	2020	6.912	10.253	19.019	9.571	24.040	NaN	-17.661	19.850	4.856	NaN	27.774	3.151	7.961	NaN	25.027	24.323
3263	2020-12-15	2020	1.580	7.786	14.977	5.613	23.103	NaN	-12.548	21.215	-0.523	NaN	27.115	-0.296	2.714	NaN	19.435	19.443

3264 rows × 190 columns

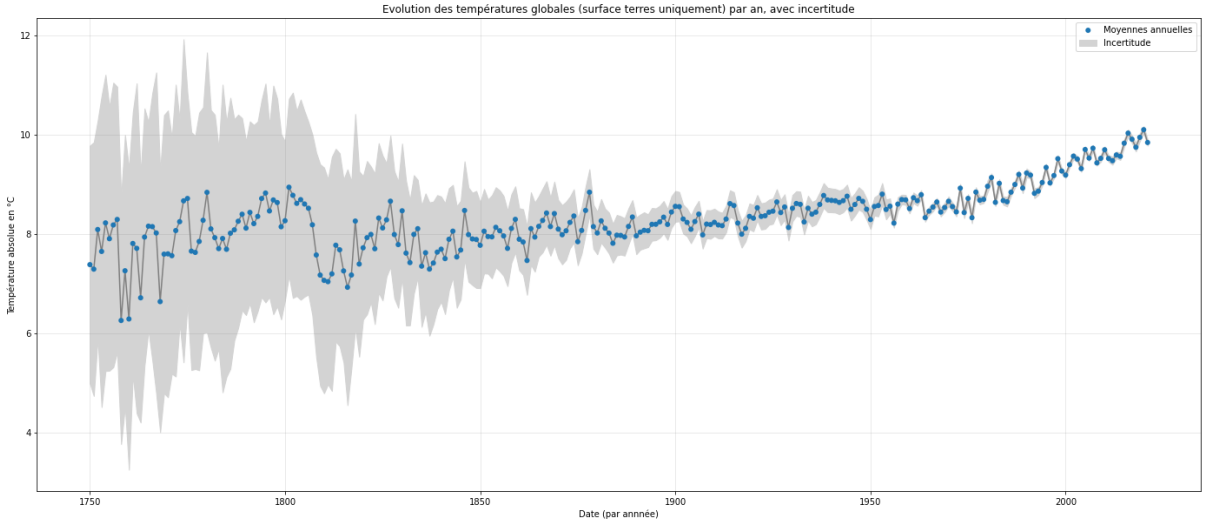
Analyse exploratoire

Données de températures

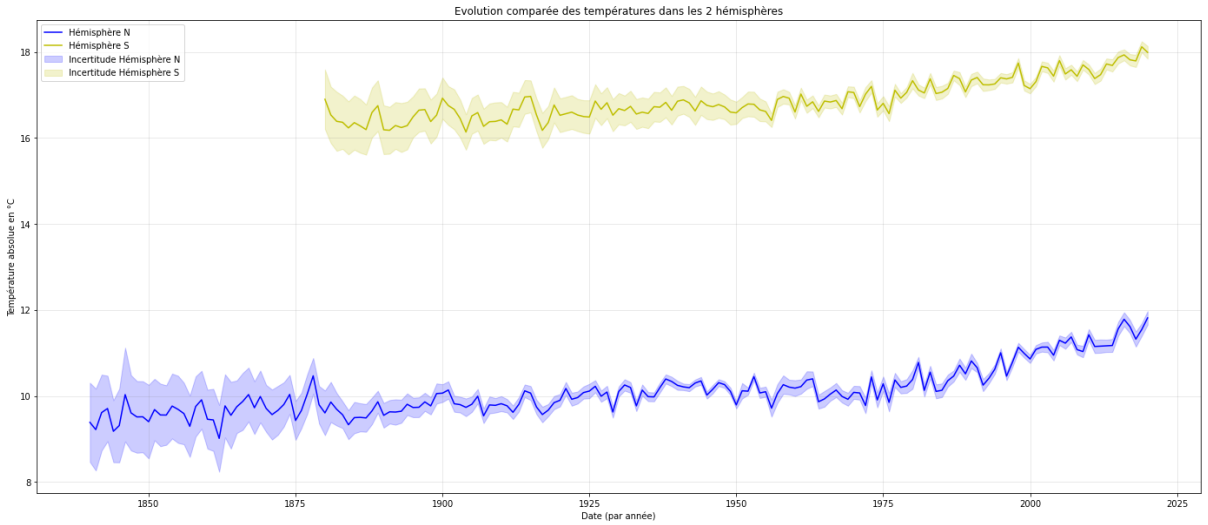
Ces données étant mensuelles, nous pouvons observer des variations saisonales en cycles annuels, matérialisées par les bandes de couleurs. la moyenne annuelle glissante, en noir dans le graphique ci-dessous :



Nous visualisons également les températures moyennes par année, en affichant la marge d'incertitude. Les écarts de température plus importants en début de dataset sont confirmés par **la forte marge d'incertitude**, en grande partie due à la rareté et au manque de précision des capteurs utilisés au 18ème siècle :



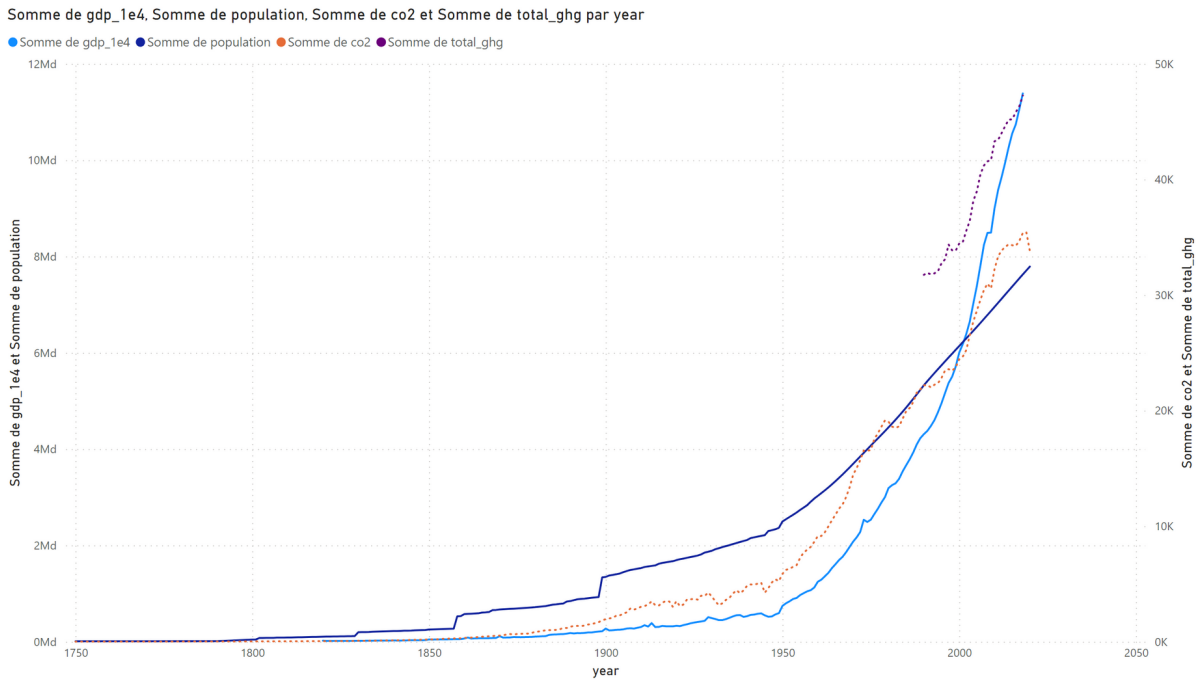
Le jeu de données obtenu en agrégeant les températures par hémisphère est aisément visualisable :



Données de dioxyde de carbone

Les **données par pays** fournissent une répartition intéressante des émissions en fonction des pays. La présence d'informations annexes telles que le produit intérieur brut et la population ouvrent également des perspectives de recherche intéressantes.

