Fundamentals of Climate Data Science

Authors: Juliane Manitz and Barbara Milewski
Github code repository: https://github.com/jmanitz/climate_data_book

---

To do: Properly cite/reference sources, such as for images and resources

---

Introduction [Barbara]

Climate is essentially weather patterns over a long timescale, such as decades or centuries compared to weather events over several hours or days. Therefore, statistics and data science play an important role in climate data analysis, such as in the likelihood that weather patterns will be in a certain way after some amount of time or the probability that Earth's climate conditions were within a certain range of values a specific number of centuries ago. In the past few decades, climate science has begun focusing on researching anthropogenic climate change, i.e. climate change caused by human activity, as that is one of the top pressing contemporary issues. Thus, it is important to understand the drivers of climate change and how those are studied and modeled by climate scientists.

The interdisciplinary nature of climate data science means it is important to have the domain knowledge needed to understand the data being worked with and what to look for when conducting analyses. Asking the right questions is vital for uncovering essential insights and having background knowledge of the field will help explain what could be the causes and effects of those trends. There are many components to climate data science for the purposes of understanding the fundamentals of climate science and modern climate change.

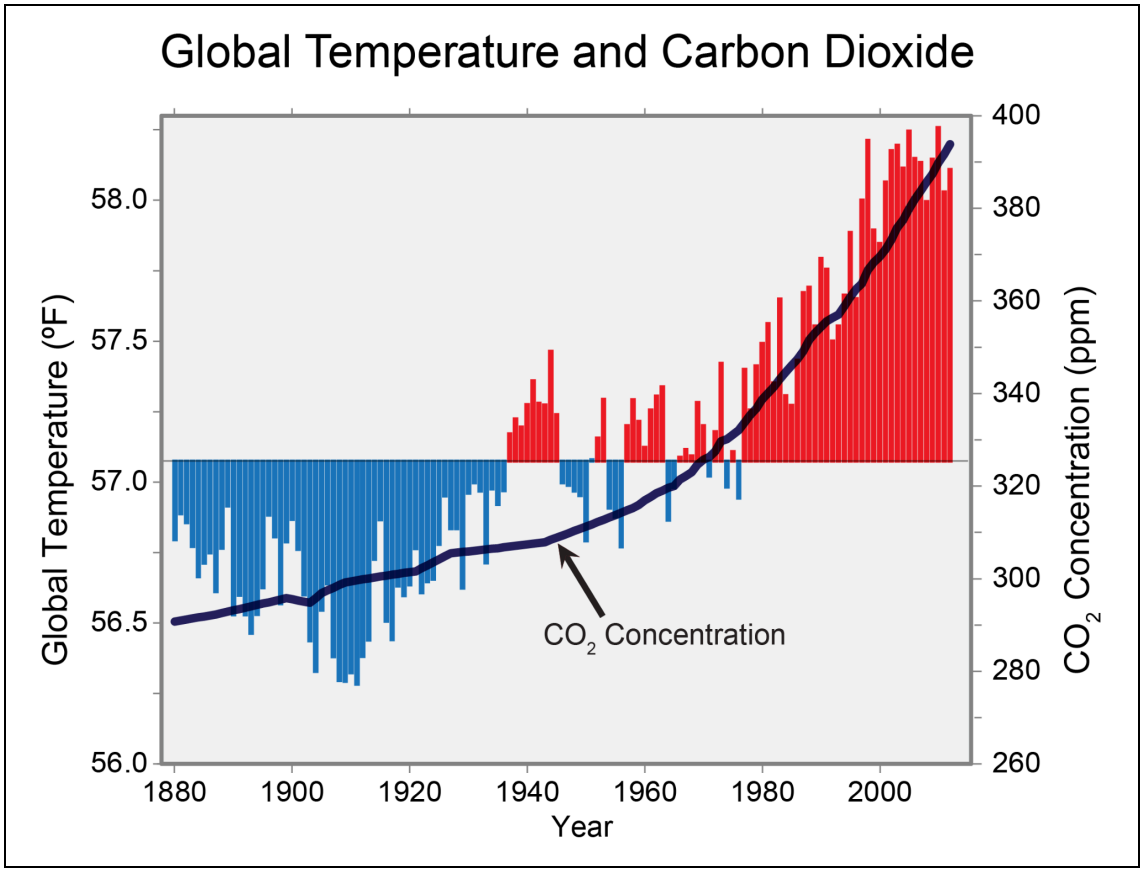


Figure: Global Temperature and Carbon Dioxide. Source: https://nca2014.globalchange.gov/report/our-changing-climate/observed-change

Scientists know that climate change is occurring by observing long-term trends in physical variables such as atmospheric concentrations of greenhouse gasses, changes in precipitation patterns, and calculating the atmospheric radiative balance. By analyzing these observed trends, the drivers behind climate change as well as its impacts can be understood. The image above shows the correlation between increasing atmospheric concentrations of carbon dioxide in parts per million and the resulting increase in average global temperature in degrees Fahrenheit. The visualization easily conveys the relationship between the two variables and it is clear that more carbon dioxide in the atmosphere is leading to global warming. This correlation is explained by the physical properties of the radiative balance of the Earth which is explained in further detail in the “Climate Models” section of this chapter. While this is a simple example, it is a good introduction to the kind of analyses needed to study and understand the fundamentals of climate science and climate change.

This chapter will aim to cover the fundamentals of climate science, relevant data sources and the uses of the data, typical key variables analyzed, common statistical methods, and the role of climate models, such as informing the Intergovernmental Panel on Climate Change (IPCC) or NASA reports on climate change.

The Physics of Climate Change

The greenhouse gas effect leading to global warming is explained by the concept of radiative forcing. Radiative forcing is important because the imbalance between the inflow of energy to Earth and the outflow of energy from Earth due to greenhouse gasses absorbing some of that outflow back into the Earth system results in net heating. The diagram below illustrates Earth's energy balance and the imbalance that occurs due to higher greenhouse gas concentrations in the atmosphere. This radiative forcing also provides the basis for the most fundamental one-dimensional climate models called energy balance models. For those interested, a more in-detail explanation of the intricacies of Earth's radiative balance is provided in Chapter 7 of the IPCC's Sixth Assessment Report.

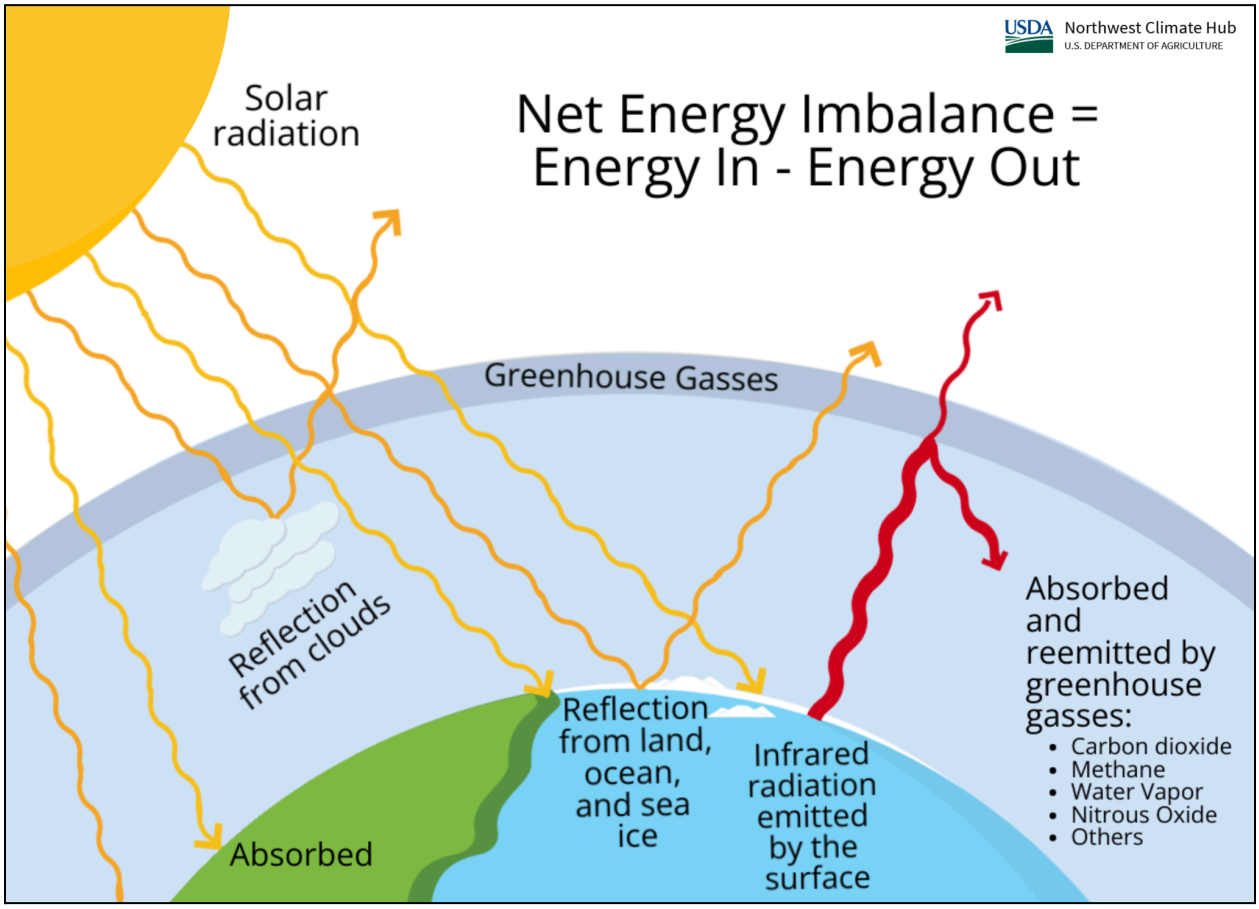


Figure: Radiative forcing. Source: https://www.climatehubs.usda.gov/hubs/northwest/topic/what-are-climate-model-phases-and-scenarios

There are numerous gasses in the atmosphere that can contribute to the net energy imbalance, but the focus is on greenhouse gasses that remain in the atmosphere for a long time and therefore are difficult to remove, as well as leading to long-lasting and compounding effects over time. Carbon dioxide is the most well-known due to its long atmospheric lifetime on a timescale of centuries and that high concentrations are emitted through human activity. While trees are often discussed as a way to sequester carbon dioxide, that is not a permanent solution due to how carbon cycles between the biosphere (e.g., trees), lithosphere (e.g., fossil fuel deposits),

and the atmosphere, referred to as ‘reservoirs’ of carbon. The carbon emitted by humans comes from the lithosphere through oil, gas, and other fossil fuel deposits that are extracted. Carbon in the lithosphere comes from organic matter that is very slowly converted to fossil fuels that then remain underground for thousands of years, making the lithosphere a long-term reservoir of carbon. Humans extract these fossil fuel deposits and burn the fossil fuels, releasing this carbon into the atmosphere. Carbon does not easily return to the lithosphere from the atmosphere and instead tends to cycle quickly between the biosphere and atmosphere through plants breathing in carbon dioxide and releasing it back after death. Furthermore, the release of carbon into the atmosphere by plants is rising due to increasingly common and severe wildfires. This means that the carbon humans are releasing from the lithosphere in just these last few decades will not return to the lithosphere naturally for thousands of years. This is why even if all anthropogenic carbon emissions were stopped today, global warming would continue due to the cumulative effect of all the carbon that has been released by humans until today. Besides the biosphere and lithosphere, there is also the uptake of carbon dioxide by the oceans that deserves discussion here.

