Topics for today's workshop:

1. Why Climate Data matters?
2. Data collection
3. Essential variables to measure, some jargons
4. Data Challenges and cleaning approaches
5. Getting started: Datasets and Resources

1. Why Climate Data matters?

Start with WHY

> Interconnectedness - Everything interacts with everything while analysing environmental factors

> Narrowing down to a single field

- Ranchers, farmers, and outdoor-recreation businesses regularly monitor drought conditions to see if the environment has sufficient water for plants and animals.

- Weather enthusiasts like to explore **extreme** storms and record-setting events.
- People who live near the coast consider how **sea level rise** might affect them.
- Weather derivatives

1. Why Climate Data matters?

Climate is the long-term pattern of weather conditions.

For climate, the expression “long-term” usually means 30 years.

Climate scientists have agreed that 30 years is a sufficient length of time to establish the usual range of conditions for different times of the year

1. Why Climate Data matters?

Climate is the long-term pattern of weather conditions.

For climate, the expression “long-term” usually means 30 years.

Climate scientists have agreed that 30 years is a sufficient length of time to establish the usual range of conditions for different times of the year

If you wanted the longest-running climate dataset, would you trust:
A. Satellites
B. Weather Stations
C. Advanced ML models

1. Why Climate Data matters?

Climate is the long-term pattern of weather conditions.

For climate, the expression “long-term” usually means 30 years.

Climate scientists have agreed that 30 years is a sufficient length of time to establish the usual range of conditions for different times of the year

If you wanted the longest-running climate dataset, would you trust:
~~A. Satellites~~
~~B. Weather Stations~~
~~C. Advanced ML models~~
D. Centuries of farmers and festival organizers

WHY?

HOW?

WHAT?

WHY ?

Burgundy harvest dates matter because harvest timing affects wine quality, farm income, and local identity—so people recorded it carefully long before modern instruments existed.

WHY ?

Kyoto cherry blossom timing matters because it’s tied to culture and public life—so the dates were preserved across generations, creating a long memory of seasonal change

HOW ?

Both are “**biological thermometers**”: grape ripening and blossom timing respond strongly to temperature across the growing season, so the calendar itself becomes a **proxy dataset.**

Eureka moment: When you plot these dates through time, you’re not graphing flowers, you’re graphing climate’s influence on biology and society.

Interesting reads:

https://www.egu.eu/news/494/burgundy-wine-grapes-tell-climate-story-show-warming-accelerated-in-past-30-years/

https://time.com/6957844/cherry-blossoms-climate-change-peak-bloom-shift/

https://science.nasa.gov/earth/climate-change/climate-change-is-shifting-wine-grape-harvests-in-france-and-switzerland/

https://www.nationalgeographic.com/environment/article/cherry-blossom-peak-bloom-climate-change

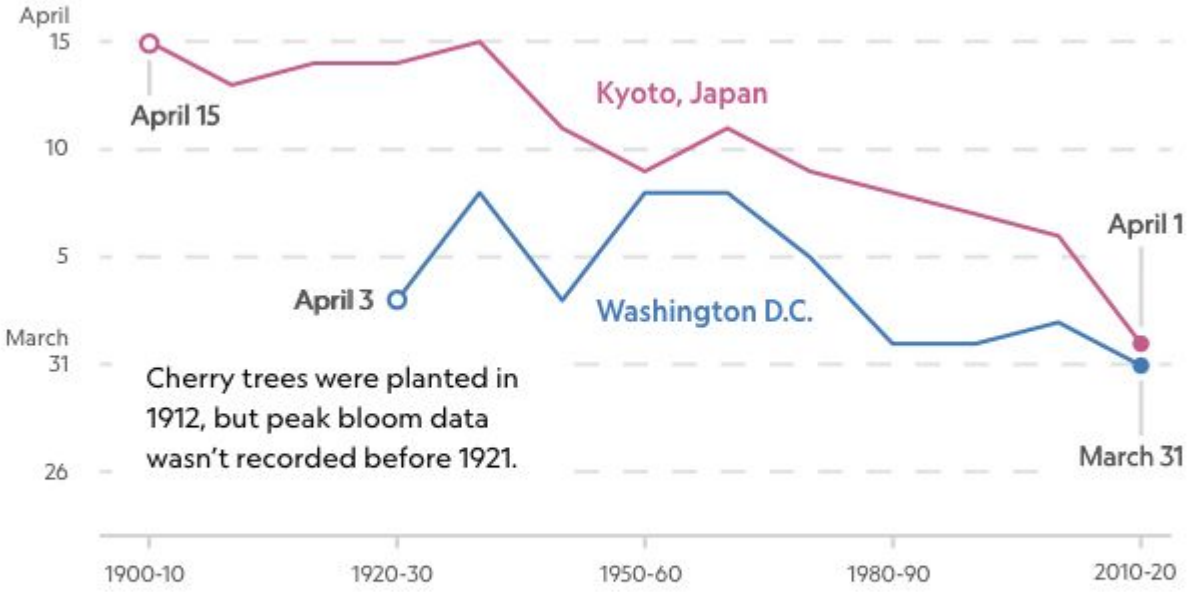
WHAT ?