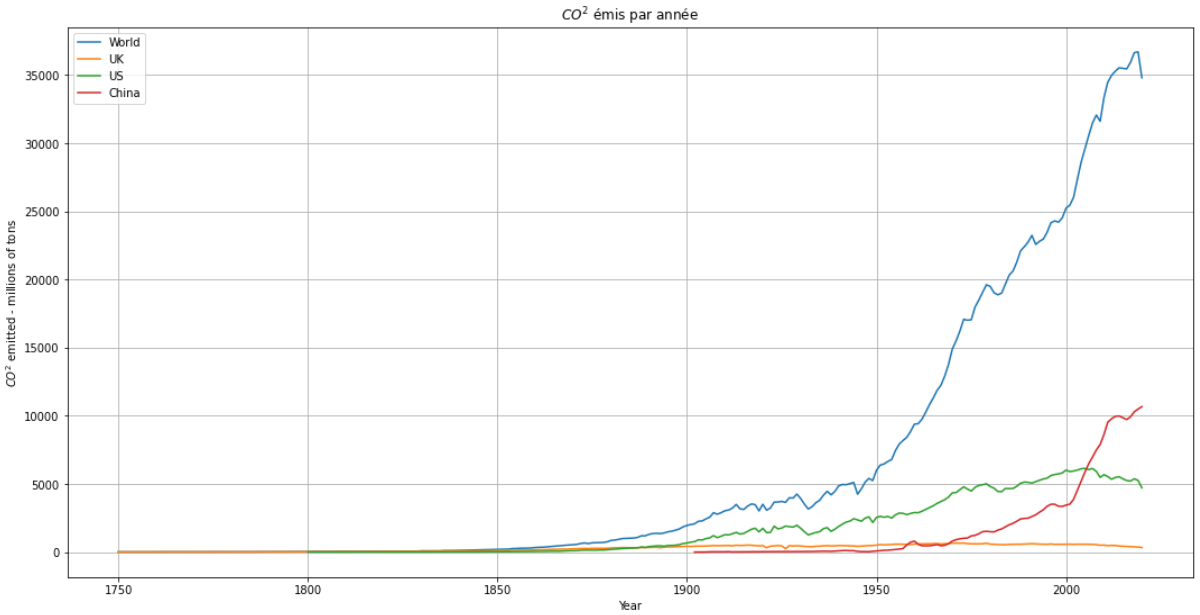
Réalisons un graphique rapide des mesures principales de ce jeu de données, avec une double échelle pour pouvoir comparer les évolutions respectives sur un même graphique. Les données annexes (population, produit intérieur brut) sont dessinées en trait continu et les émissions (CO_2, gaz à effet de serre) en trait pointillé.



Ces informations ne seront pas utilisées dans le cadre de notre étude du réchauffement climatique, mais pourraient utilement être exploitées pour de prochaines explorations plus approfondies. Nous notons enfin que la quantité totale de gaz à effet de serre (colonne total_ghb) ne commence que dans les années 1990 car les mesures dont elle est composée (méthane, ciment, gaz, transport maritime..) ne commencent qu'à cette date là.

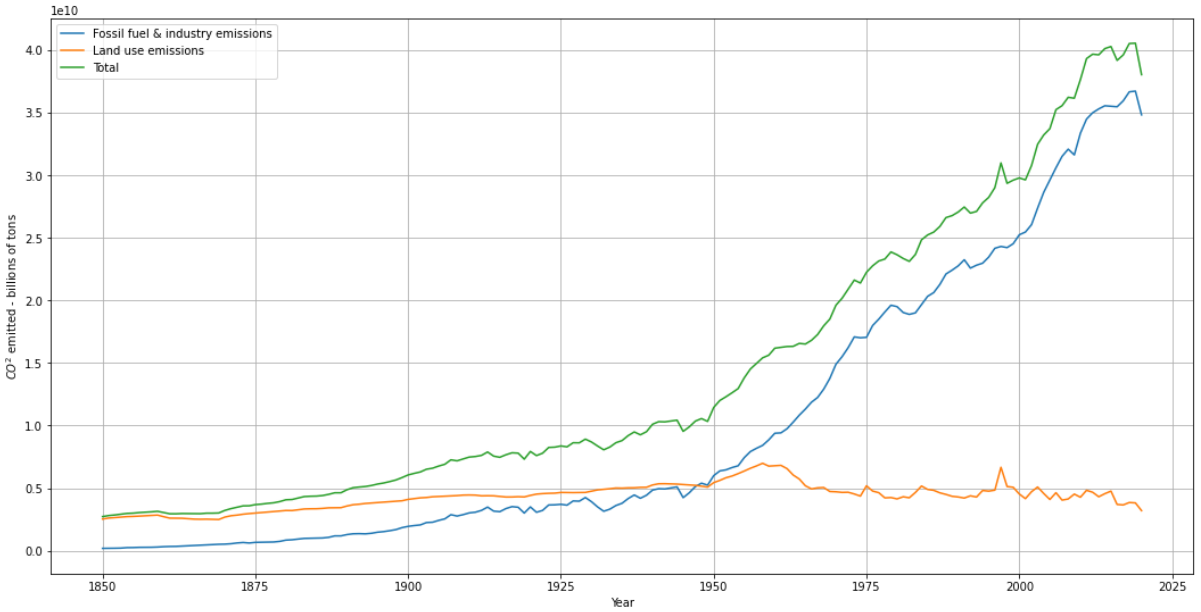
Afin de travailler sur une période de temps plus grande (à partir de 1750) nous utiliserons seulement les informations de CO_2 régionales (par pays et au niveau mondial).

Affichons un graphe rapide de la production de CO_2 pour quelques pays :



Le **jeu de données globales** est plus complet (i.e. ne contient aucune valeur manquante), et s'étend sur une durée plus courte – il commence près d'un siècle plus tard. On observe une différenciation marquée des deux mesures individuelles dès le début du jeu de données.

Les émissions industrielles commencent avec l'ère industrielle, aux alentours de 1850, et augmentent beaucoup sur les 70 dernières années. De manière comparative, les émissions liées à l'utilisation de la terre sont relativement stables. Nous obtenons l'évolution suivante pour ces deux mesures :



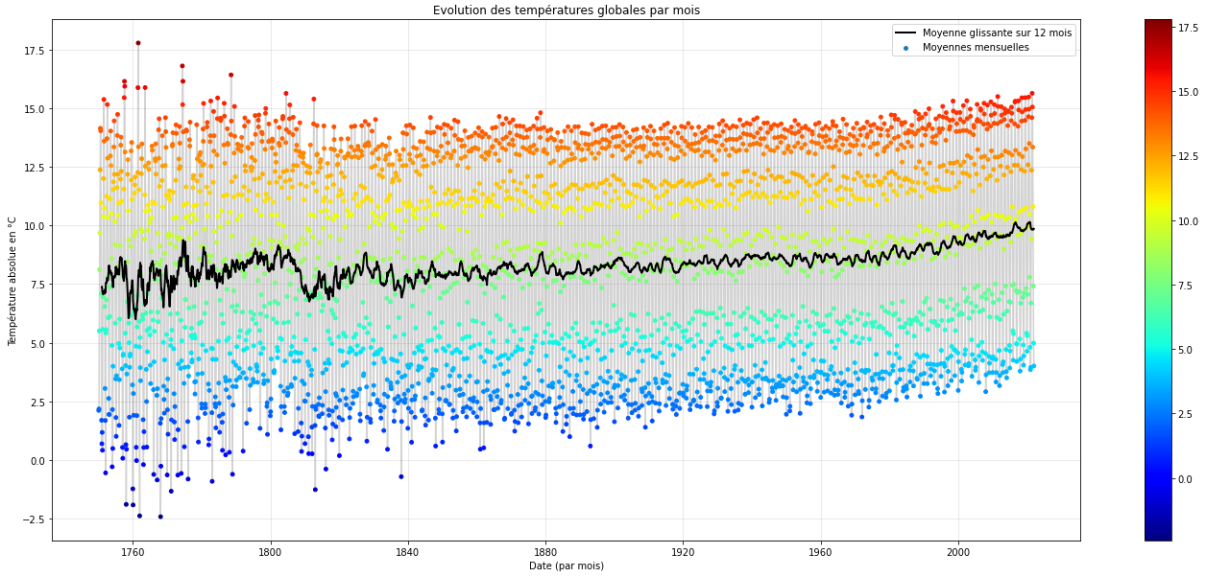
Analyse et Modélisations

Qu'est-ce que le réchauffement climatique?

Le réchauffement climatique est un phénomène de changement climatique caractérisé par une augmentation générale des températures moyennes à la surface de la Terre, qui modifie l'équilibre climatique et les écosystèmes.

Pouvons-nous confirmer le phénomène de changement climatique ?