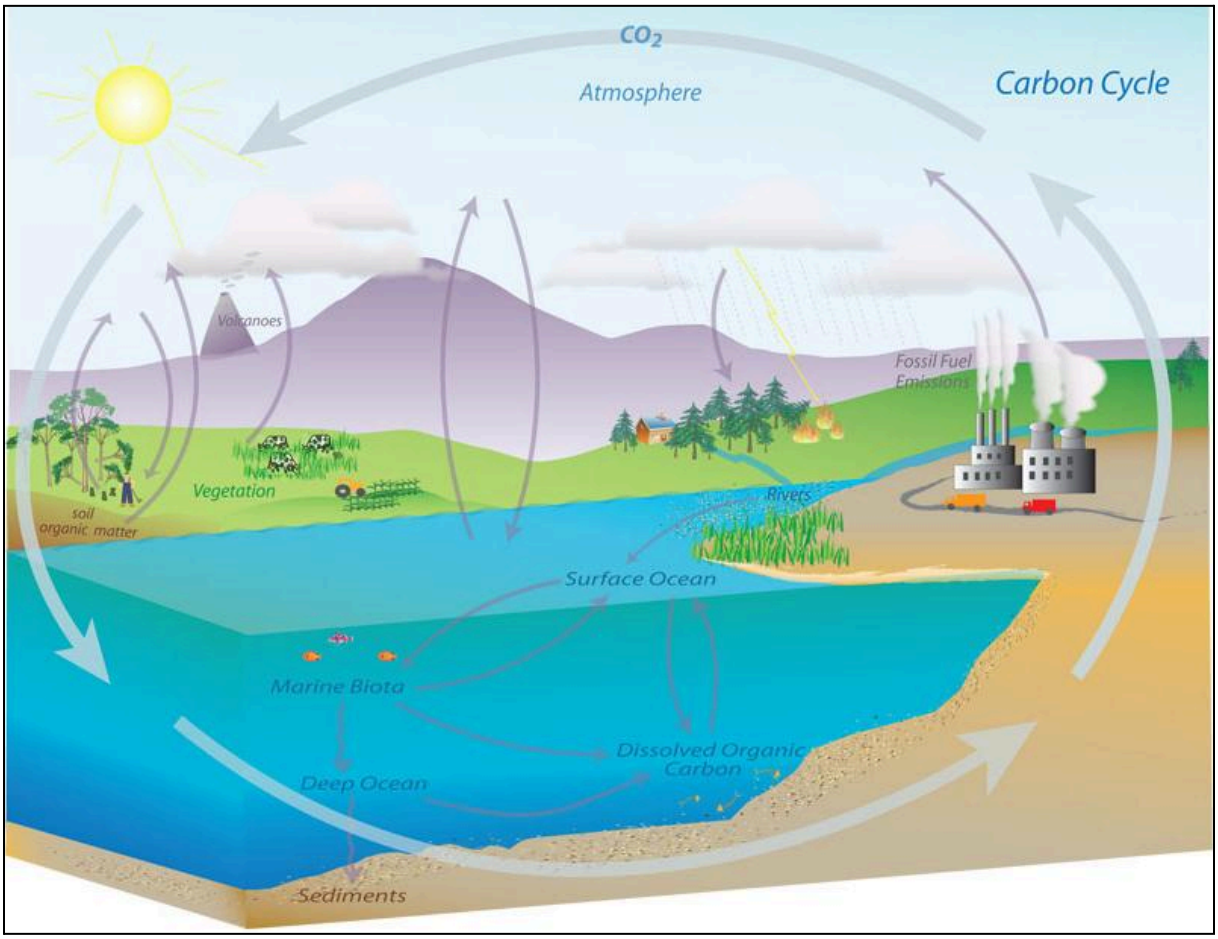


Figure: The carbon cycle showing the biosphere, atmosphere, lithosphere, and ocean reservoirs of carbon. Source: https://www.noaa.gov/education/resource-collections/climate/carbon-cycle.

Current estimates put how much of anthropogenic carbon dioxide emissions the oceans absorb at about 25% to 30%. This emphasizes the oceans as a significant contributor to mitigating the

impacts of these emissions and slowing down global climate change, but this also results in the gradual acidification of the oceans, leading to hostile environmental conditions for marine life. The chemistry of how carbon dioxide is absorbed by the oceans is as follows:

$$CO_2 \rightleftharpoons CO_2 * H_2O$$

$$CO_2 * H_2O \rightleftharpoons H^+ + HCO_3^-$$

$$HCO_3^- \rightleftharpoons H^+ + CO_3^{2-}$$

Most of the dissolved CO_2 is in the form of HCO_3^-, meaning the amount of atmospheric carbon dioxide the ocean can continue to take up is determined by the overall equilibrium:

$$CO_2 + CO_3^{2-} \rightleftharpoons 2HCO_3^-$$

Essentially, as the oceans take up more carbon dioxide, they become more acidic, acting as a feedback loop leading to a reduced capacity to continue absorbing carbon dioxide. This means that while they are a significant buffer against the accumulation of carbon dioxide in the atmosphere today, even that cannot be relied on as a long-term solution.

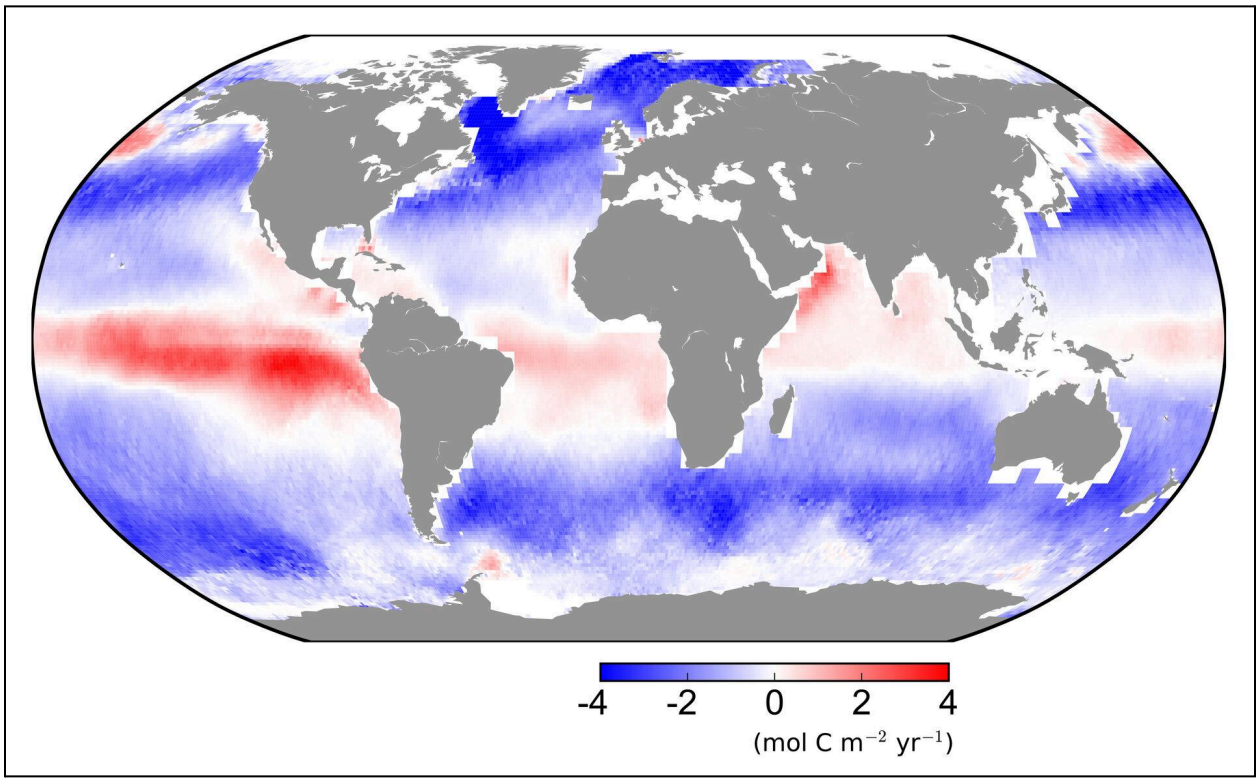


Figure: The flow of carbon dioxide between the ocean and atmosphere per year. Blue areas indicate where more carbon dioxide enters the ocean than leaves it, and make up most of the area depicted. Source: https://www.esa.int/Applications/Observing_the_Earth/Can_oceans_turn_the_tide_on_the_climate_crisis.

Another important feedback loop in climate change is the melting of ice, snow, and permafrost. Ice and snow are important for increasing the albedo, or reflectivity, of the Earth's surface, counteracting the greenhouse gas effect by reflecting more received energy back into space. As ice and snow melt, the albedo of the surface decreases and more energy is absorbed by Earth. This is particularly insidious when looking at how glaciers melt. First, a warmer spot starts to melt, resulting in water pooling. This water is darker than the ice around it, leading to even more energy being absorbed and even more heating taking place, accelerating the melting. This

process continues until the glacier breaks apart, meaning that even a small amount of melting at the beginning can result in significant losses of ice and reflectivity later on. The melting of permafrost has another detrimental impact on climate change: its loss releases greenhouse gasses such as methane that were stored within the permafrost. Therefore, permafrost is not only important for biodiversity but also for acting as a reservoir of greenhouse gasses that would otherwise contribute to global warming.

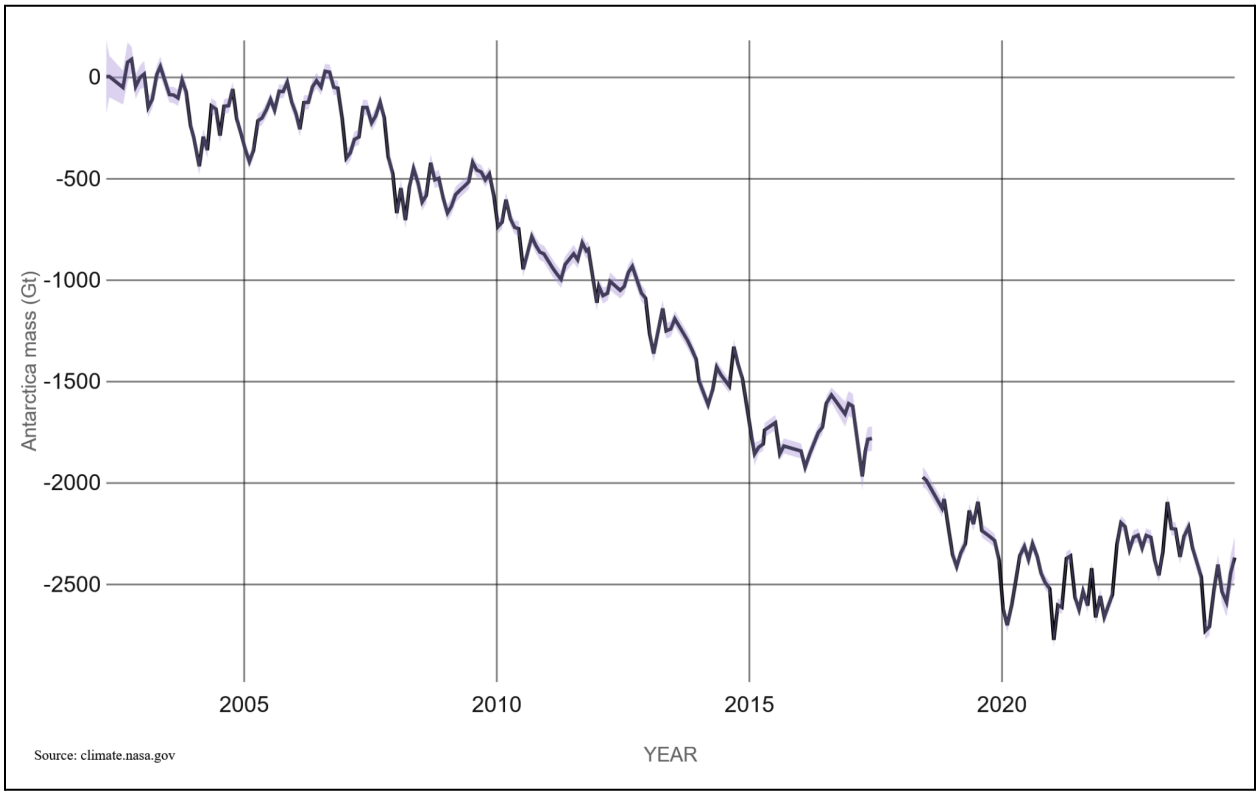
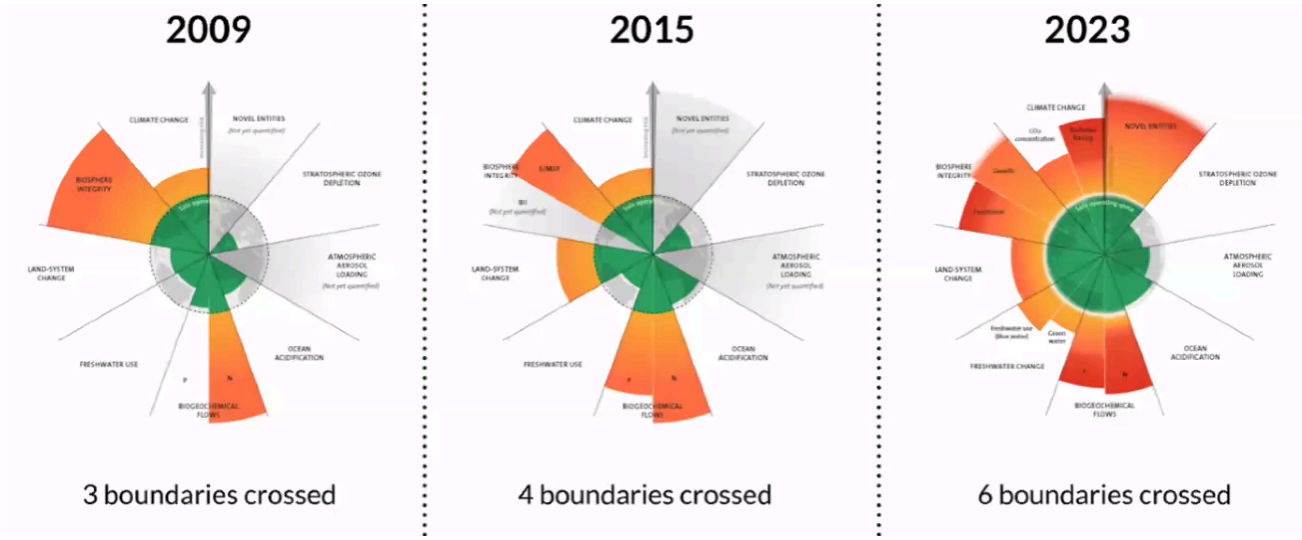