For Burgundy, long historical records show harvests shifting earlier, with studies reporting grapes being picked markedly earlier in recent decades, consistent with accelerated warming.

For Kyoto, widely cited analyses note very early peak bloom years in the modern era (including exceptionally early peaks), which stand out when compared against the long historical record.

Date of peak cherry blossom bloom

Average by decade, from 1900 to 2020



Lucas Petrin, NGM Staff

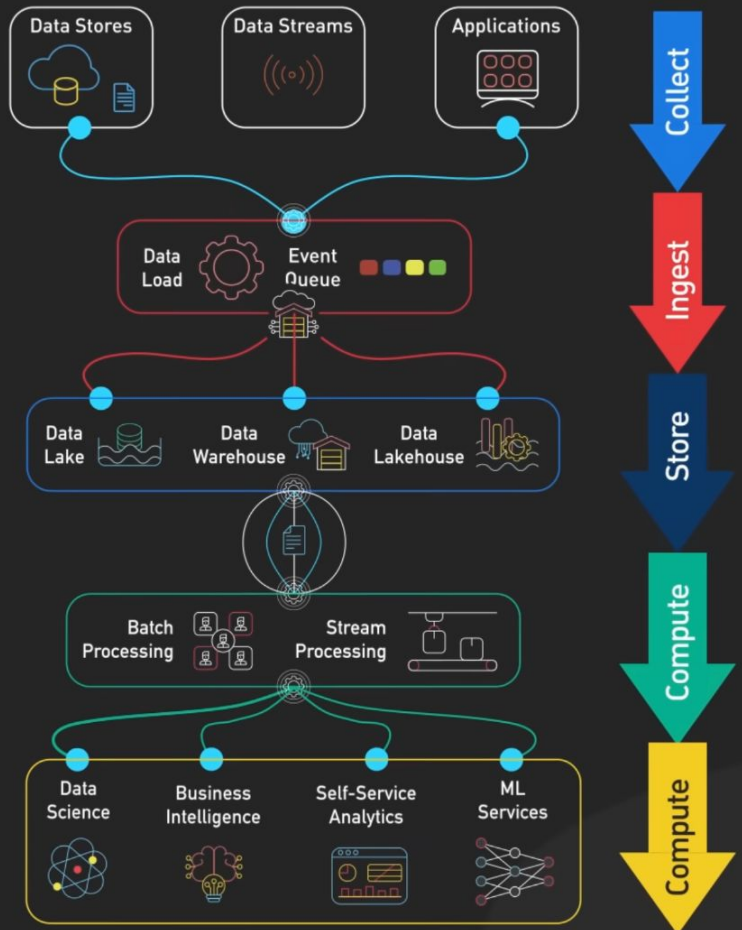
Sources: Yasuyaki Aono and Shizuka Saito, International Journal of Biometeorology, 2010; National Parks Service; Environmental Protection Agency; Japan Meterological Agency; National Ocean and Atmospheric Administration

2. Data Collection

- ETL -> Extract - Transform - Load
- ELT -> Extract-Load-Transform
- **ELTL ->Extract, Load, Transform, Load**
- Streaming


On
TEMP
TIME
PREHEAT
START

The Amazing Data Pipeline



Extract

Transform

Load

Personally, food analogies are the most palatable. :)

Rest time

Let’s brainstorm and ask questions!

Types Of Data Collection

- Vegetation-Phenology-&-Health-Analysis-(NDVI/EVI)
- Soil-Moisture-Anomaly-Analysis: Crop stress, warming-driven drying in semi-arid areas
- Land-Surface-Temperature-(LST)-Analysis: Identifying hot-spots in urban heat analysis
- Wildfire-&-Burn-Severity-Analysis
- Land-Use-&-Land-Cover-Change-(LULC)-Modeling
- Glacier-&-Ice-Sheet-Mass-Balance
- Sea-Ice-Extent-&-Concentration-Analysis
- Permafrost-Deformation-Analysis-(InSAR)
- Snow-remote-sensing
- Sea-Surface-Temperature-(SST)-Analysis
- Ocean-Altimetry-&-Sea-Level-Rise-Analysis
- Ocean-Color-&-Chlorophyll-Analysis
- Greenhouse-Gas-Monitoring-(GHG)
- Aerosol-Optical-Depth-(AOD)-Analysis
- Precipitation-&-Cloud-Microphysics
- Groundwater-Storage-Analysis-(Gravimetry)