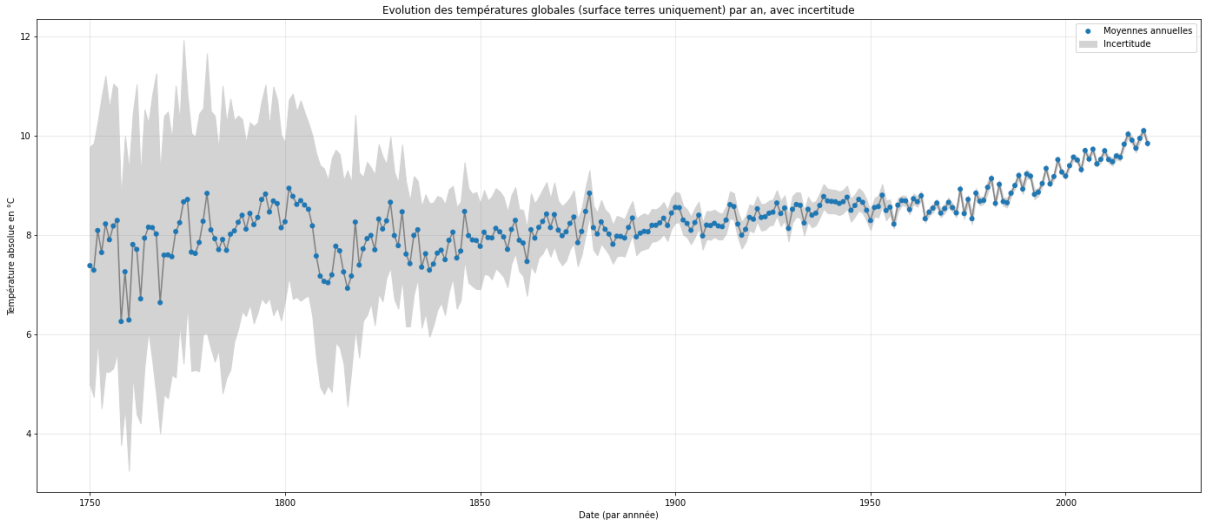
Commençons notre analyse en tentant de confirmer ce phénomène. A cet effet, à partir des données du data set: “global_land” nous présentons un graphique sur l'évolution des températures mensuelles globales, de 1750 à 2021. Egalement, nous allons présenter la moyenne annuelle glissante sur l'ensemble des données du même data set, rendant plus visible la tendance générale :



Nous pouvons observer des variations saisonnières correspondant à chaque année, matérialisées par les bandes de couleurs pour chaque variation annuelle de température. Également, une tendance croissante est visible en suivant les points d'une même couleur (i.e. les valeurs saisonnières augmentent globalement).

Approximativement, jusqu'à la moitié du graphique (année 1880) la dispersion des bandes de couleurs est beaucoup plus importante, qui s'accompagne d'une plus grande incertitude. Les données collectées à cette époque sont moins fiables, à cause des capteurs utilisés et

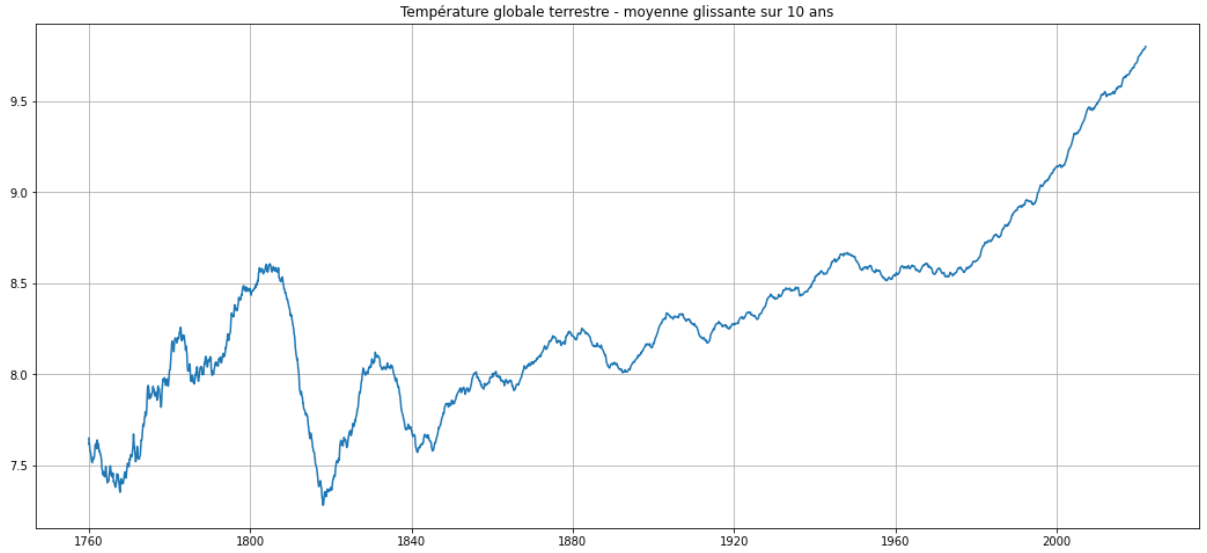
d'un nombre de stations météorologiques restreint. Cette incertitude est clairement représentée dans le graphique suivant, obtenu à partir du même data set :



La tendance est encore plus visible sur la moyenne annuelle glissante : si jusqu'en 1880 la moyenne oscillait autour de 8 °C, elle remonte sur les dernières décennies autour de 10 °C. Nous pouvons identifier une augmentation approximative de 2 °C en à peine plus d'un siècle.

Bien que cette valeur puisse paraître relativement petite (au vu, par exemple, des oscillations de la température entre jour et nuit, ou même été et hiver), les conséquences sur l'équilibre global sont énormes.

Afin de mieux se rendre compte de cette tendance, et pour lisser un peu les variations annuelles que l'on peut observer, nous calculons et présentons sur le graphique suivant une moyenne glissante sur 10 ans:



Les premières années sont assez chaotiques, cela étant directement lié à l'incertitude que nous avons déjà observée. À partir de ces observations, nous sélectionnons une période plus stable, par exemple à partir des années 1850, et calculons les différences avec les 10 dernières années :

- La différence de température moyenne sur 10 ans entre 1850 et 2022 est de 1.943°C.
- La différence de température moyenne sur 10 ans entre 1900 et 2022 est de 1.63°C.

Donc, **oui, le réchauffement climatique est bel et bien une réalité** !

Le réchauffement commence-t-il au même moment sur l'ensemble du globe ?

Après l'observation des graphiques précédents, il n'est pas possible d'établir précisément le début du réchauffement climatique. Il s'agit d'ailleurs d'un important sujet de désaccord entre experts, qui depuis des années ne parviennent pas à une réponse unique. Certaines recherches le corrèlent avec la révolution industrielle occidentale, telles que les travaux de Abram et al. ou ceux du groupe PAGES. D'autres études indiquent un début plus précoce.

Le réchauffement climatique est très graduel, et subit des variations cycliques qui rendent difficile une datation précise.

Néanmoins, en étudiant à nouveau le graphique, nous pouvons observer une tendance beaucoup plus explicite, forte et continue à partir des années 1975. De plus, dans les décennies précédant les années 1970, les températures moyennes mondiales semblent être assez stables, ce qui a suscité de vives controverses dans le domaine de la climatologie. En fait, il existe même quelques études publiées entre 1965 et 1979, qui prévoyaient une baisse des températures. En 1975, le New York Times titre aussi bien :

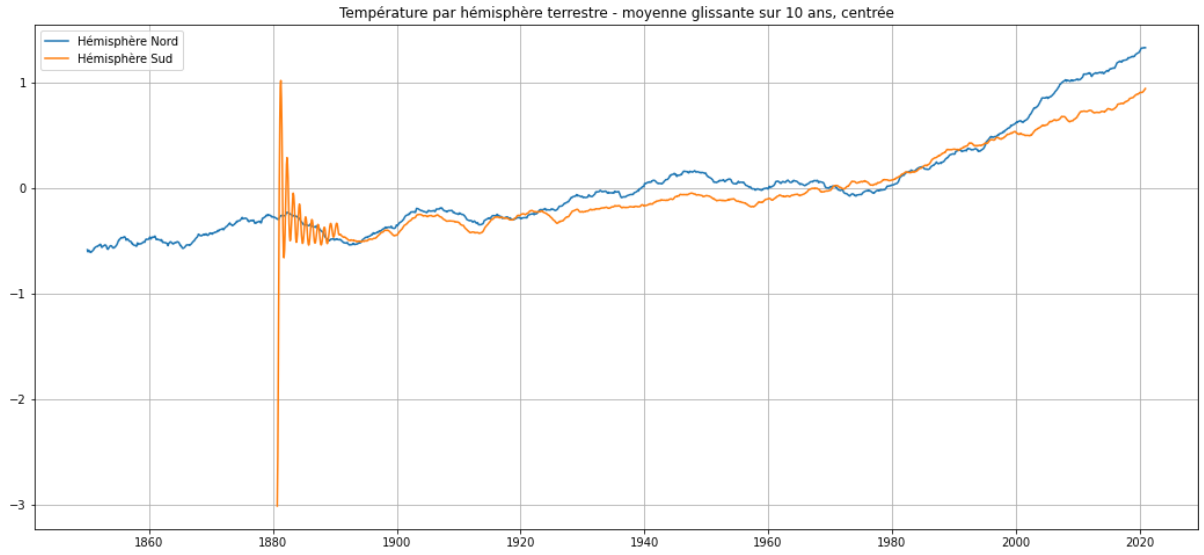
> « Des scientifiques s'interrogent sur les raisons du changement climatique : un refroidissement majeur pourrait être en vue »,

que :

> « Une tendance au réchauffement est observée : deux études contredisent l'idée d'une prochaine période froide ».