Figure: The variation of the mass of ice in Antarctica since 2002, measured by NASA's GRACE satellites. Source: https://climate.nasa.gov/vital-signs/ice-sheets/.

Climate change is just one of nine planetary boundaries, which define nine critical thresholds within which humanity can safely operate to maintain a stable Earth system. Alarmingly, six of these nine boundaries have already been crossed (see Figure xz), signaling that human activities are destabilizing key environmental processes beyond the carbon cycle. Other planetary boundaries include biosphere integrity (biodiversity loss), land-system change (deforestation), biogeochemical flows (nitrogen and phosphorus cycles), novel entities (chemical pollution), and freshwater use. Each boundary breach represents a systemic crisis, and they are deeply interlinked. For instance, deforestation not only contributes to biodiversity loss but also amplifies climate change by reducing carbon storage. Similarly, chemical pollution affects freshwater ecosystems, exacerbating biodiversity decline. The interconnectedness of these boundaries means that exceeding one often pushes others toward their limits, amplifying global risks. Urgent, systemic action is required to address these intertwined crises and restore planetary balance, as exceeding multiple boundaries increases the likelihood of triggering irreversible changes in the Earth system.



Further Reading

- How much carbon dioxide the oceans take up: https://news-oceanacidification-icc.org/2024/11/11/how-we-discovered-that-the-oceans-surface-absorbs-much-more-carbon-dioxide-than-previously-thought/
- Further explanation of the chemistry of the oceans' uptake of carbon dioxide: https://www.pmel.noaa.gov/co2/files/dickson_thecarbondioxidesysteminseawater_equilibriumchemistryandmeasurementspp17-40.pdf
- More climate feedback loops: https://earthhow.com/climate-feedback-loops/
- Watch how glaciers melt: https://www.nasa.gov/science-research/earth-science/the-anatomy-of-glacial-ice-loss/

Key Variables Analyzed [Juliane]

In climate data science, key variables analyzed are crucial for understanding and predicting climate patterns, assessing impacts, and developing mitigation strategies. These variables span multiple domains such as atmospheric, oceanic, terrestrial, and socioeconomic factors. In this section, we summarize some of the most important categories and specific variables analyzed in climate data science.

Atmospheric Variables

Temperature is one of the most critical indicators of climate change. Climate scientists track global, regional, and local temperature trends over time, focusing on surface temperatures, and atmospheric temperatures at various altitudes.

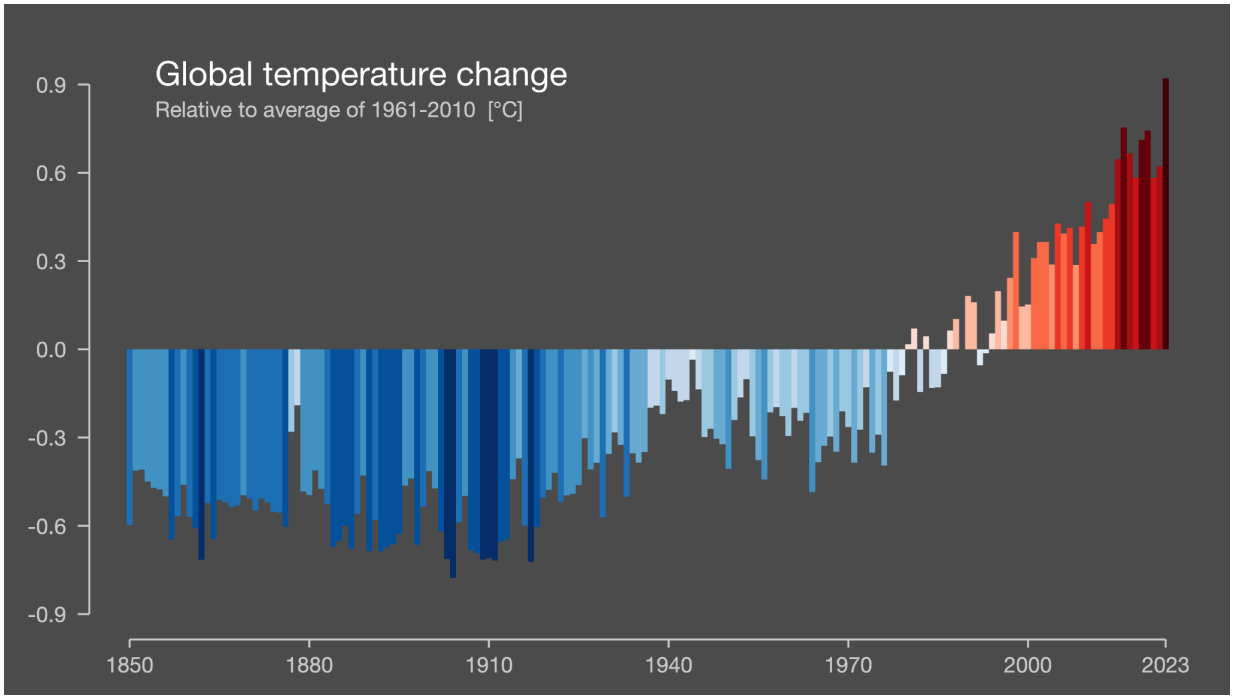


Figure: Global Temperature Change over Time Relative to Average of 1961-2010. Source: https://showyourstripes.info/c

Climate Change is caused by the increase of greenhouse gas (GHG) concentrations. Thus, the levels of gasses like carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) are central to climate models, as these gasses trap heat and contribute to global warming. GHG concentrations are typically measured in parts per million (ppm) using spectroscopy, which identifies each type of gas molecule absorbing a unique set of light wavelengths.

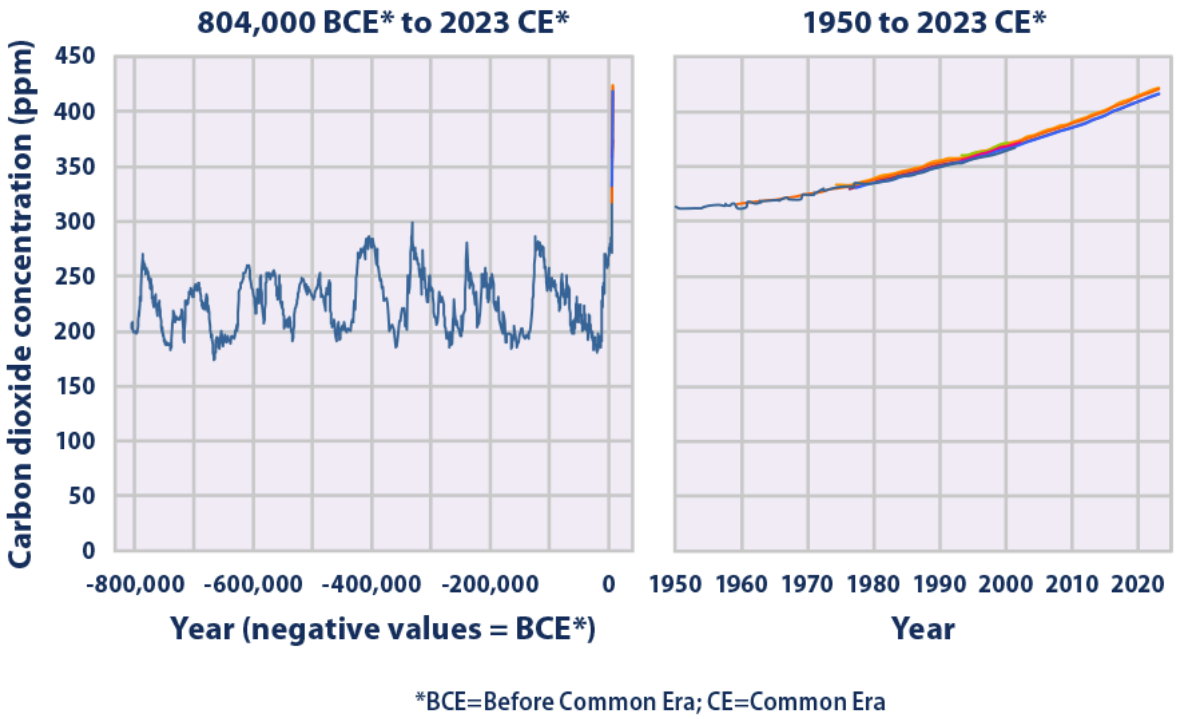


Figure: Global Atmospheric Concentrations of Carbon Dioxide Over Time. Source: https://www.epa.gov/climate-indicators/climate-change-indicators-atmospheric-concentrations-greenhouse-gases

Other atmospheric variables include precipitation and wind patterns, as well as humidity. Rainfall and snow patterns are studied to understand shifts in hydrological cycles, including changes in droughts and flood occurrences. Analyzing wind speeds and directions can provide insights into changes in storm activity, jet streams, and circulation patterns. Finally humidity is important, because as temperature rises, the air's capacity to hold moisture increases, influencing humidity levels. This can affect weather patterns, cloud formation, and precipitation.

The climate crisis is altering weather patterns and causes more frequent and extreme weather events, such as heatwaves, storms, hurricanes, floods and droughts. Soil moisture data is vital for understanding drought conditions, agricultural productivity, and water resource management. We are also interested in wildfire frequency and intensity, because as temperatures rise, wildfires are becoming more frequent and severe, affecting ecosystems and releasing stored carbon into the atmosphere. By analyzing trends in these events, scientists can assess future risks and inform strategies for mitigation and adaptation. Additionally, the links between changing precipitation patterns, rising sea levels, and temperature fluctuations make this data essential for evaluating the impacts on food security, infrastructure, and human health.