Climate Data Sources and methods

1. **Weather Station Observations**
   Weather station provide the most accurate climate measurements available. However, their utility is limited by practical constraints

   Advantages:

   - High measurement accuracy
   - Direct observational data
   - Temporal continuity at specific locations

   Limitations:

   - Expensive establishment and maintenance costs
   - Limited global coverage, especially in developing regions
   - Concentrated in accessible areas (near cities, in valleys)
   - Inadequate spatial representation for forest stands and ecological plots

Climate Data Sources and methods

2. Interpolated climate data

Spatial interpolation methods extend point observations from weather stations to continuous climate surfaces, providing global or regional coverage at specified spatial resolutions

Methodology: Interpolation algorithms estimate climate values at grid cells based on nearby weather station observations, incorporating spatial relationships and terrain influences.

Quality factors:

- Density of weather station network
- Interpolation algorithm performance
- Spatial resolution (finer resolution = higher detail but greater uncertainty in datasparse areas)

Spatial resolution challenges:

In mountainous regions, climate changes dramatically with elevation over short distances. Coarse-resolution datasets may miss critical local climate variation

Climate Data Sources and methods

3. Climate Models:

Climate models represent Earth's climate system through differential equations based on fundamental physics, fluid dynamics, and chemistry. They simulate both current climate conditions and future projections under different emission scenarios.

Climate models exist on a continuum of complexity:

- Simple models: Radiative heat transfer models treat Earth as a single point, averaging energy balance
- Intermediate models: Radiative-convective models add vertical atmospheric stratification Expand horizontally to capture spatial variability
- Complex models (Global Climate Models - GCMs): General Circulation Models (GCMs) solve full equations for mass and energy transfer and radiative exchange Discretize equations for fluid motion and energy transfer, integrating over time Include both atmospheric and oceanic processes in coupled models
- Atmospheric GCMs (AGCMs) model atmospheric processes with sea surface temperatures prescribed as boundary condition
- Coupled Atmosphere-Ocean GCMs (AOGCMs) combine atmospheric and oceanic models internally, capturing atmosphere-ocean interactions critical for realistic climate simulation

3. Essential variables to measure, some jargons

Climate Variables and Time Scales

Essential Climate Variables (ECVs) are climate measurements that critically characterize Earth's climate system. Terrestrial ECVs include:

1. Temperature (mean, minimum, maximum)
2. Precipitation (rainfall and snowfall)
3. Wind speed and direction
4. Radiation (shortwave and longwave)
5. Atmospheric pressure
6. Water vapor

For ecological and forest modeling applications, temperature, precipitation, and their derived variables are most commonly used

Climate Variables and Time Scales

Observed and Derived Climate Variables

Primary observed variables are measured directly at weather stations:

- Mean monthly, seasonal, and annual temperature
- Mean minimum and maximum temperature
- Total monthly, seasonal, and annual precipitation

Derived biologically-important variables are calculated from primary observations:

- Degree-days (growing degree-days, chilling degree-days)
- Frost-free periods
- Extreme temperature events
- Climatic moisture deficit (CMD)
- Bioclimatic indices

Time Scales and Data Availability

Climate variables are available at multiple temporal resolutions, each with distinct advantages and limitations.

Daily climate data:

- Most detailed temporal resolution
- Generated from weather station observations
- Used to derive monthly, seasonal, and annual variables
- Can be used to calculate biologically-important variables
- Disadvantage: Large file sizes make distribution and modeling impractical

Monthly, seasonal, and annual data:

- Most commonly used in ecological models
- Moderate file sizes, practical for model applications
- Suitable for species distribution models and climate niche modeling

Time Scales and Data Availability

Climate variables are available at multiple temporal resolutions, each with distinct advantages and limitations.

Annual climate variables:

- Special significance for forest and ecological applications
- Mean annual temperature (MAT) and mean annual precipitation (MAP) are most widely used
- Annual degree-days, growing season indices, and climatic moisture deficit are biologically critical

Decadal and normal period (30-year) averages:

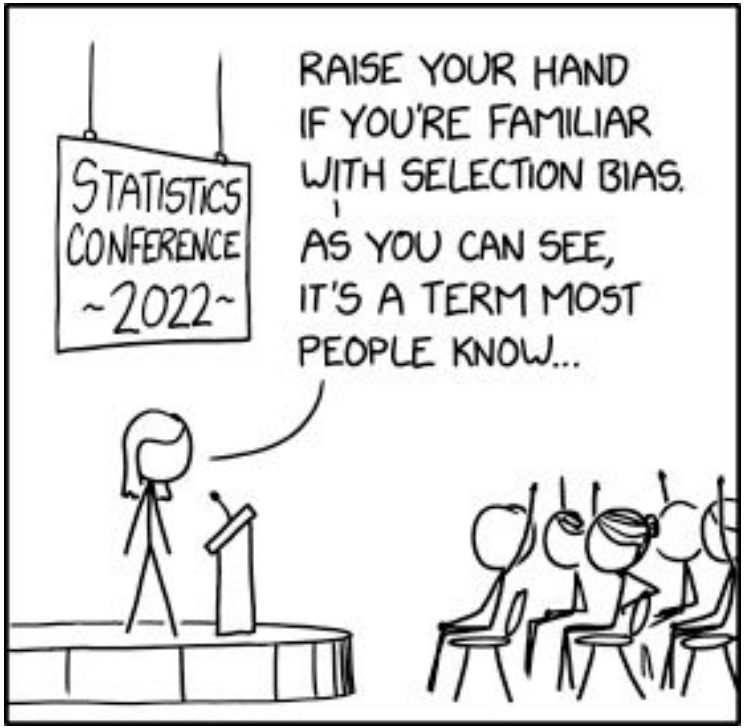
- Used for climate change analysis and establishing baseline conditions
- Follow the convention that climate represents long-term conditions (typically 30- year averages per the World Meteorological Organization)

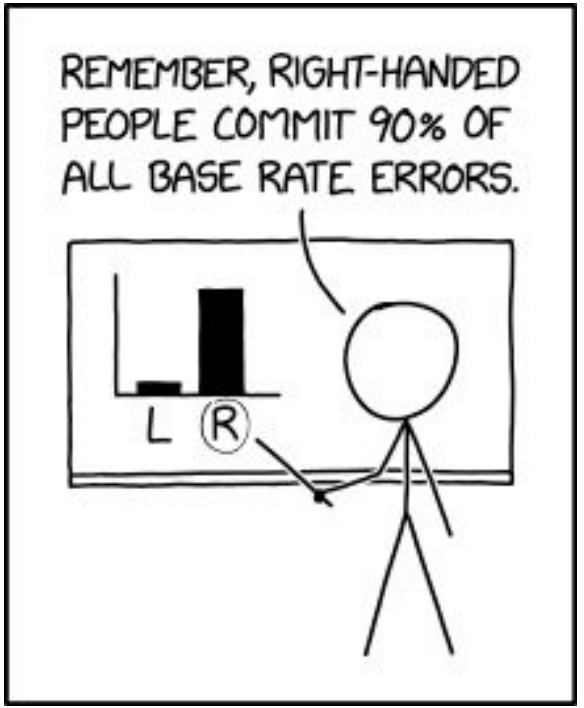
Recommendations for Climate Data Use

- Application and scale: Match spatial resolution and variable selection to modeling requirements
- Topographic complexity: Use topographically-informed datasets (PRISM, ClimateNA) in mountainous regions
- Temporal requirements: Select appropriate time intervals (daily, monthly, annual) based on ecological question
- Variable availability: Ensure required climate variables (temperature, precipitation, derived variables) are available
- Uncertainty assessment: Use multiple datasets or GCMs to characterize uncertainty
- Validation: Compare model projections against independent observations in study region
- Data lineage: Understand dataset methodology, training data, and interpolation approach
- Currency: For forward projections, use recent CMIP6 outputs rather than CMIP5 where available

4. Data Challenges and cleaning approaches

4. Data Challenges and cleaning approaches





- Illu. by xkcd

4. Data Challenges and cleaning approaches

Environmental change has differentiated impacts on different genders, and there is increasing evidence that women, girls, and other marginalized communities disproportionately suffer from climate change and environmental disasters.

Gender Data: Gender data is any type of data that can at least be disaggregated by male/female.

4. Data Challenges and cleaning approaches

Sex-Disaggregated Data	Gender-Inclusive Data	Diverse and Comprehensive Data	Gender-Balanced Data Collection
Data are collected and presented by sex* as a primary and overall classification	Data reflect gender issues	Data are based on concepts and definitions that adequately reflect the diversity of women and men and capture all aspects of their lives	Data collection methods take into account stereotypes and social and cultural factors that may induce gender bias in the data

Report: Gender data & Climate

4. Data Challenges and cleaning approaches

Elements to consider while working with Gender data:

1. Does the available data allow to analyze in depth the denfer gap that we want to highlight?

YES	No
Who produced it? What was the purpose? What is the methodology of data collection	Start Small Search for proxy datasets