Dès la fin de cette décennie, le quotidien américain tranche et prend position dans un éditorial, le 12 juillet 1979, relayant sans équivoque l'inquiétude des chercheurs devant l'imminence du réchauffement climatique.

Pour déterminer si le réchauffement se produit au même moment sur l'ensemble du globe ou pas, nous allons analyser sur le graphique l'évolution des températures sur les deux hémisphères, obtenu à partir du data set “hems” :



De la même manière que pour les températures globales, dans ce graphique par hémisphère nous ne pouvons identifier avec précision un point de départ du réchauffement climatique. Les deux hémisphères montrent une tendance croissante, mais il est intéressant de remarquer que leur comportement est différent entre l'un et l'autre. La hausse de température dans l'hémisphère Sud est graduelle et constante, tandis que dans l'hémisphère Nord d'importantes variations apparaissent.

Sur cette dernière période (1970 - 2021), dans l'hémisphère Sud la température moyenne passe de 16,9 °C à 18 °C, soit une augmentation de 1,1 °C, tandis que dans l'hémisphère Nord la température passe de 10 °C à 11,8 °C, soit 1,8 °C d'augmentation sur la même période.

A cause de ces variations, il est difficile d'établir si le réchauffement a débuté plus tôt dans un hémisphère que dans l'autre, mais **nous pouvons observer globalement un réchauffement plus rapide de l'hémisphère Nord**.

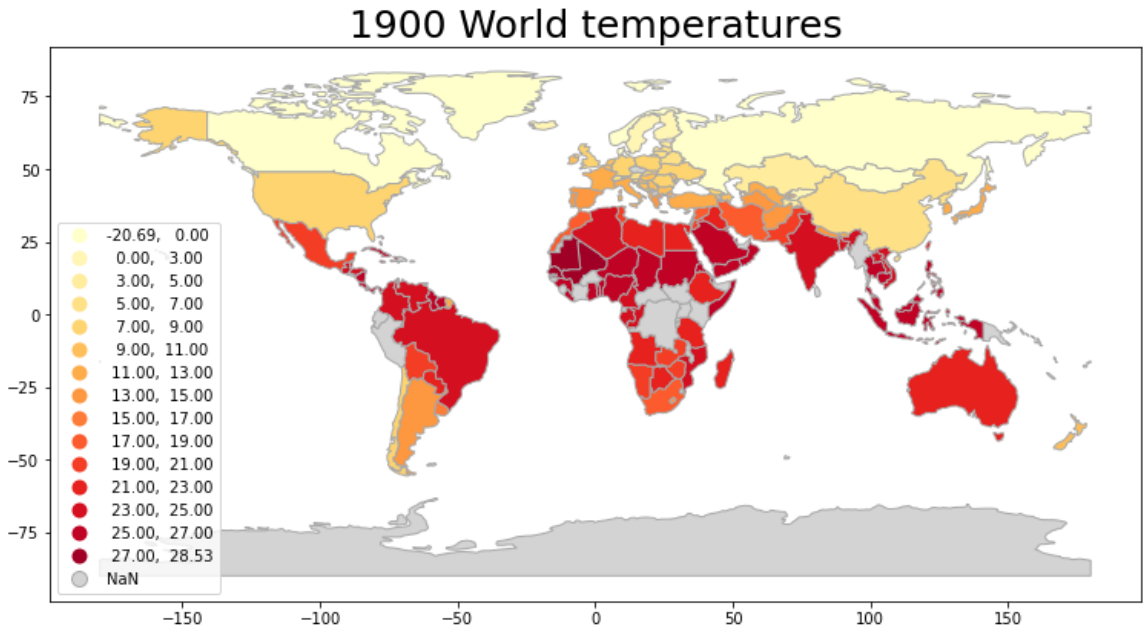
Afin de valider cette interprétation, et grâce aux données obtenues du data set “temp_countries” nous allons étudier plus en détail l'évolution des températures dans les différents pays et continents. Nous avons choisi un pays par continent, le plus représentatif à notre avis, en plus de Greenland, que nous avons présenté dans le graphique suivant :

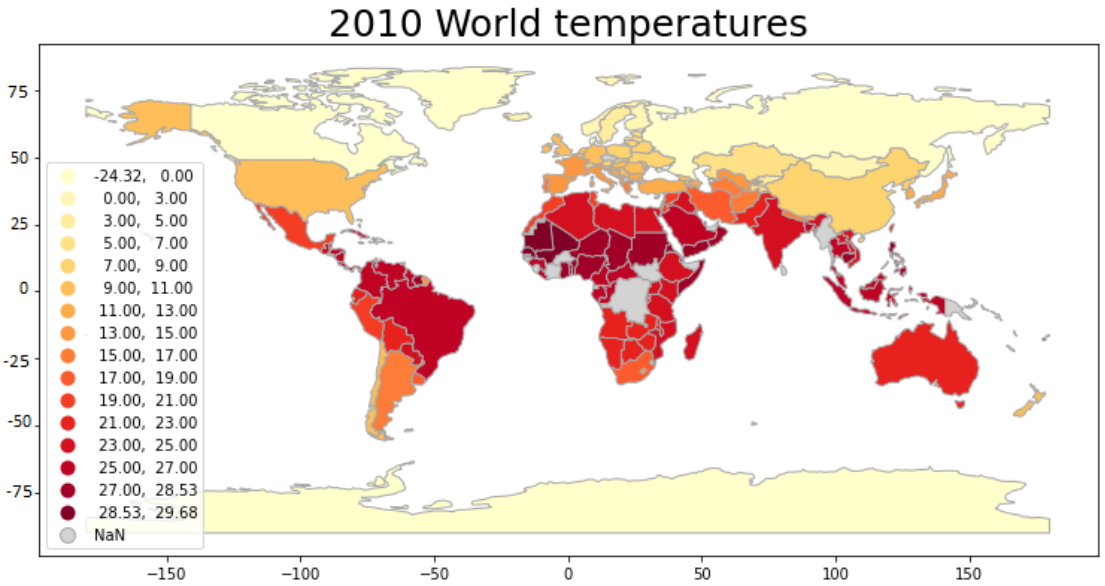
Dans ce graphique nous pouvons confirmer les éléments observés précédemment : le réchauffement est plus rapide dans l'hémisphère nord (courbes en trait plein) que dans l'hémisphère sud (courbes en trait pointillé). Sur l'ensemble des pays sélectionnés, on peut même observer que **plus les pays sont situés au Nord, plus la différence de température est visible**.

Le réchauffement est-il uniformément distribué sur le globe ?