Oceanic Variables

Sea Level Rise is a central variable in climate data science. Monitoring sea level is critical for assessing the impacts of melting glaciers and polar ice caps, and thermal expansion of water due to warming. This is key for predicting coastal flooding and erosion risks.

Projections under different scenarios based on the assessment presented in the IPCC Sixth Assessment Report are available in the NASA sea level projection tool. Assuming intermediate GHG emissions (SSP2-4.5, CO_2 emissions around current levels until 2050, then falling but not reaching net zero by 2100), we can expect a sea-level rise of between 43 and 76 cm (17 - 30in), although the exact amount depends on several factors, such as ice sheet dynamics and regional variations.

Example: **Variation in Sea-Level Rise**

In this example, we explore the geographical variation in sea-level rise at different locations world-wide. We use data from the NASA sea level projection tool.

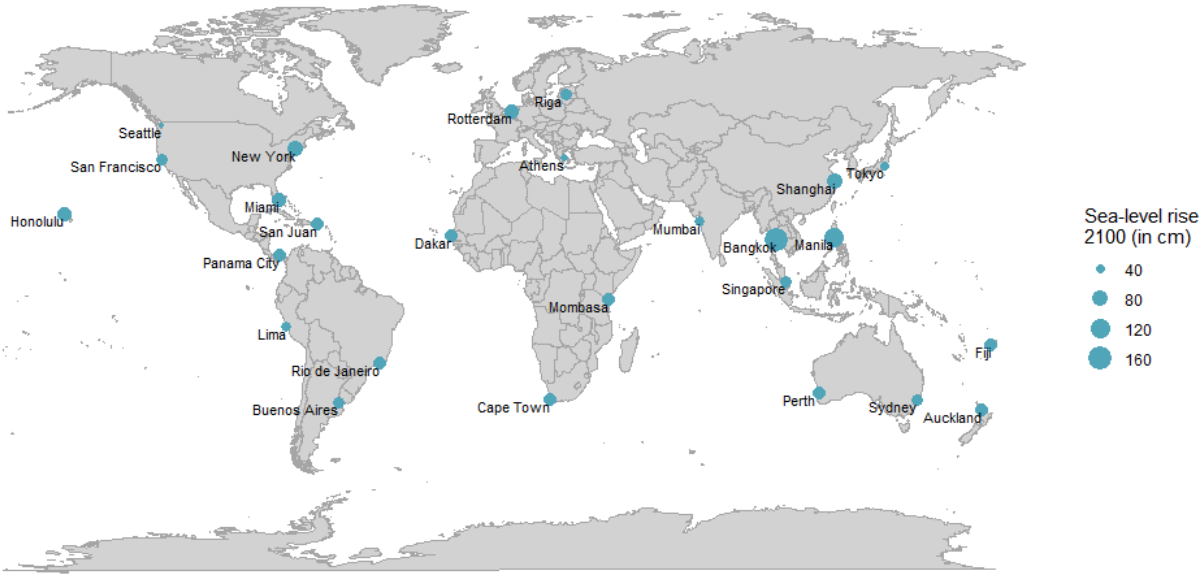
```
# Load sea-level data
dt <- read.csv("sea-level-projection-data.csv", skip=3)
# Add geo location to the dataset
dt[c("lon","lat")] <- ggmap::geocode(dt$location)

## i <https://maps.googleapis.com/maps/api/geocode/json?address=New+York&key=xxx>
## i <https://maps.googleapis.com/maps/api/geocode/json?address=Miami&key=xxx>
## . . .
```

Next we plot the information on the world map with point size proportional to local sea-level rise projection

```
# World map data
world <- map_data("world")

ggplot() +
  geom_polygon(data = world,  aes(long, lat, group = group),
               color = "darkgrey", fill = "lightgrey", linewidth=0.1) +
  geom_point(data=dt, aes(lon, lat, size=sea_level_proj), color="#50A7BA") +
  geom_text(data=dt, aes(lon, lat, label=location), hjust=1, vjust=1, size=3) +
  theme_void() + labs(size = "Sea-level rise\n2100 (in cm)")
```



Other relevant oceanic variables include:

- Sea Surface Temperatures (SST): Rising SSTs are closely linked to stronger storms and hurricanes, coral bleaching, and changing marine ecosystems. SSTs also play a role in ocean circulation patterns.
- Ocean Currents: Ocean circulation, including thermohaline circulation, influences global climate systems. Changes in currents like the Gulf Stream can impact weather patterns globally.
- Ocean Acidification: Increasing CO_2 levels lead to more acidic oceans, which harms marine life, particularly organisms that rely on calcium carbonate for their shells or skeletons.

Cryospheric Variables

The cryosphere is an umbrella term for those portions of Earth's surface where water is in solid form. This includes sea ice, ice on lakes or rivers, snow, glaciers, ice caps, ice sheets, and frozen ground. Measuring their magnitude is crucial to inform important feedback loops. For example, the size and mass of glaciers and polar ice sheets (in Greenland and Antarctica) are crucial for understanding long-term sea level rise and albedo effects, where melting ice reduces the Earth’s reflectivity and accelerates warming. Tracking seasonal snow cover helps assess changes in freshwater availability, impacts on ecosystems, and albedo changes. Finally, the thawing of permafrost releases stored methane and carbon dioxide, amplifying greenhouse gas concentrations and contributing to further warming. More details will be provided in the climate models section at the end of this chapter.

Terrestrial Variables

Analyzing changes in forests, grasslands, and urbanization helps assess carbon sinks, habitat loss, and changes in the carbon cycle which are important as natural methods of sequestering carbon dioxide from the atmosphere. Specifically, the Normalized Difference Vegetation Index (NDVI) is a widely-used metric to quantify the health and density of vegetation using sensor data. It is calculated from spectrometric data at two specific bands: red and near-infrared. The spectrometric data is usually sourced from remote sensors, such as satellites. An open-access dashboard is Global Forest Watch. It is an online platform that provides data and tools for monitoring deforestation, forest degradation, and reforestation efforts. By harnessing cutting-edge technology, it allows anyone to access near real-time information about where and how land-use and forests are changing around the world. The project supports governments, businesses, researchers, and activists to make informed decisions on conservation strategies, policy decisions, and sustainable land management.

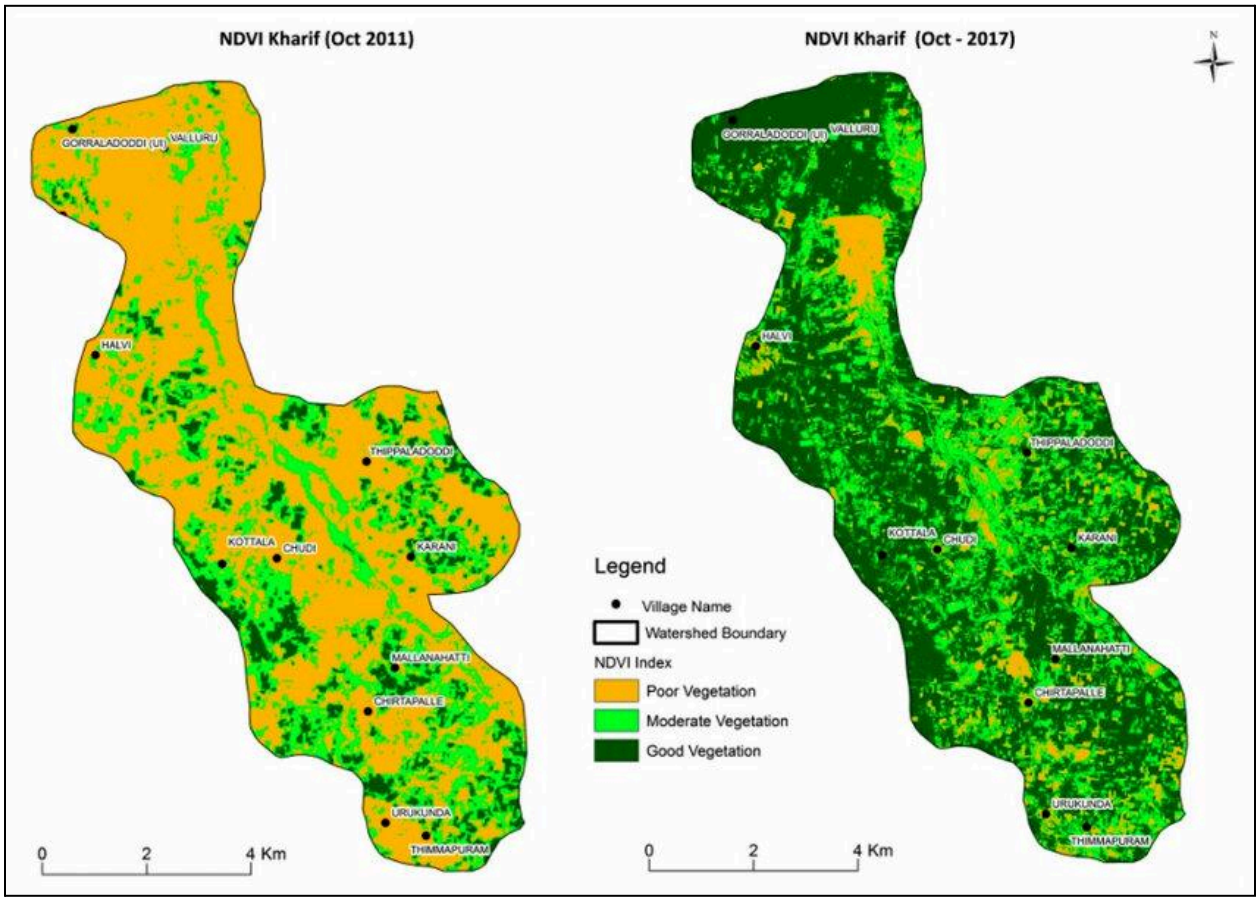


Figure: Satellite imagery of the vegetation cover in Kurnool District shows an increase in healthy vegetation, measured in NDVI, after the adoption of a watershed program encouraging the implementation of soil and water conservation activities. Source: https://www.researchgate.net/figure/Comparison-of-Normalized-Difference-Vegetation-Index-NDVI_fig4_356717309

Biodiversity and Ecosystem Health

In climate data science, biodiversity and ecosystem health reveal the broader ecological impacts of the climate crisis. For example, species distribution is a key factor, as warming temperatures and shifting habitats force plants and animals to migrate or adapt to new conditions. Tracking changes in species ranges provides insight into how ecosystems are being disrupted.

Furthermore, scientists study periodic events in biological life cycles and how these are influenced by seasonal and interannual variations in climate, as well as habitat factors. This field is called phenology and includes the timing of biological events such as flowering or animal migration, also shifts in response to changing climate conditions, serving as a clear indicator of environmental changes. Erratic weather patterns caused by the climate crisis disturb balanced ecosystems. For example, plants may bloom before butterflies emerge to pollinate them, or caterpillars may emerge before migratory birds arrive to feed them to their young. Additionally, the spread of vector-borne diseases is a growing concern, as the climate crisis alters the habitats of disease vectors like mosquitoes, expanding the risk of diseases such as malaria, dengue, and Lyme disease to new, previously unaffected regions.

Socioeconomic Variables

Tracking carbon emissions from industrial, transportation, and residential sources is central to understanding the drivers of the climate crisis. By monitoring these emissions, climate data scientists can evaluate the effectiveness of mitigation policies, such as carbon pricing and the transition to renewable energy. Changes in emission patterns reveal how successfully nations and regions are reducing their carbon footprints, allowing for more targeted policy interventions. Urban areas, which are responsible for the majority of global carbon emissions, are key targets for emission reduction strategies. For example, Google's **Environmental Insights Explorer** utilizes remote sensing data and satellite imagery to track emissions from transportation, industry, and buildings in order to estimate carbon footprints for cities globally.