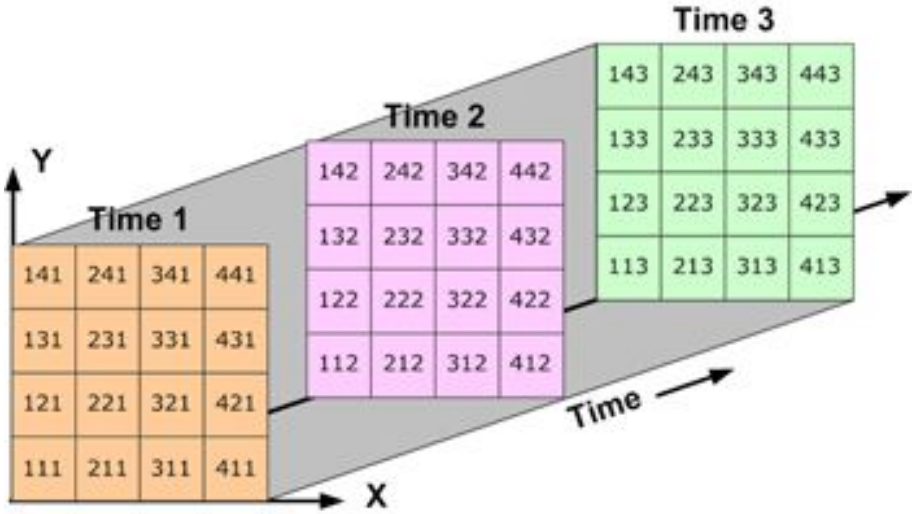
Resource: Mapping Gender data gaps in the Environment and climate change : https://data2x.org/wp-content/uploads/2023/10/Data-Gaps-in-Environment-and-Climate-Change-WR-251023.pdf

Some data samples

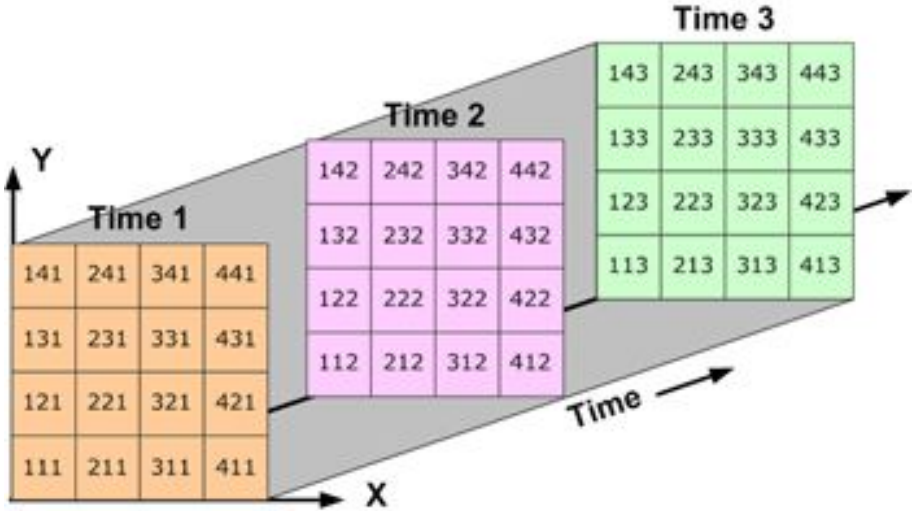
1. NetCDF (Network Common Data Form)
   NetCDF is a self-describing, hierarchical data format developed and maintained by UCAR (University Corporation for Atmospheric Research). It is the most widely used format in the climate science community

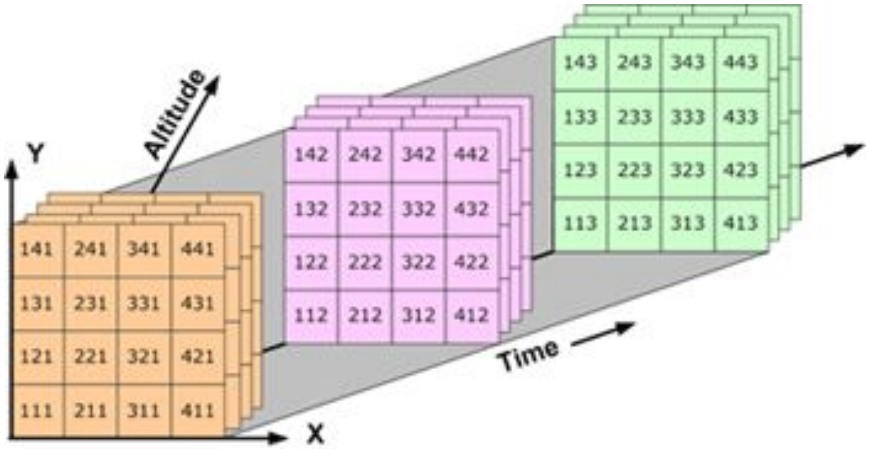
   Key Characteristics:
   - Multi-dimensional support: Stores data with multiple dimensions (latitude,
   - longitude, time, pressure levels)
   - Self-describing: Metadata is embedded within the file structure
   - Platform-independent: Data can be read on any machine with NetCDF libraries
   - Hierarchical structure: Supports nested variables and attributes
   - Versions: NetCDF3 (classic), NetCDF4 (based on HDF5 with enhanced compression)