Températures mondiales

Afin de mieux appréhender les différences de température au niveau mondial et à l'aide de l'outil geopandas (projet source intégré à la librairie pandas), nous visualisons les températures dans l'ensemble des pays du monde (pour lesquels nous avons des données dans le data set “temp_countries”) au début du siècle dernier, en 1900, et aujourd'hui (données accessibles jusqu'en 2020). Nous utilisons la même échelle de couleur pour permettre une bonne comparaison :

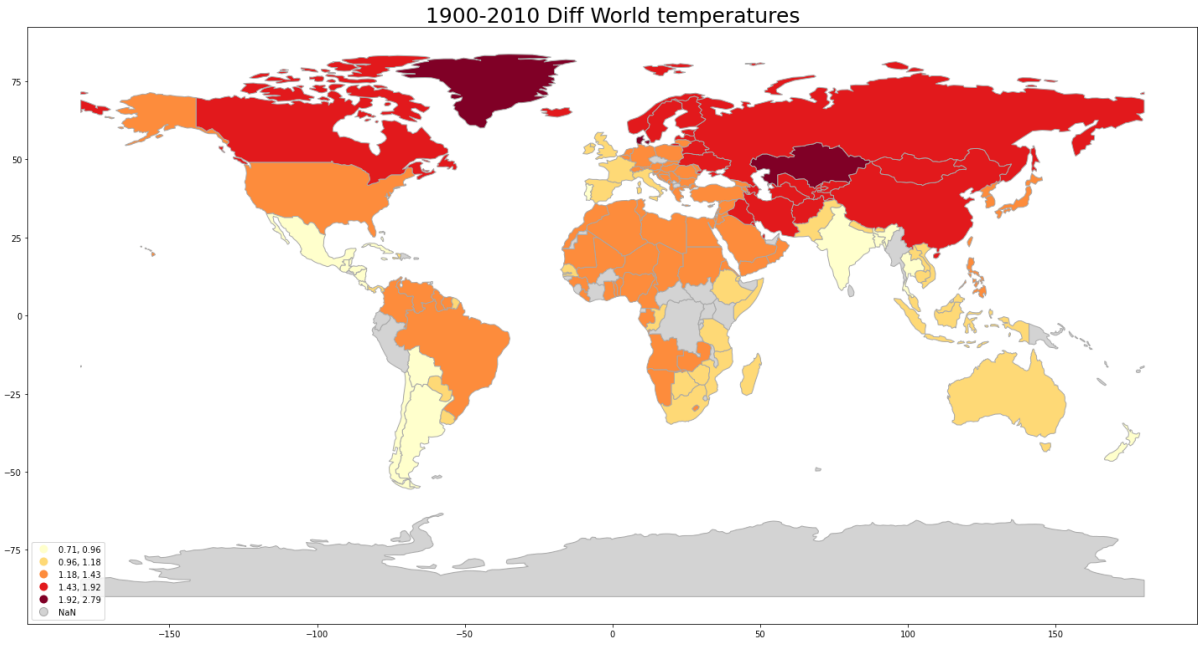




L'augmentation est visible ; beaucoup de pays ont une couleur plus prononcée en 2010 - la quasi-totalité des pays a pris au moins une teinte de couleur plus sombre, et l'algorithme a même dû ajouter une catégorie de température (> 28.58°C).

Différence sur un siècle

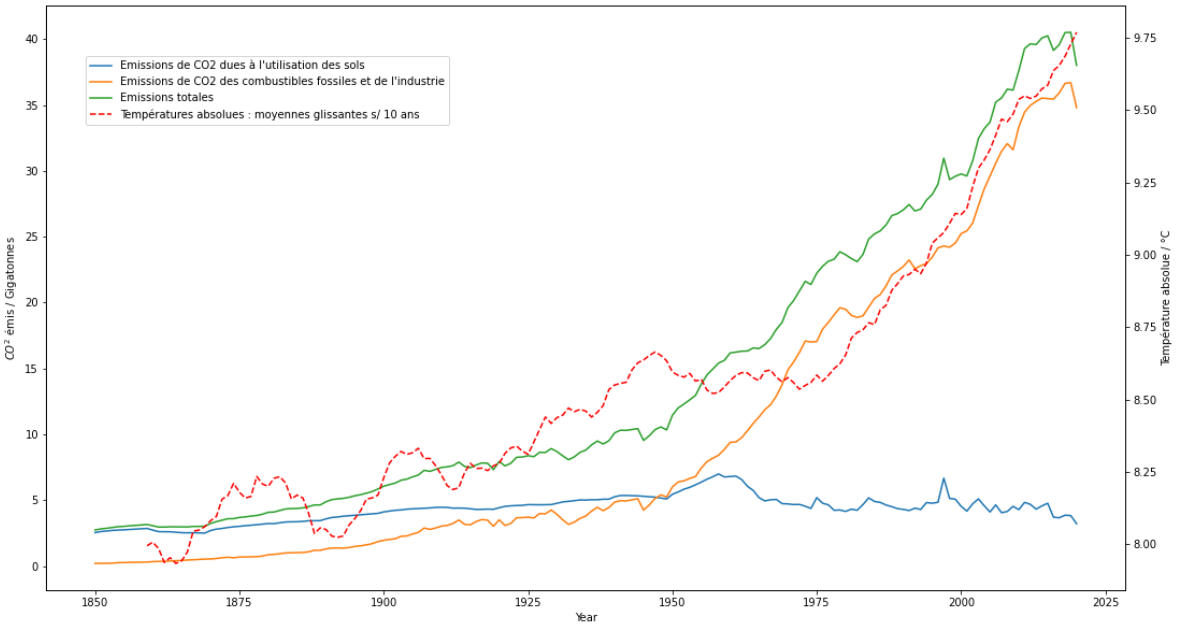
Nous voulons identifier l'augmentation de température sur l'ensemble des pays, en prenant comme référence les températures observées au début du siècle précédent (1900) et en les comparant aux températures observées ces dernières années (2010) :



Nous retrouvons les observations précédentes : **l'augmentation de température est en moyenne plus importante dans l'hémisphère nord, et croît globalement en remontant vers le Nord**. Cela n'est pas uniforme, cependant, d'autres paramètres doivent entrer en compte. Nous savons que la climatologie est une science complexe, et doit considérer des paramètres locaux (type environnement local, ou régulations de certains pays) autant que systémique par les effets globaux du climat, tels que les modifications des courants océaniques et atmosphériques.

L'évolution des températures est-elle corrélée aux émissions de CO_2 ?

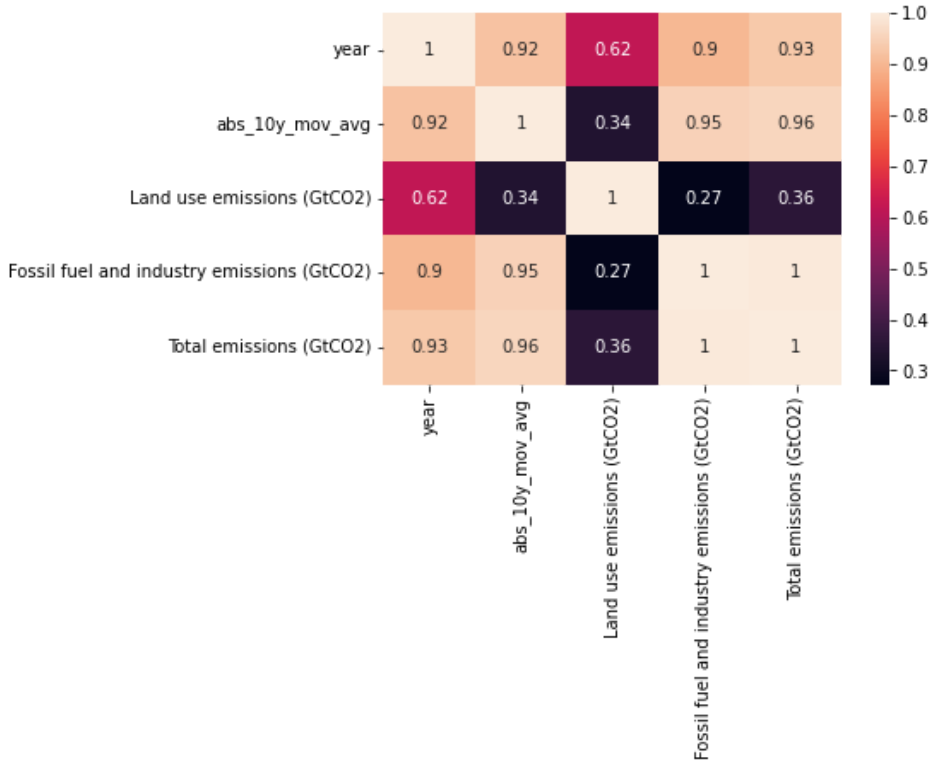
Afin d'analyser la relation entre ces deux variables, nous utilisons un DataFrame réduit (1850 - 2020), “co2_global”, dans lequel nous calculons la moyenne glissante sur 10 ans des températures globales, ainsi que la somme des émissions de CO_2 dues à l’utilisation des sols et de celles liés à la combustion d’énergie fossiles. Nous pouvons comparer l’évolution de ces 4 variables dans le graphique suivant :



La hausse des températures semble suivre celle des émissions totales de CO_2, bien que de manière moins linéaire. Nous constatons également une nette baisse des émissions de CO_2 en 2020, liée à la baisse globale d'activité pendant la crise Covid. Malheureusement, cette baisse ponctuelle n'aura eu aucun effet sur le climat, face à l'accumulation des rejets de CO_2 dans l'atmosphère pendant plusieurs décennies (article ici).