Socioeconomic variables also play a pivotal role in climate data science, because impacts of the climate crisis often intersect with environmental and social justice issues. Socioeconomic factors, including population density, income levels, and access to resources, are essential for evaluating how communities are affected by climate impacts and their capacity to adapt. Low-income and marginalized communities often face disproportionate risks from the climate crisis due to limited access to infrastructure and resources. These variables are critical for assessing vulnerability and resilience, informing climate justice initiatives aimed at protecting the most vulnerable populations from the adverse effects of extreme weather, rising sea levels, and other climate-driven challenges.

Low-income and marginalized communities (also called environmental justice communities) are more likely to live in urban heat islands, neighborhoods where temperatures are significantly higher than surrounding regions due to poor urban planning, fewer trees and green spaces. As a result, residents face greater risks of heat-related illnesses and fatalities, especially during

heatwaves exacerbated by the climate crisis. Vulnerable communities are also more likely to be exposed to higher levels of pollution as they are often located near polluting industries such as factories, refineries, and landfills. This proximity exposes residents to higher levels of air, water, and soil pollution, which can lead to increased rates of respiratory illnesses, cancer, and other health problems. This is in particular damaging since polluted air traps more heat and thus worsen during extreme heat events. In addition, these vulnerable populations often live in areas more prone to flooding, hurricanes, or heatwaves and may lack adequate infrastructure or financial means to recover from climate-related disasters. For example, stormwater overflow events release untreated sewage and wastewater into nearby water bodies and occur when combined sewer systems are overwhelmed by heavy precipitation. The EPA estimates that more than 65% of active CSO outfalls are in Census block groups with high concentrations of people of color, low income, and other demographic indicators of vulnerability.

Analyzing these variables allows climate data scientists to build comprehensive models, identify trends, and predict future climate scenarios. This understanding is critical for guiding mitigation and adaptation strategies, helping policymakers and society prepare for the impacts of climate change. In the next section, we describe statistical concepts for common analysis techniques. Afterwards, we will explain how climate models work and the role they play in climate data science.

Statistical Methodological Concepts for Analyzing Climate Data
[Juliane]

Statistical methodologies form the backbone of climate data analysis, providing the tools needed to understand complex environmental patterns, make predictions, and inform policy and market decisions. Climate data is vast and diverse as introduced in the previous section. In this section, we describe several key statistical methodologies central to analyzing these datasets.

Time Series Analysis

Climate data is typically recorded over long periods, making time series analysis a foundational approach. Time series methods analyze data points collected sequentially over time, allowing researchers to detect trends, seasonal patterns, and cycles. Time series analysis is essential for identifying long-term trends in temperature, sea level rise, and extreme weather events, as well as projecting future changes.

Generally, time series data is modeled as a stochastic process Y_t. For exploratory analysis, time series can be decomposed into the sum of three components representing trend T_t, seasonality S_t and white noise error ε_t: $Y_t = S_t + T_t + \varepsilon_t$.
Scientists are also interested in autocorrelation, i.e. correlation of the time series with lags of itself, which is commonly used to determine whether the time series is stationary. Stationary time series do not have an overall trend, no relevant seasonal effect and their variance does not increase over time. Many statistical techniques require stationarity to make valid inferences and reliable predictions, e.g ARMA models consisting of autoregression (AR) and moving average (MA) to smooth out the data, reduce noise and improve forecast accuracy. For non-stationary time series autoregressive integrated moving average (ARIMA) models can help to forecast future values. The "integrated" (I) part indicates that the data values have been replaced with the difference between each value and the previous value. Periodic or seasonal components in the ARIMA framework can be eliminated by the seasonal differencing.

Time series analysis is also used for detecting unexpected values and anomalies that deviate significantly from the regular pattern in the time series. These can indicate errors, rare events, or significant changes. For example, statistical techniques like Z-scores or setting confidence intervals can highlight values that deviate beyond expected limits, marking them as anomalies.

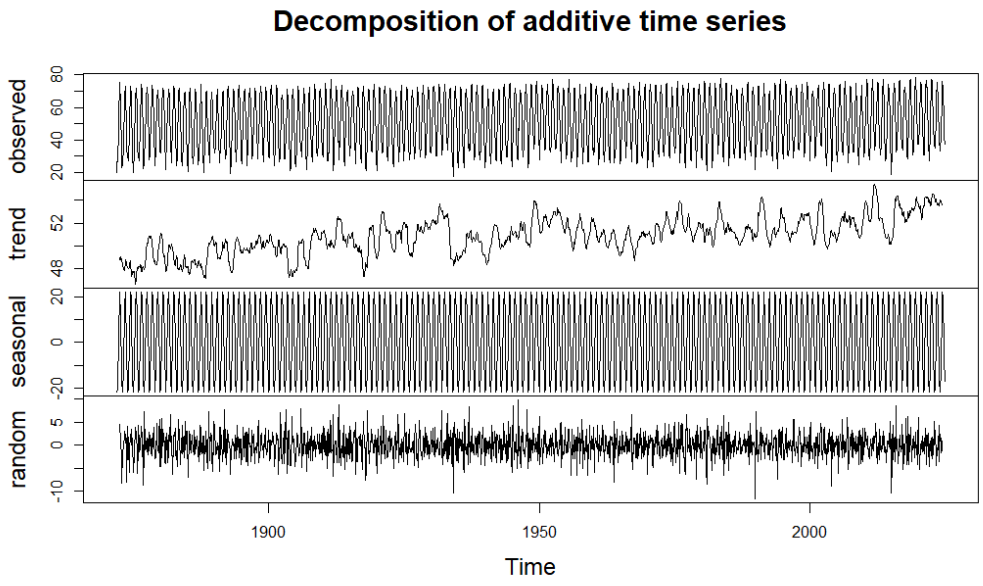
Example: **Average Temperature in Boston, MA**

In this example, we study the annual average temperature in Boston MA, which has been recorded by NOAA since 1872. We decompose the time series into trend, seasonality and random error term.

```
load("TempBostonTS.RData") # NOAA weather data, dtM
```

```
ts_decomposed <- dtM %>% select(meanTemp) %>%
  # Create time-series objects with freq 12 => Monthly data.
  stats::ts(frequency = 12, start = c(1872, 1)) %>%
  # Missing value imputation by Kalman smoothing
  imputeTS::na_kalman() %>%
  # Decompose a time series into seasonal, trend and error
  decompose(type = "additive")

plot(ts_decomposed)
```

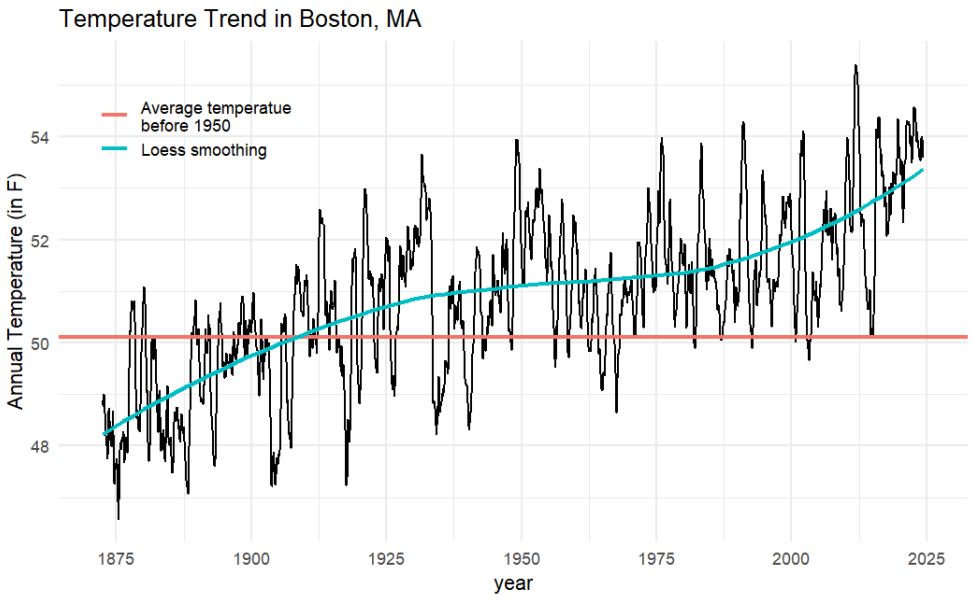


We look in more detail into the trend component of the time series by plotting the mean temperature before 1950 as reference and adding the locally estimated smoothed trend to visualize the rising temperatures over time.

```
# Calculate mean Temperature before 1950
ref1950 = dtM %>% filter(Year<1950) %>%
            summarize(mean(meanTemp, na.rm=TRUE)) %>% pull()

# Plot time series trend component
data.frame(trend=ts_decomposed$trend, date=time(ts_decomposed$x)) %>%
  ggplot(aes(x=date, y=trend)) + geom_line() +
  # Add reference: mean temperature before 1950
  geom_hline(aes(color = "Average temperatue\nbefore 1950",
             yintercept = ref1950), size = 1) +
  # Add loess smoothed trend
  stat_smooth(aes(color = "Loess smoothing"), method = "loess", se=FALSE) +
```

```
scale_x_continuous(breaks=seq(1850, 2025, 25)) +
labs(x="year", y="Annual Temperature (in F)",
     title="Temperature Trend in Boston, MA", color="") +
theme_minimal() + theme(legend.position = c(.15,.85))
```



Spatial Analysis and Geostatistics

Spatial analysis techniques are critical in climate science due to the geographic nature of environmental data, which includes variables like temperature, rainfall, and vegetation cover across different locations.

Different data types such as polygon (areal) maps and point pattern data, require different analysis techniques. Polygon or areal data represent areas, such as administrative regions, ecosystems, or climate zones, and are often visualized on maps with defined boundaries. Point pattern data, on the other hand, represent specific locations or "points" where measurements are taken, such as weather stations, pollutant sources, or species observations. Point data capture localized information and are useful for detecting fine-scale patterns. Both data types complement each other in climate science, where polygon maps offer a broad perspective on areal trends, and point data provide precise, localized insights.

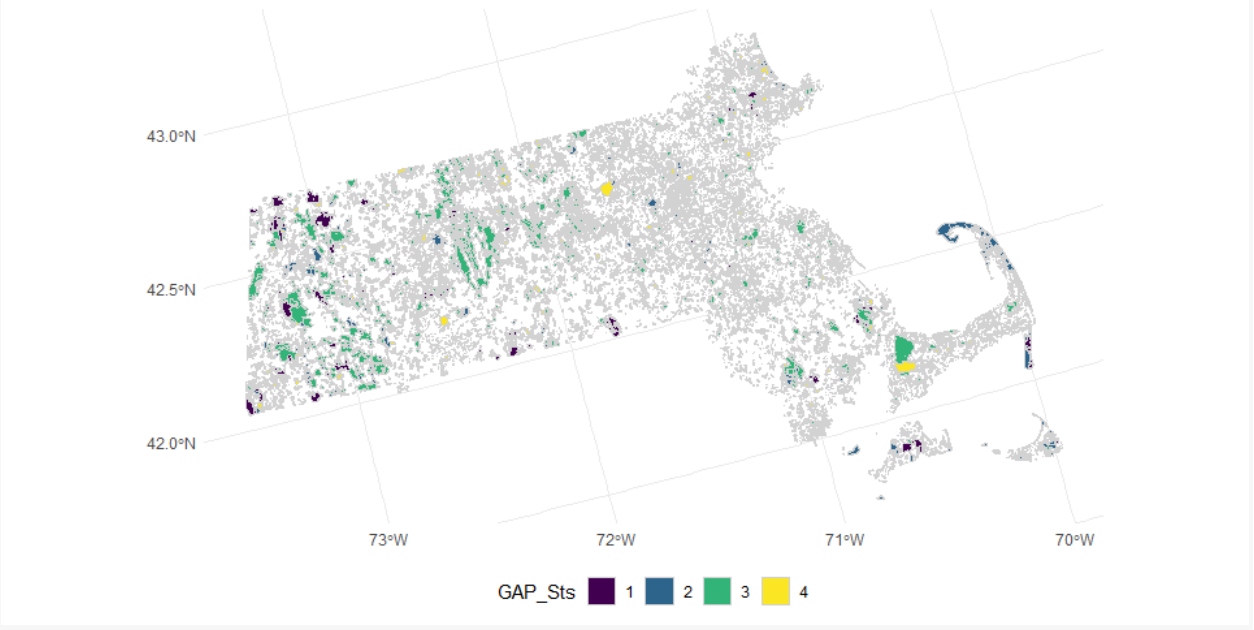
Example: **Biodiversity Protection Status in Massachusetts**