   Data Types Supported:
   - Gridded geospatial data (rasters)
   - Terrain data
   - Time-series data (single station over multiple time periods)
   - Climate variables: temperature, humidity, precipitation, wind direction



READ :
https://desktop.arcgis.com/en/arcmap/latest/manage-data/netcdf/essential-netcdf-vocabulary.htm






```
>> netcdf_file =
'https://opendap1.nodc.no/opendap/physics/point/cruise/nansen_legacy-sing
_profile/NMDC_Nansen-Legacy_PR_CT_58US_2021708/CTD_station_P1_
EG01-1_-_Nansen_Legacy_Cruise_-_2021_Joint_Cruise_2-1.nc'

>> xrds = xr.open_dataset(netcdf_file)

>> xrds
```

xarray.Dataset

▸ Dimensions: (**PRES**: 320)

▾ Coordinates:

PRES (PRES) float32 1.0 2.0 3.0 ... 318.0 319.0 320.0

▸ Data variables:
(33)

▸ Indexes: (1)

▸ Attributes: (73)

Some data samples

2. GRIB (Gridded Binary ro General Regularly-distributed Information in Binary Form)

GRIB is the World Meteorological Organization (WMO) standard format for weather and climate data, first defined in 1985. It was designed to exchange and store large volumes of gridded data efficiently.
Each record contains one field (e.g., temperature on one level/time) plus its own metadata, optimized for transmission and storage efficiency.

Key Characteristics:

- Gridded data: Raster-based structure representing measurements across geographic grids
- Machine-independent: Requires encoding/decoding software
- Highly compressed: Efficient storage for large numerical datasets
- Multi-parameter support: Can contain multiple data layers (bands) in a single file
- Versions: GRIB1 (older), GRIB2 (current standard with enhanced capabilities)

Typical Data Variables:

- Temperature, wind speed, precipitation, wave height
- Pressure fields and wind components
- Geopotential heights

Some data samples

3. GeoTIFF (Georeferenced Tagged Image File Format)

Overview
GeoTIFF is an extension of TIFF format that includes geospatial referencing information, making it ideal for raster spatial data

Key Characteristics:

- Georeferenced: Contains geographic coordinate system information
- Raster-based: Grid-structured spatial data
- TIFF-compatible: Standard image format with geospatial extensions
- Widely supported: Compatible with most GIS software
- Single time slice: Typically represents one time period (unlike NetCDF time series)

Use Cases:

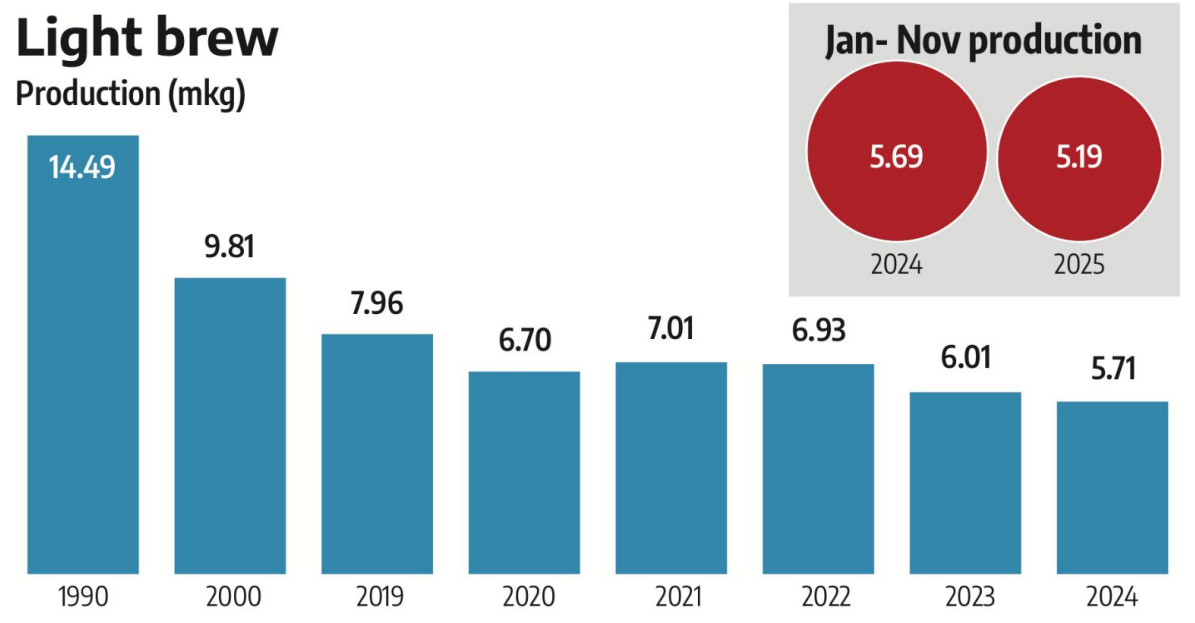
- Land cover classification
- Sea surface temperature maps
- Satellite imagery and remote sensing data
- High-resolution precipitation observations (e.g., CHIRPS)
- Vegetation indices and environmental parameters
- Individual time steps from climate datasets

<Add GeoTiff visual sample>

Supplementary Data Formats

1. ASCII: Simple text-based format for tabular data with one measurement per line
2. CSV (Comma Separated Values) : Tabular format with comma-delimited values, easily imported into spreadsheets and databases
3. JSON (JavaScript Object Notation): Structured text format with key-value pairs, increasingly used for web-based climate data access.