Corrélation

Nous appliquons le **test de Pearson** (qui permet de déterminer le degré de corrélation entre deux variables continues) à ce jeu de données, et affichons une **HeatMap** (fonction graphique de SeaBorn permettant d'afficher un tableau, colorisé en fonction des résultats de ce test pour chaque paire de variables d'un DataFrame) pour les repérer :



Le calcul de la p-value permet de confirmer les résultats des tests de Pearson. Nous vérifions celle-ci pour les coefficients de corrélation de chaque paire de variables qui nous intéresse :

	Coeff test Pearson	P-value
Températures / Années	0.920264	1.108349e-66
Total émissions CO2 / Années	0.933693	2.742281e-77
Températures / Total émissions CO2	0.958577	1.256212e-88
Emissions CO2 par utilisation sols / Années	0.618689	1.924302e-19

Elles sont pour chaque test très proches de 0. En conséquence, nous rejetons l'hypothèse nulle selon laquelle les variables de chaque paire sont indépendantes entre elles.

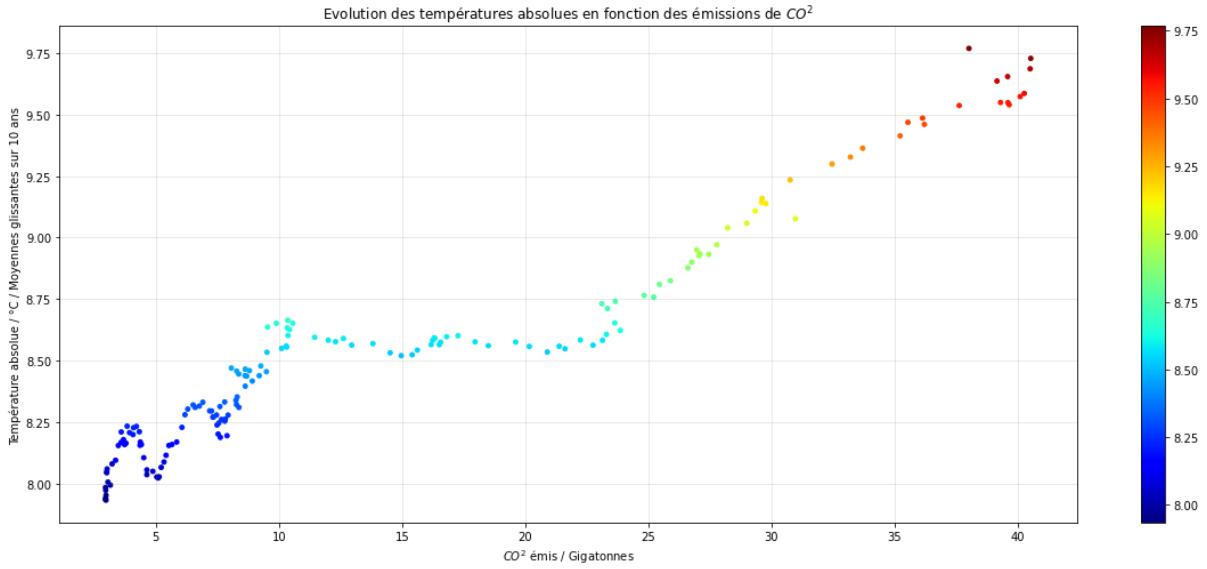
Ainsi, nous pouvons confirmer que :

- Les émissions totales de CO_2 et les températures moyennes sur 10 ans présentent chacune une forte corrélation aux années : coefs > 0.9. Les années passant, les émissions de CO_2 et les températures augmentent.
- Les températures moyennes sur 10 ans et les émissions totales le sont encore plus entre elles : coef > 0.95. Au-delà de la tendance générale à la hausse de ces 2 variables, cela confirme que globalement, les variations de l'une suit les variations de l'autre.

- Les émissions dues à l'utilisation des sols ne sont que moyennement corrélées aux années : coef = 0.62. Sur le graphique, nous observons en effet une baisse de celles-ci sur le dernier tiers de la période étudiée (1960 - 2020).

Régression

Intéressons nous de plus près à la relation qu'entretiennent émissions de CO_2 et températures. Nous pouvons la visualiser à l'aide d'un **scatterplot** (nuage de points) :



Nous avons vu grâce aux tests statistiques que ces 2 variables sont fortement corrélées, il n'est pas surprenant d'observer une relative linéarité dans cette représentation graphique.

Une régression linéaire simple a pour objectif d'expliquer une variable Y par le moyen d'une autre variable X.

Nous modélisons le lien entre nos deux variables avec la fonction **LinearRegression** de **SciKit-learn**, qui repose sur ce principe : $Y = \tilde{\beta}_0 + \tilde{\beta}_1 * X_1$, où X_1 est la variable explicative et Y la variable expliquée (variable cible). Nous obtenons les coefficients $\tilde{\beta}_0$ et $\tilde{\beta}_1$ égaux respectivement à 7.98864401361582 et 0.03814822, ce qui nous donne l'équation suivante:

Température (°C) = 7.989 + 0.0381 × émissions de CO_2 (Gt)

Evaluation du modèle

Notre modèle de régression linéaire obtient un score R^2 (coefficient de détermination indiquant la qualité d'une régression linéaire) proche de 0.92. **Il est donc performant** et confirme une linéarité d'environ 92% entre nos deux variables. Les émissions de CO_2 constituent donc bien un facteur majeur de la hausse des températures.

Forts des résultats des tests statistiques de Pearson, et du score obtenu par le modèle de régression linéaire, nous sommes à présent en mesure d'affirmer que **statistiquement, la hausse des températures est très fortement liée à celle des émissions de CO_2**.

Attention cependant : dans le cadre d'une étude statistique comme la nôtre, **corrélation ou linéarité ne signifient pas nécessairement causalité**.

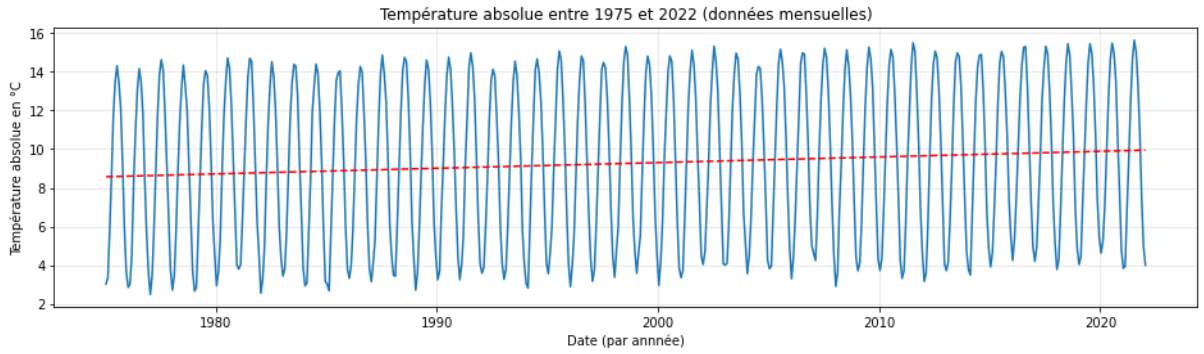
Pouvons-nous prédire les températures sur les prochaines décennies ?

Parmi différents modèles de prédiction, nous avons choisi **Facebook Prophet**.