The 30x30 goal is a global conservation initiative aimed at protecting 30% of the planet's land and oceans by 2030. In this context, we analyze the proportion of protected areas in the Commonwealth of Massachusetts.

```r
# Read polygon map data
map_dt <- st_read("map_data/PADUS3_0Combined_StateMA.shp") %>% st_drop_geometry()
```

```
Reading layer `PADUS3_0Combined_StateMA' from data source using driver `ESRI
Shapefile'
Simple feature collection with 26347 features and 45 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 1830830 ymin: 2302826 xmax: 2137684 ymax: 2477959
Projected CRS: USA_Contiguous_Albers_Equal_Area_Conic_USGS_version
```

```r
# Plot GAP status
ggplot(data = map) + geom_sf(aes(fill = GAP_Sts), color="lightgrey") +
  scale_fill_viridis(discrete = TRUE) +
  theme_minimal() + theme(legend.position = "bottom")
```




```r
# helper function to calculates proportion of land in Massachusetts with
# input variable x = area (1 acre = 0.0015625 sq mi); MA area = 10,565 sq.mi
calc_prop <- function(x){ as.numeric(x * 0.0015625 / 10565)[1]*100 }

# Calculate proportion of protected land
map_dt %>% group_by(GAP_Sts) %>%
  summarize(pct = sum(GIS_Acres) %>% calc_prop()) %>%
  gt() %>% fmt_number("pct")
```

GAP_Sts	pct
1	1.75
2	4.20
3	11.32
4	3.75

The analysis shows that in Massachusetts protected areas (GAP 1+2 = 5.6%) and GAP3 = 11.2%, most are under state management. Only GAP 1 and 2 are typically considered protected from logging and mining, a far cry from the nationwide 30x30 pledge. However, GAP 3 areas are governed under multiple use mandates (e.g., wildlife, forestry and mining) and may have particular potential to advance biodiversity and climate protections more quickly through administrative mechanisms (Rosa and Malcom, 2020).

Kriging and similar geostatistical techniques predict climate variables in unmeasured areas by using the spatial correlations among nearby measurements, creating continuous surfaces of temperature, rainfall, or pollution data from discrete observations. Similarly, spatial autocorrelation analysis quantifies how climate variables are similar or dissimilar across distances, helping identify trends or hotspots of change.

In climate data science, spatial data is often obtained through remote sensing, a powerful technique that gathers information about the Earth's surface without direct contact, typically using satellites, drones, or aircraft. This technology captures extensive datasets across multiple spectral bands, enabling researchers to monitor environmental changes over time and across different locations. Advanced methodologies, including spectral analysis, machine learning, and change detection, facilitate the processing and interpretation of this data, allowing for the identification of trends and anomalies. Remote sensing is particularly valuable for providing real-time, global insights that support environmental monitoring, disaster response, and resource management efforts. For instance, it plays a crucial role in tracking deforestation, glacial melt, urban expansion, and variations in ocean temperatures.

Extreme Value Theory (EVT)

Extreme weather events, such as hurricanes, heatwaves, and heavy rainfall, are becoming more frequent and intense due to climate change. A mathematical representation of extremes, or events with a low probability of occurrence. An extreme weather event is an occurrence that deviates substantially from typical weather at a specific location and time of year. Extreme value theory provides a framework for modeling and predicting the occurrence of rare events, helping researchers estimate the probability, frequency and magnitude of future extremes.

Analysis of extreme weather is made more difficult by the fact that extreme events are, by definition, rare, and therefore reliable data is limited. Thus, analyzing these rare, high-impact events requires statistical approaches that focus on the extremes rather than the average. In

other words, we are focusing on the tail of standard distributional assumptions. The Generalized extreme value (GEV) method utilizes a distribution of standardized maxima

$$G(z) = exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}_{+}^{-1/\xi}\right],$$

with location μ, scale σ, and shape ξ, e.g. Gumbel (ξ → 0), Frechet (ξ >0), or Weibull (ξ <0) distribution. The GEV distribution is used when we are interested in the most extreme values over a fixed period, like the maximum yearly temperature or the strongest storm in a decade. Another approach, called Generalized Pareto (GP) focuses on the events that exceed a certain threshold, specifically considers the probability of exceeding a predetermined threshold

$$H(x) = 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_{+}^{-1/\xi}$$

with Exponential (ξ → 0), Pareto (ξ > 0) or Beta (ξ < 0) distribution. In calculating the model fit, it is useful to determine whether the model distribution remains the same as time progresses. Figure 1.3 shows how the probability of the extremes might change in the future under possible proposed climate scenarios. This change can be incorporated into the model as a change with respect to time; μ, x, σ, and ξ can be represented as some sort of function of time.

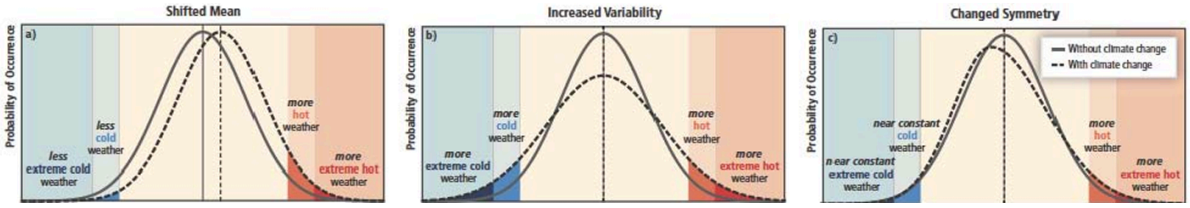


Figure 1.3: Temperature magnitudes and probability of occurrence in the context of a warmer climate. Plots obtained from IPCC Summary for Policymakers (2012), Figure SPM.3

Machine Learning and Predictive Modeling

With the increase in available climate data, machine learning has become an essential tool in climate science for predictive modeling and pattern recognition. Machine learning methods complement traditional statistical techniques, especially when handling large and high-dimensional climate datasets. These methods can be useful for automated extraction of information from images that would otherwise take a lot of manual labor. Algorithms such as random forests, gradient boosting and Support Vector Machines (SVMs), can identify complex, nonlinear patterns within large climate datasets, making them valuable for tasks like forecasting temperatures, rainfall, or even crop yields. Deep learning techniques, particularly Convolutional Neural Networks (CNNs), are also employed for analyzing spatial data, such as satellite imagery, to monitor deforestation, ice melt, or land-use changes.

Understanding basic concepts is essential for developing accurate and reliable models, especially in fields like climate data science. Supervised and unsupervised learning represent two main approaches to model training, each suited for different types of tasks. In supervised

learning, models are trained on labeled data, meaning each training example has an input and a known output.

Given the complexity of the models, one key concept is the division of data into training and test sets. The training data is used to build and "teach" the model, helping it learn patterns within the dataset, while the test data is used to evaluate the model's performance on unseen data, giving an indication of its generalization ability. Cross-validation is a technique used to improve model reliability by dividing the dataset into multiple subsets, allowing the model to be trained and tested on different combinations of data, which reduces overfitting and ensures that performance metrics are not biased by a single train-test split. These concepts form the backbone of reliable model building, ensuring that results are both valid and applicable in real-world climate scenarios.

Another foundational concept is distinguishing between regression and classification. In regression tasks, models predict continuous outcomes, such as temperature changes or CO_2 levels, where errors are often measured by metrics like mean squared error (MSE). In classification tasks, the goal is to predict discrete categories, such as whether an area is forested or deforested; here, errors are tracked by classification accuracy, precision, or recall. Error metrics like MSE (Mean Squared Error) and classification accuracy are ways to measure how well a model's predictions match reality.

In contrast, unsupervised learning is used with unlabeled data, where the goal is to find patterns, structure, or relationships within the data without specific output labels. This approach is valuable for tasks like clustering regions with similar climate patterns or grouping seasons with similar temperature profiles. Unsupervised learning algorithms can reveal hidden structures within large climate datasets, which can provide new insights or help define further supervised learning tasks. Together, these learning methods allow machine learning to adapt to a variety of climate data challenges, enhancing predictive accuracy and pattern recognition.

Further Reading

- Python data science book: https://jakevdp.github.io/PythonDataScienceHandbook/
- Time series analysis
    - https://r-statistics.co/Time-Series-Analysis-With-R.html
- Spatial analysis and geostatistics
    - L. Anselin, G. Morrison, A. Li, K. Acosta (2023). Hands-On Spatial Data Science with R. GitHub eBook (Creative Commons Licensed). Online version.
    - https://www.oreilly.com/library/view/applied-geospatial-data/9781803238128/
- Extreme value theory
    - https://grotjahn.ucdavis.edu/EWEs/extremes_primer_v9_22_15.pdf
- Machine Learning and Predictive Modeling
    - James, Witten, Hastie and Tibshirani (and Taylor). An Introduction to Statistical Learning with Applications in R or Python. PDF is available online: https://www.statlearning.com/

- Boehmke, B., & Greenwell, B. (2020). Hands-On Machine Learning with R. CRC Press. Freely available from: https://bradleyboehmke.github.io/HOML/

Climate Models [Barbara]

Modeling is used in climate science to simulate Earth's climate system to understand its past and how it is likely to evolve in the future, especially in the context of anthropogenic climate change. This helps provide not only a scientific basis for the understanding of climate change but also to guide policies and action to mitigate global warming and assess the impacts of climate change across the globe. While we have real-world observations of Earth's climate conditions before significant human activity and afterwards, we must rely on climate models to show what Earth's climate would be like if not affected by human activity to prove that the observed climate change is indeed caused by humans.

For example, the Intergovernmental Panel on Climate Change (IPCC) uses climate models to create their reports on the different climate scenarios based on greenhouse gas emissions and atmospheric concentrations and what climate action is needed to mitigate the damage caused. These simulations are useful for informing policies on what action should be taken and to determine if current action is enough for achieving the desired outcomes.