Using LLMs to extract Climate Data



Data Cleaning

Common data quality issues in Climate datasets

1. Measurement errors: Electronic sensor faults, calibration drift, and instrument malfunctions
2. Transmission errors: Data loss or corruption during collection and transmission
3. System changes: Equipment upgrades, station relocations, or methodology changes
4. Outliers: Extreme or anomalous values caused by equipment failure or data entry errors
5. Missing values: Data gaps due to sensor downtime, communication failures, or collection gaps
6. Inconsistencies: Unit mismatches, formatting variations, or temporal gaps
7. Redundant records: Duplicate entries from multiple sources or repeated transmissions

Data Cleaning

Data Profiling and Initial Assessment

- **Structure analysis:** Examine rows, columns, and data types
- **Descriptive statistics**: Calculate mean, median, std deviation, quartiles, and range for each variable
- **Missing data pattern:** Quantify missing values and identify temporal or spatial patterns
- **Outlier screening:** Identify extreme values requiring investigation
- **Temporal analysis**: Check for gaps, duplicates, and temporal ordering
- **Variable relationships:** Analyze correlations between related climate variables
- **Distribution analysis:** Examine whether variables follow expected statistical distributions

Core Data Cleaning Techniques

1. Handling Missing Values

Missing Data Assessment

- Quantify percentage of missing values for each variable
- Distinguish between missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)
- Identify temporal patterns (consecutive gaps vs. sporadic missing points)
- Determine if missingness correlates with seasons, events, or specific equipment

Core Data Cleaning Techniques

Imputation Methods

For sporadic missing values (1-2 observations):

- Mean imputation: Replace with the variable's long-term mean (may underestimate variance)
- Seasonal decomposition imputation: Use seasonal mean for the missing period
- Interpolation: Linear interpolation between adjacent observations
- Forward-fill/backward-fill: Propagate last known observation (suitable for short gaps)

For extended missing periods (multiple days/weeks):

- Autoregressive models: Use ARIMA-based imputation leveraging temporal autocorrelation
- Spatial interpolation: Use measurements from nearby stations
- Machine learning approaches: K-nearest neighbors or multiple imputation by chained equations (MICE)
- Deletion: Remove records if gap exceeds acceptable threshold and alternative
- methods are unsuitable

Imputation Validation

Core Data Cleaning Techniques

Outlier Detection and Treatment

- Z-score method: Flag values exceeding 3 standard deviations from mean (assumes normality)
- Interquartile range (IQR): Flag values beyond 1.5 IQR from quartiles (robust to non-normality)
- Modified Z-scores: Use median absolute deviation instead of standard deviation for better robustness
- Winsorization: Cap extreme values at specified percentiles (e.g., 1st and 99thpercentile)

Outlier Type	Cause	Treatment
Isolated spike	Measurement/transmission error	Replace or interpolate
Multiple consecutive	Sensor failure	Impute using appropriate method
Extreme weather event	Genuine meteorological occurrence	Retain with notation; analyze separately
Impossible value	Data entry error	Correct or delete

Advanced Consideration for Climate data

Handling Non-Stationarity and Temporal Dependence:

- Detect stationarity: Apply Augmented Dickey-Fuller (ADF) test or KPSS test
- Detrending: Remove long-term trends using polynomial fitting or seasonal-trend decomposition
- Differencing: Calculate first or second differences to achieve stationarity if needed
- Preserve autocorrelation: Avoid operations that disrupt legitimate temporal structure during cleaning
- Seasonal decomposition: Separate trend, seasonal, and residual components for targeted analysis