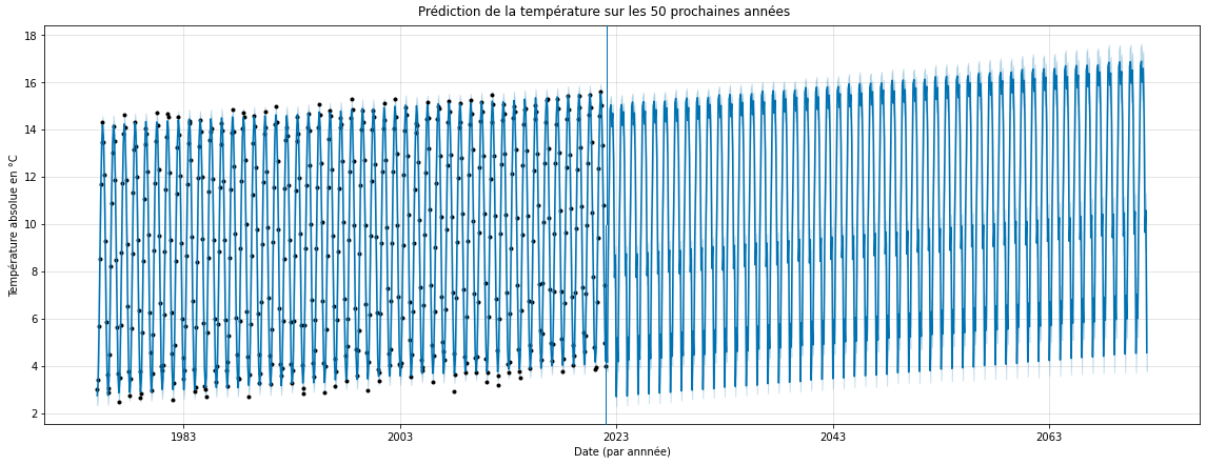
Prophet permet la prédiction de données de séries temporelles. Nous avons testé le modèle multiplicatif et le modèle additif de Facebook Prophet et nous avons vu que le deuxième correspond mieux à nos données: aux tendances non-linéaires s'ajoute une composante liée à la saisonnalité. Cette librairie est donc particulièrement adaptée à notre étude.

Nous allons nous concentrer sur la période où la tendance croissante de la température est constante et forte, c'est-à-dire à partir de 1975.

Dans le graphique suivant nous visualisons bien que pendant la période comprise entre 1975 et 2022 la tendance de la température absolue est explicite, croissante, constante et forte (augmentation de la tendance de 8,7 °C en 1975 jusqu'à 10 °C en 2022) :



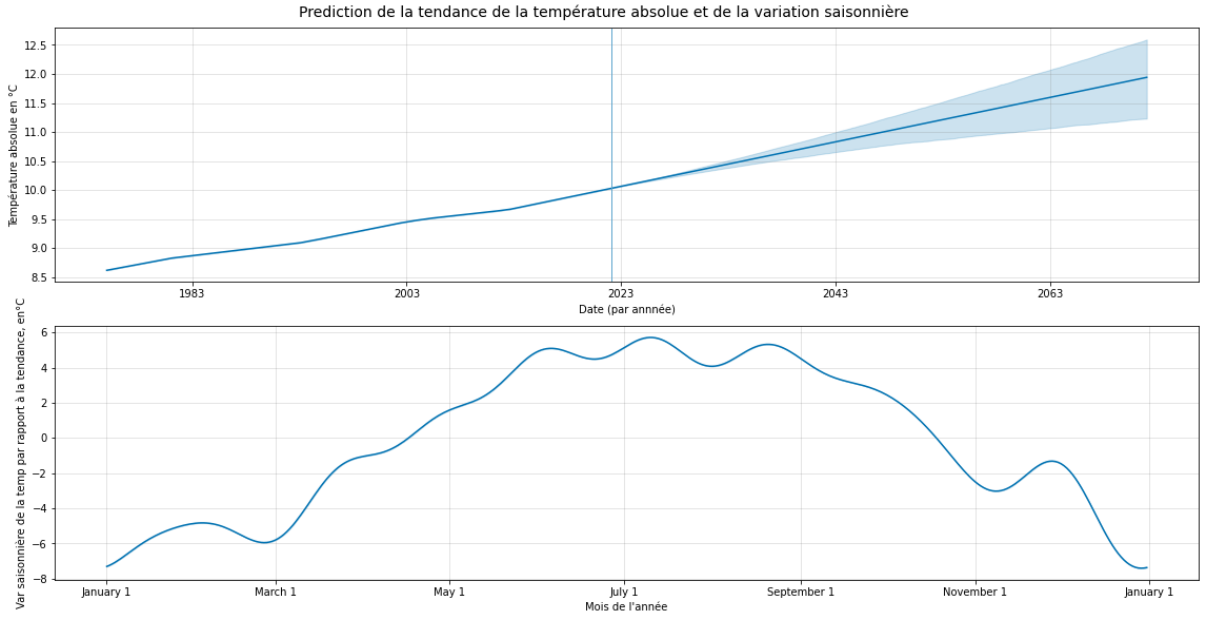
A partir de nos données, Facebook Prophet calcule la prédiction suivante pour les températures sur les 50 prochaines années :



La prédiction qui va à partir de la barre verticale de l'année 2021 du graphique nous montre toujours une évolution de la température à la hausse.

Pour plus de détail, visualisons 2 composantes de la prédiction :

- La tendance ;
- Les variations saisonnières de cette tendance :



Notre prévision repose sur le principe que la température globale absolue va suivre la même tendance que ces 46 dernières années, sans action climatique majeure. Dans ce cas, notre modèle prédit une croissance approximative de 1.9 °C dans 50 ans (10 °C en 2021, jusqu'à 11,9 °C en 2072) . En comparaison avec la multitude d'études prévisionnelles qui existent actuellement, notre résultat semble juste et raisonnable, mais bien peu optimiste.

Bilan & Suite du projet

Ce sujet est déjà connu, largement étudié et débattu par des experts hautement qualifiés. Nous souhaitions principalement savoir si la Data Analyse nous permettrait d'atteindre des résultats concordants.

Grâce à l'analyse statistique de données fiables et concrètes, nous avons pu répondre aux problématiques posées en début de projet

- Nous avons pu confirmer et quantifier le réchauffement climatique : c'est une réalité indiscutable.
- Le phénomène est graduel depuis la révolution industrielle de 1880, et s'accélère à partir de 1975, de manière plus forte encore dans l'hémisphère nord.
- Les émissions de CO_2 ont une forte influence (corrélation / linéarité) sur la hausse des températures.
- Nos prédictions ne sont pas optimistes. D'après notre modèle Prophet, la température moyenne globale augmentera de près de 2°C dans les 50 prochaines années. Ces résultats rejoignent ceux de la majorité des études disponibles sur le changement climatique :
    - La hausse de la température globale s'est encore accentuée, selon le dernier rapport du GIEC
    - Réchauffement climatique : voici l'ampleur des hausses que vous connaîtrez
    - Réchauffement climatique : les prévisions alarmantes de Météo France

Afin de compléter notre étude, de nombreux autres facteurs impactants pourraient être analysés, notamment :

- Émissions de Gaz à Effet de Serre autre que le CO_2 (méthane, etc.).
- Évolution de la répartition entre sols naturels et sols exploités par et pour l'activité humaine.

Malheureusement, nos conclusions, et a fortiori celles des experts en climatologie, ne sont positives ni pour notre planète Terre (UnhapPy Earth), ni pour l'humanité.

Bibliographie

Données

- Data Overview - Berkeley Earth
- GISS Surface Temperature Analysis (GISTEMP v4)
- Our World in Data

Documentation

- 6e rapport du GIEC : quelles solutions face au changement climatique ? - Réseau Action Climat
- Le véritable coût du changement climatique | National Geographic
- The Elusive Absolute Surface Air Temperature (SAT)
- Global Carbon Project
- Changement climatique et effet de serre | Insee
- Early onset of industrial-era warming across the oceans and continents
- Le réchauffement climatique anthropique aurait débuté au tout début de la révolution industrielle / Actualités scientifiques / Actualités / Accueil - IPSL
- Covid-19 et baisse des émissions de CO2 | Cairn.info
- Hoax climatique #3 : quand les scientifiques prévoyaient un refroidissement
- GeoPandas
- Prophet | Forecasting at scale.

Travail d'équipe

Nous avons fortement collaboré durant l'exécution de ce travail :

- Lors de chaque phase, après analyse des besoins courants, nous avons décidé de **tâches qui pouvaient être réalisées indépendamment** (recherche de sources, analyse et nettoyage de jeux de données, production de visuels..) et nous nous les sommes affectées, puis **nous avons fait un suivi** afin de nous entraider et tous apprendre des expériences de chacun.
- Des **réunions hebdomadaires** nous ont permis de décider régulièrement des tâches à exécuter, suivre leur progression et nous adapter aux aléas : difficultés, compétences spécifiques, indisponibilité passagère.
- Tous nos rendez-vous ont donné lieu à des **comptes-rendus de réunion**, avec l'état des lieux et les actions affectées à chacun : https://pad.castalia.camp/mypads/?/mypads/group/earth-temp-w4u0n6s/view
- L'utilisation **d'outils collaboratifs** (Google Colab, GitHub, Etherpad) nous a permis de travailler indépendamment, et de fusionner notre travail régulièrement afin d'assurer la cohérence du résultat. Nous avons procédé à des **revues de code** régulières pour assimiler le travail de chacun et tous les documents ont été **collaborativement revus et validés**.