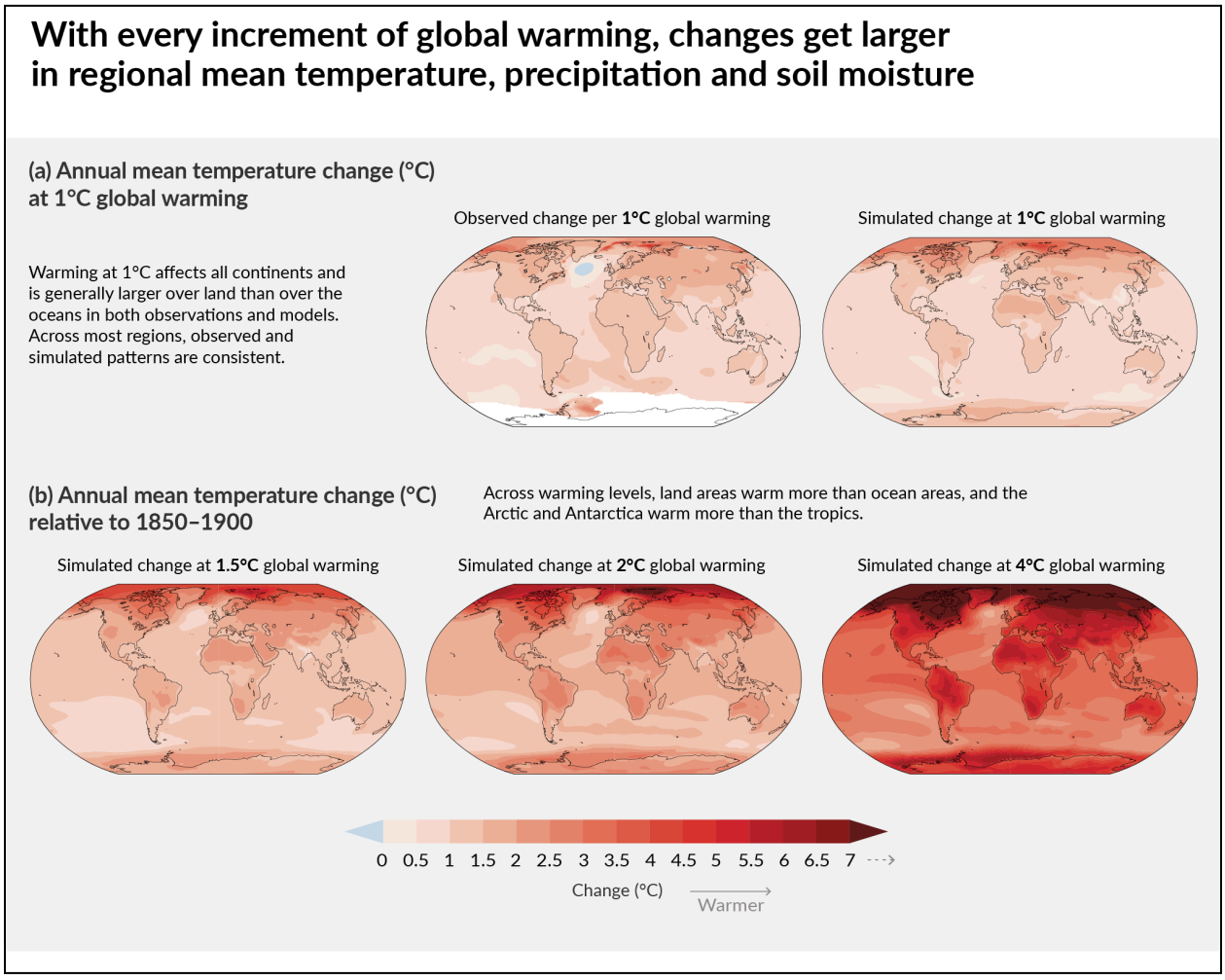


Figure: Changes in annual mean surface temperature, precipitation, and soil moisture. Source: https://www.ipcc.ch/report/ar6/wg1/chapter/summary-for-policymakers

Types of Climate Models

Climate models are essentially computer programs that simulate the interactions between different components of Earth's climate system using mathematical equations based on basic physical principles, such as the conservation of mass and energy, to approximate the dynamics. There are different kinds of climate models with varying use cases and levels of complexity, the simplest one being a one-dimensional radiative-convective or energy balance climate model. This model calculates the flow of energy between different components of a defined climate system to quantify the radiative balance of the Earth and the resulting radiative forcing (also called climate forcing).

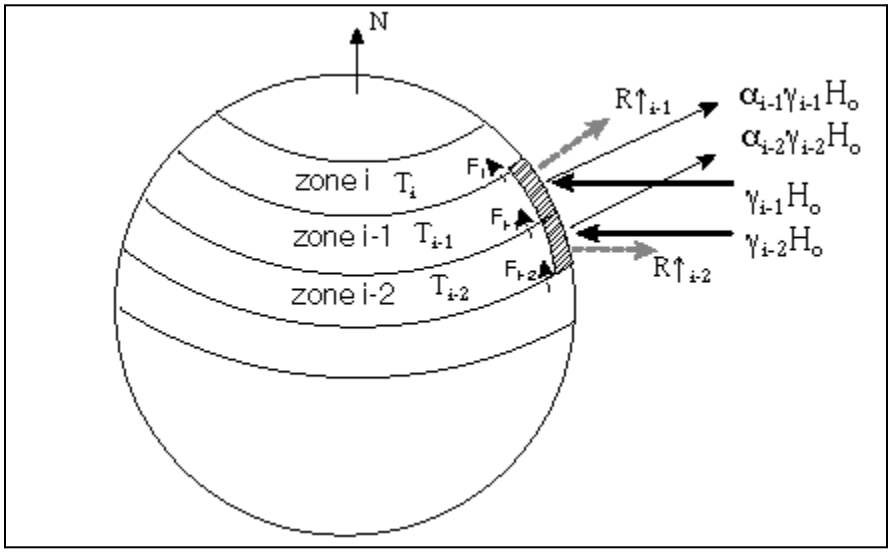


Figure: A diagram of a one-dimensional energy balance model. Source: http://www.shodor.org/master/environmental/general/energy/application.html.

There are also general circulation models (GCMs), sometimes called global climate models, that approximate the physical dynamics between many parts of the Earth climate system in three dimensions, such as the atmosphere, oceans, land surface, and ice. These models are the ones often used by the IPCC and other organizations when studying and discussing anthropogenic climate change. GCMs are probabilistic, meaning they calculate the statistics of weather pattern trends over an extended period of time and the probability that atmospheric or surface conditions will be in a certain range of values at the simulated time(s) for each variable.

Traditionally, this is calculated on a grid of cells representing a surface area of a region of (or whole) Earth, such as a 100 x 100 km square per cell. This determines the resolution of the model and to make the cell smaller (higher resolution) requires more computing power in a non-linear fashion. A size of about 10 km squared is considered high-resolution, but even that is not nearly small enough to explicitly model important atmospheric phenomena, such as clouds and precipitation. Therefore, approximations called parameterizations are used to simulate the atmospheric effects of these phenomena. These parameterizations are not perfectly able to capture the nuances of these small-scale atmospheric phenomena and so conditions such as deep convection (storm formation) are not modeled exactly like observed in the real world.

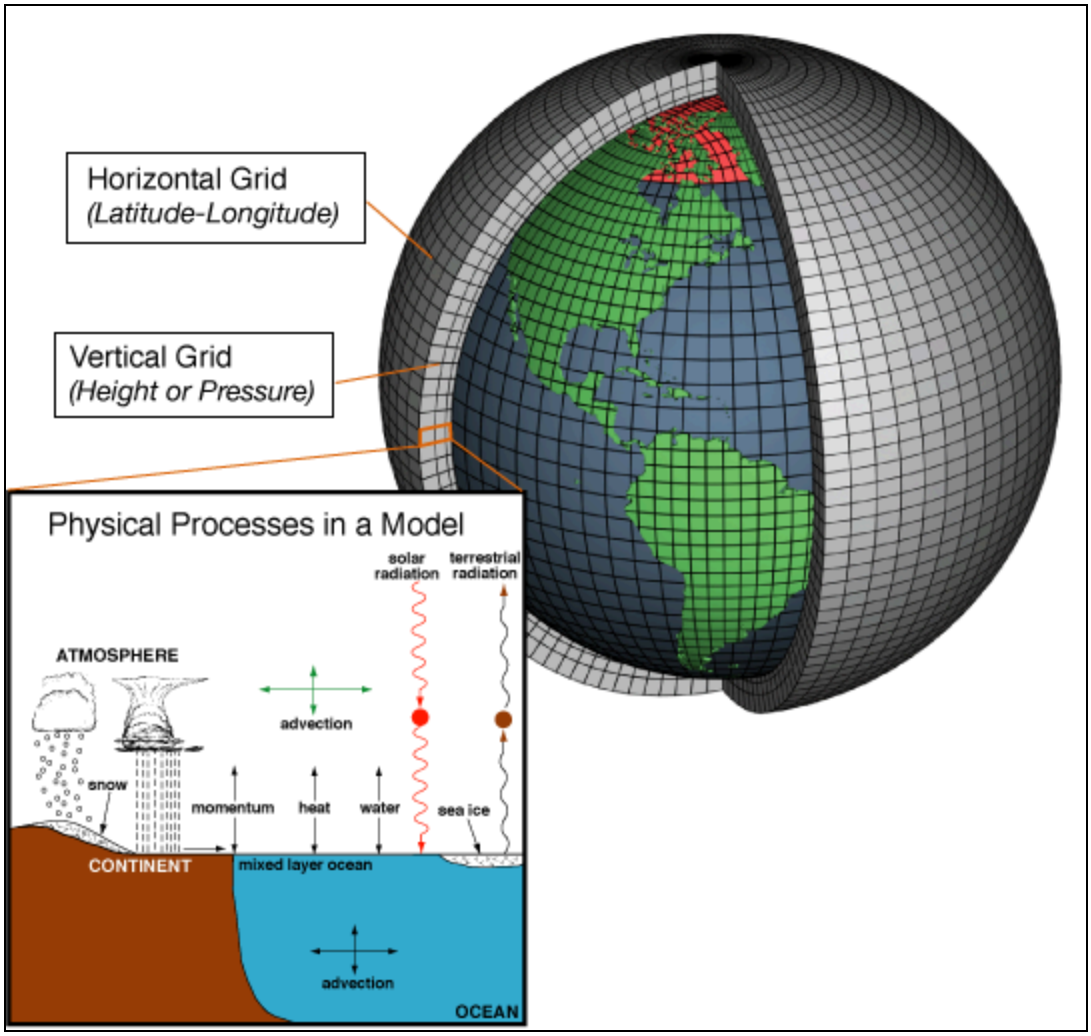


Figure: Climate model schematic. Source: https://celebrating200years.noaa.gov/breakthroughs/climate_model/modeling_schematic.html

To initialize the model, real-world climate data is assimilated into it to provide the initial conditions from which it will progressively calculate values (for a forecast) iterating over a set interval of time using physical equations. The quality of the data is essential because even a small difference from the real-world conditions at the beginning can result in significant differences over a long-scale simulation. This uncertainty about the exact conditions arises from the fact that there are so many variables to account for in how the climate will evolve over time that even small changes can accumulate and result in significant differences over a long period of time. This uncertainty has led to the popularity of ensemble forecasting, where multiple models are initialized with the same conditions or multiple runs are conducted with slightly different conditions to determine an average probability of what forecasted conditions will be.

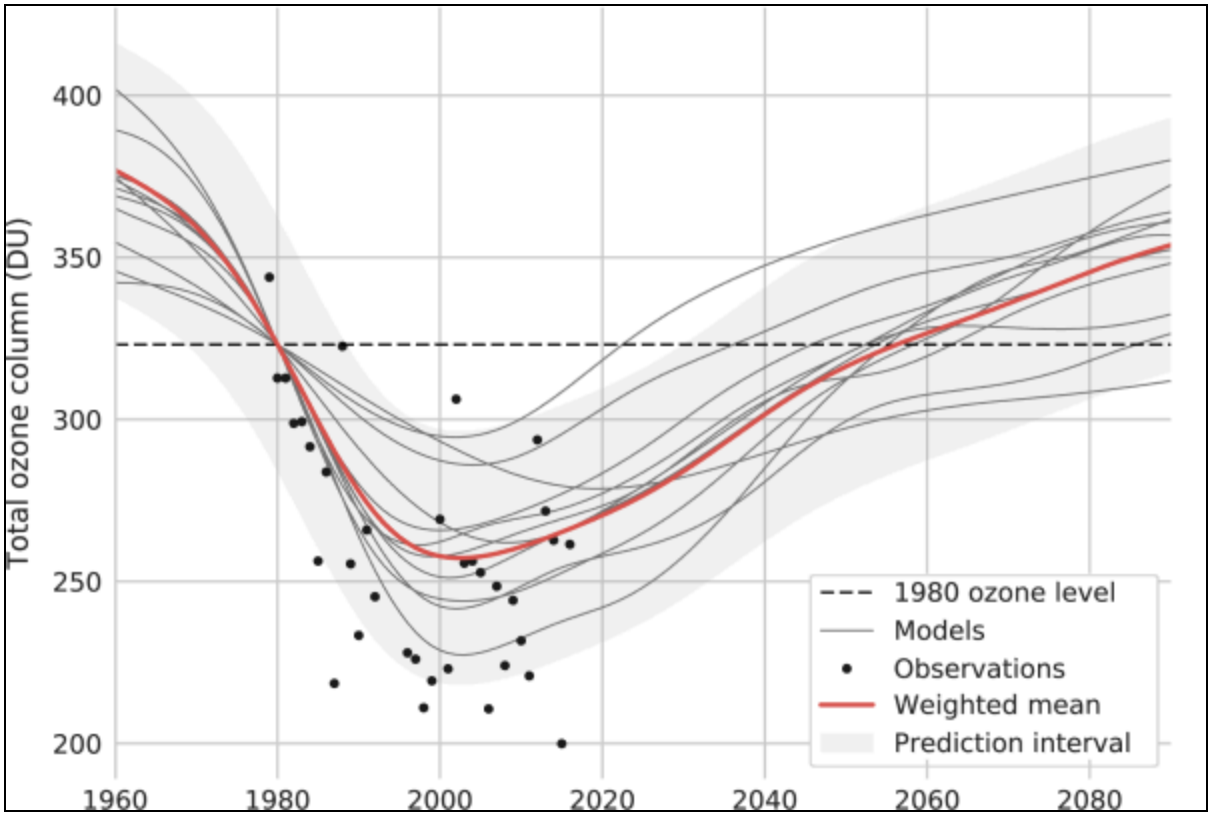


Figure: The output of an ensemble of climate models compared with real-world observations. The red line indicates the average of all the models and tends to be more accurate than any single model. Source: https://www.lancaster.ac.uk/data-science-of-the-natural-environment/blogs/how-to-use-ensembles-of-climate-models.

Because iterating over the data is computationally expensive, these models are typically run on supercomputers, especially those that are higher resolution. For example, doubling the horizontal resolution of a model requires 8 times as much computing power and if vertical resolution is included, that balloons to 16 times as much. Therefore, climate (and weather) models are often run on some of the most powerful supercomputers available.

Examples of Models

Climate data can also be used for models that make more short-term forecasts, such as flood or storm surge prediction models. One example is the National Hurricane Center's "Sea, Lake, and Overland Surges from Hurricanes" (SLOSH) model used by NOAA for over 3 decades to simulate storm surge from tropical cyclones. Storm surge is an increasingly important issue as sea level rise and more frequent storms lead to worse flooding. The model processes historical climate data and data on flooding and storm surges to predict the risk of storm surge due to a tropical cyclone across the US.

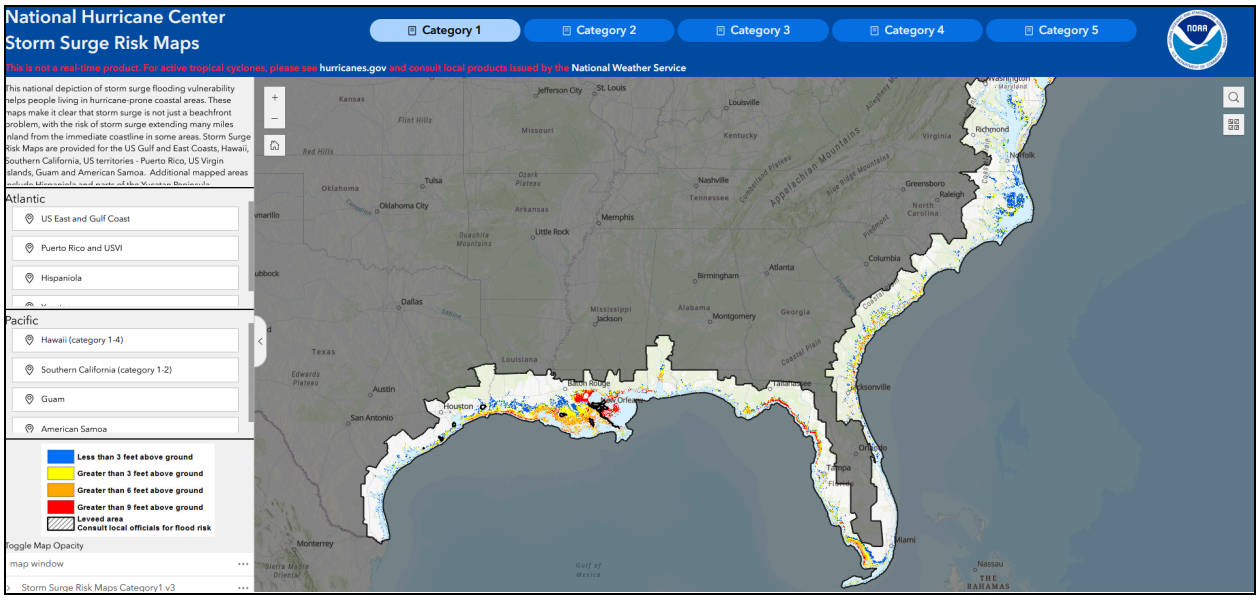


Figure: National Hurricane Center interactive storm surge risk map. Source: https://www.nhc.noaa.gov/nationalsurge/.

For anthropogenic climate change specifically, an open-source model by Project Drawdown available on GitHub is used to calculate the percentage contributions to greenhouse gas emissions by sector: electricity, food and agriculture, industry, transport, buildings, and other energy. This provides a data-driven basis for the list of feasible solutions to the impacts of anthropogenic greenhouse gas sources on the Project Drawdown website. As described on the organization's website, to evaluate a proposed solution, a review of relevant scientific literature is conducted and that information used to simulate the amount of greenhouse gas reduction, implementation costs, and operation savings calculated by their models. These findings are then reviewed and if they meet the standards for inclusion, they are added to the solution library. The documentation about the model is available on Project Drawdown's GitHub website for those who are interested in running it.

SOLUTION	SECTOR(S)	SCENARIO 1 *	SCENARIO 2 *
Abandoned Farmland Restoration	Land Sinks	12.48	20.32
Alternative Cement	Industry	7.70	15.56
Alternative Refrigerants	Industry / Buildings	42.73	48.75
Bamboo Production	Land Sinks	7.70	19.60
Bicycle Infrastructure	Transportation	2.73	4.63
Biochar Production	Engineered Sinks	1.36	3.00
Biogas for Cooking	Buildings	4.65	9.70
Biomass Power	Electricity	2.62	3.59
Bioplastics	Industry	1.33	2.48
Building Automation Systems	Electricity / Buildings	9.55	14.01
Building Retrofitting	Electricity / Buildings		
Carpooling	Transportation	9.06	11.07
Clean Cooking	Buildings	31.38	76.34
Coastal Wetland Protection	Food, Agriculture, and Land Use / Coastal and Ocean Sinks	1.20	1.62

Figure: The Project Drawdown model is used to give estimates of gigatons CO2 equivalent reduced / sequestered (2020–2050) for each solution listed in their solutions library for two scenarios depending on how much action is taken to reduce greenhouse gas emissions. Source: https://drawdown.org/solutions/table-of-solutions

Another model that can be downloaded and run is the University Corporation for Atmospheric Research (UCAR) Community Earth System Model (CESM), a fully coupled global climate model that includes atmosphere, land surface, ocean, and sea ice components. There is a free asynchronous distance learning course available on UCAR MetEd on how to run the model, modify components, and analyze the outputs which is linked in the Further Reading section. The user should have some familiarity with unix-like operating systems, Python libraries, and compiling code, specifically Fortran and C. The model download page links to a quickstart guide which details the download and compiling process further. The CESM also has an active community involved in its use and development through working groups and meetings which are a great resource for advancing one's knowledge and understanding of climate modeling as well as ensuring that the model is kept up-to-date.

Examples of Data Sources

Climate data is important for understanding how Earth's climate was in the past and how it will evolve in the future, especially considering the significance of climate change. Data from natural sources and human observations of weather patterns over long periods of time are instrumental in informing climate scientists about historical trends in important variables such as surface temperature, atmospheric concentrations of greenhouse gasses, and precipitation patterns.

There are many natural sources of historical climate data including ice cores, tree rings, and ocean and lake sediments. The unique ability of these records to extend far back through time makes them invaluable for paleoclimatological uses. Paleoclimatology, the study of Earth's climate thousands or millions of years ago, is important because it gives perspective to the

current climate change we are experiencing right now and provides insights into past and current drivers of different eras in Earth's climate. This data can also be used to validate model outputs in tandem with satellite observations of atmospheric phenomena and contemporary trends in climate, such as to better model the influence of the stratosphere on the North Atlantic Oscillation.

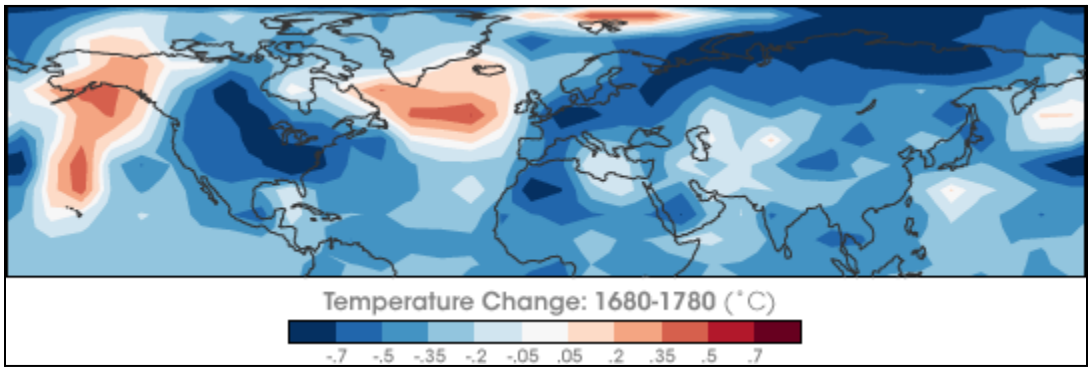


Figure: Lowered temperatures in the Northern Hemisphere due to the Maunder Minimum. Source: https://earthobservatory.nasa.gov/features/Paleoclimatology_Understanding

A great resource for paleoclimatological data is NOAA's webpage for natural records of past climate, featuring datasets from an wide set of natural sources spanning an extensive timescale. NOAA also has a climate data primer webpage, linked in the "Further Reading" section, that provides several tools for climate data analysis and visualization along with estimated difficulty ratings and explanations of what each is used for and how.

In more recent times, observations by humans of weather conditions have also become important for climate science. There are many sources of this kind of data online, though because of its significantly shorter timescale compared to paleoclimatological data, this data is often used for conducting analyses focused on climate change and human activity in the industrial or contemporary era. Here are a few examples of sources of observational climate data:

- National Climatic Data Center Climate Data Online: https://www.ncei.noaa.gov/cdo-web/
- Climate Toolbox Historical Climate Tracker: https://climatetoolbox.org/tool/Historical-Climate-Tracker
- NCEI Global Historical Climatology Network: https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily
- World Bank Group Climate Change Knowledge Portal: https://climateknowledgeportal.worldbank.org

Further Reading

1. IPCC AR6 Model List: https://psl.noaa.gov/ipcc/cmip6/models.html
2. IPCC 2007 Model Comparison: https://www.ipcc.ch/site/assets/uploads/2018/02/ar4-wg1-chapter8-1.pdf

3. IPCC Sixth Assessment Report, Chapter 7: https://www.ipcc.ch/report/ar6/wg1/chapter/chapter-7/
4. National Hurricane Center's SLOSH model: https://www.nhc.noaa.gov/nationalsurge
5. Project Drawdown GitHub: https://github.com/ProjectDrawdown
6. Project Drawdown methodology page: https://drawdown.org/solutions/methods
7. Project Drawdown model documentation: https://projectdrawdown.github.io/solutions/index.html
8. UCAR's CESM: https://www.cesm.ucar.edu/
9. CESM model releases: https://www.cesm.ucar.edu/models/releases
10. CESM on GitHub: https://github.com/ESCOMP/CESM
11. CESM MetEd distance learning course: https://www.meted.ucar.edu/education_training/lesson/1363
12. UCAR guide to climate data analysis tools and methods: https://climatedataguide.ucar.edu/climate-tools
13. Exascale computation: https://sab.noaa.gov/wp-content/uploads/2021/08/ExascaleWeatherClimateNOAAComputationsEKalnay_Final-1.pdf
14. NOAA Climate Data Primer: https://www.climate.gov/maps-data/climate-data-primer/visualizing-climate-data