Répartition de l'effort

- Olga :
    - Apport des idées pour définir les questions et la direction de notre recherche.
    - Préparation d'un cheminement du projet.
    - Étude, compréhension et sélection des données à importer.
    - Nettoyage de données et analyse des données manquantes.
    - Constitution de la moyenne glissante sur 12 mois.
    - Travail d'analyse et interprétations des résultats et graphiques obtenus.
    - Travail rédactionnel.
    - Traitement préliminaire des questions de recherche 6 :
        - Pouvons-nous prédire les températures sur les prochaines décennies?

- Boris :
    - Définition des questions de recherche et de la stratégie.
    - Identification et pré-traitement des sources de données de températures.
    - Nettoyage de données et analyse des données manquantes.
    - Identification et pré-traitement des sources de données de CO_2.
    - Traitement préliminaire des questions de recherche 2 et 3 :
        - Le réchauffement commence-t-il au même moment sur l'ensemble du globe ?
        - Le réchauffement est-il uniformément distribué sur le globe ?
    - Travail d'analyse et interprétations des résultats et graphiques obtenus.
    - Maintenance des outils collaboratifs (GitHub, Etherpad).

    - Travail rédactionnel.

- Nicolas :
    - Définition du contexte du projet et des questions de recherches.
    - Repérage, acquisition et familiarisation avec les jeux de données.
    - Définitions de fonctions de visualisation adaptées à nos jeux de données
    - Paramétrage de ces fonctions et productions graphiques.
    - Interprétations et analyse des observations graphiques.
    - Traitement de la question de recherche 4 :
        - L'évolution des températures est-elle corrélée aux émissions de CO_2 ?
    - Travail rédactionnel.

Difficultés rencontrées

Bien qu'aujourd'hui il existe énormément d'études sur le sujet du réchauffement climatique, de nombreuses controverses persistent entre climatologues professionnels. Il n'y a pas de réponses uniques aux problématiques relatives à ce phénomène.

Au début du projet, nous n'étions pas sûrs que notre niveau d'expérience sur le sujet nous permettrait d'obtenir des résultats intéressants, logiques et réalistes. Pourtant, grâce aux compétences acquises au cours de la formation et à un travail de groupe fructueux, nous considérons que nous y sommes parvenus.

Difficultés prévisionnelles :

De nombreuses sources de données sont disponibles sur internet. Par conséquent, nous avons passé plus de temps que prévu à étudier et nous familiariser avec ces différentes sources et data sets, pour pouvoir choisir ceux qui s'adapteraient le mieux à l'analyse que nous envisagions.

Difficultés concernant l'acquisition des jeux de données :

Les jeux de données initiaux utilisent des formats variés et peu évidents au premier abord (espaces optionnels en début de ligne, séparation des champs par espaces, mesures de référence insérées en en-tête, etc.), et nous avons dû utiliser des expressions régulières complexes pour l'extraction de toutes les données. Par ailleurs, de nombreuses tables ont dû être concaténées ou fusionnées après nettoyage, au moyen de pivots et non sans mal (structure et nettoyage des données).

Difficultés liées à nos compétences techniques ou théoriques :

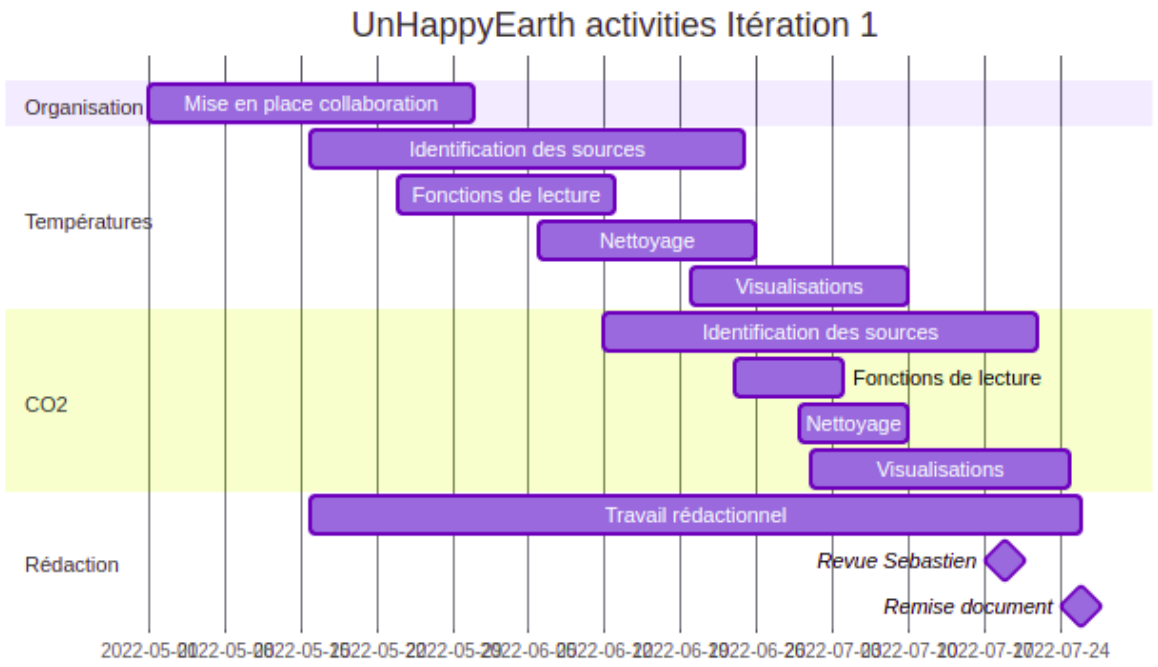
Plusieurs notions utilisées durant le projet n'ont pas été proposées ou pas suffisamment abordées en formation, comme les librairies GeoPandas et Facebook Prophet, ou pour la

construction de certains graphiques. Néanmoins, grâce aux bases acquises en formation et à la documentation disponible sur internet, nous avons su les utiliser et les appliquer.

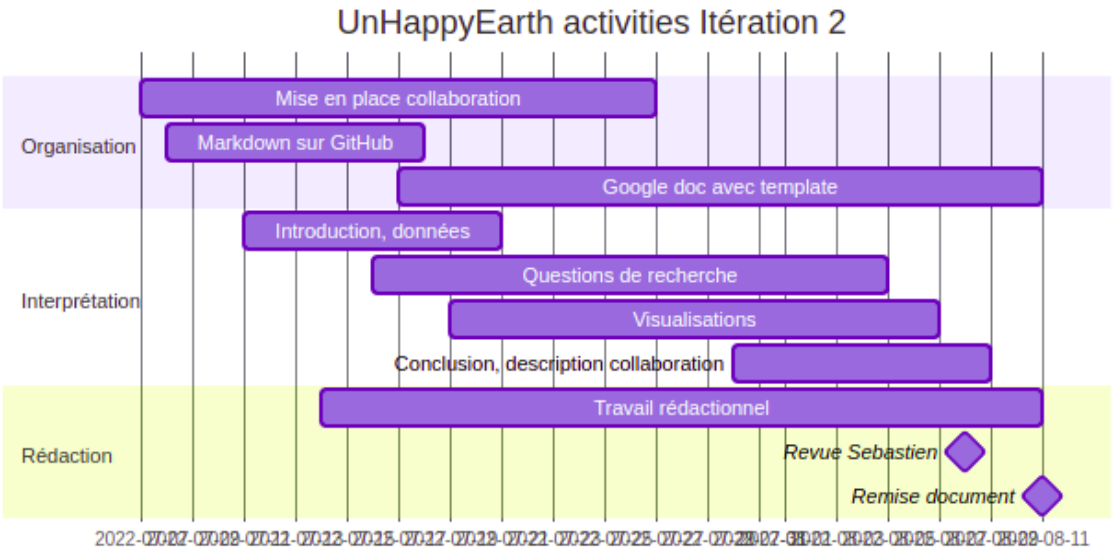
Annexes

Diagramme de Gantt

Itération 1 : Analyse exploratoire



Itération 2 : Modélisation



Description des fichiers de code

- **Itération_1_-_Analyse_Exploratoire.ipynb** :

Acquisition, nettoyage et analyse exploratoire des données.
Lien sur le référentiel GitHub : Itération 1 - Analyse Exploratoire.ipynb

- **Itération_2_-_Modélisation.ipynb** :
  Data Analyse, modélisations et conclusions.
  Lien sur le référentiel GitHub : Itération 2 - Modélisation.ipynb