Handling Extreme Values and Non-Normal Distributions

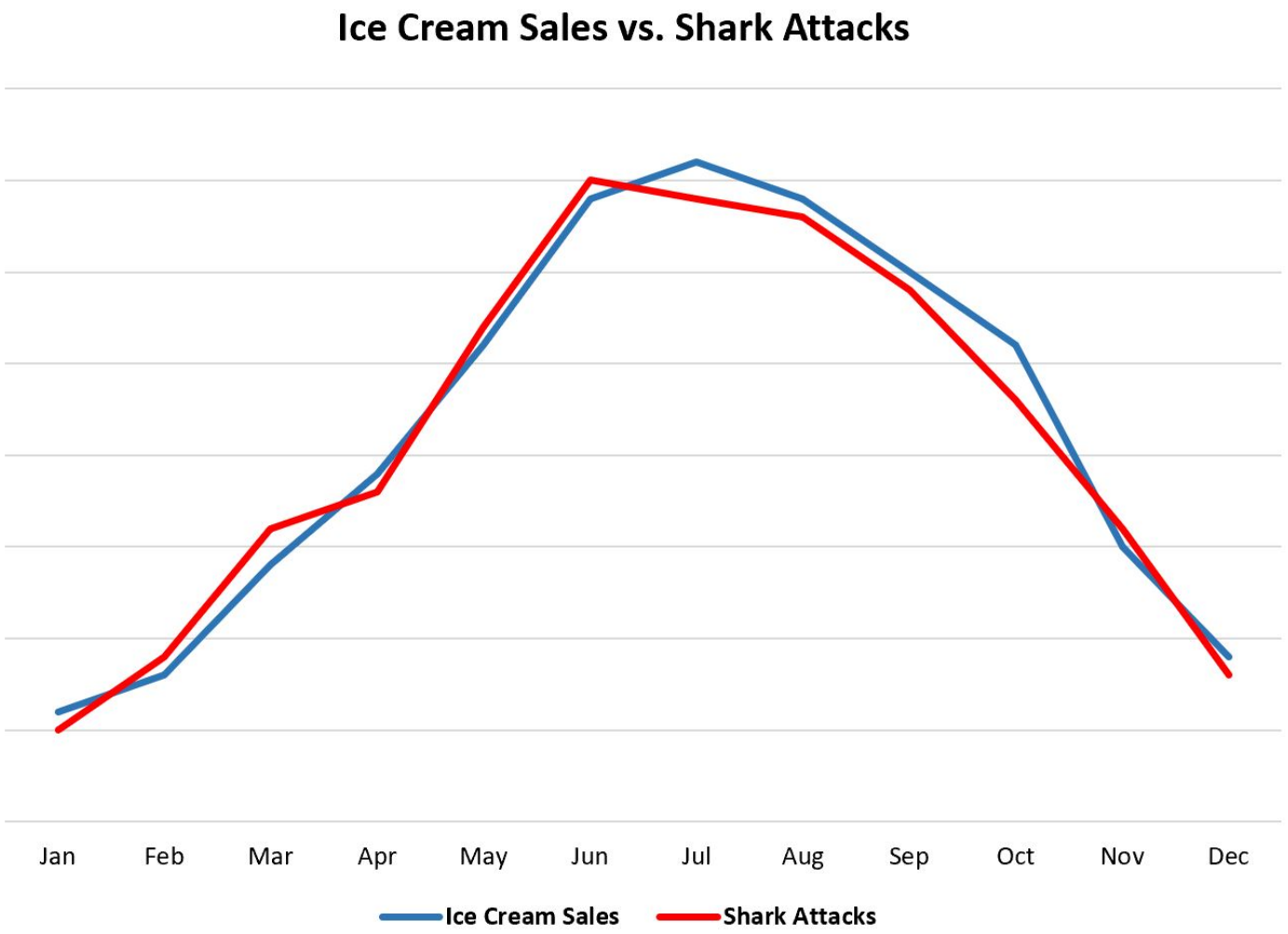
- Distribution analysis: Assess normality using Shapiro-Wilk test, Q-Q plots, or histogram visualization
- Extreme value theory: Apply generalized extreme value (GEV) distribution for tail analysis
- Transformation: Consider log, square-root, or Box-Cox transformations for skewed variables
- Robust statistics: Use median and IQR instead of mean and standard deviation for non-normal data
- Tail analysis: Distinguish between physically plausible extremes and measurement errors

Best Practices

1. Preserve raw data: Always maintain original unmodified datasets for comparison and validation
2. Document procedures: Record all cleaning steps, parameters, and decisions for reproducibility
3. Automate where possible: Use software tools to ensure consistency and efficiency
4. Quality prioritization: Focus efforts on critical variables and time periods affecting analysis objectives
5. Validate results: Compare cleaned data against independent sources to confirm accuracy
6. Domain expertise: Involve meteorologists/climatologists in decision-making for ambiguous cases
7. Balance effort: Avoid over-cleaning; assess cost-benefit ratio before extensive processing
8. Transparency: Publish data quality metadata and cleaning methodology alongside results
9. Version control: Track dataset versions and maintain audit trails of changes
10. Continuous improvement: Monitor cleaning effectiveness and refine procedures based on analysis feedback

Common Pitfalls to Avoid

1. Removing all outliers indiscriminately: Climate extremes are real and important; distinguish errors from genuine events
2. Ignoring seasonal patterns: Climate data has strong seasonality; use season-specific thresholds
3. Over-imputation: Excessive imputation creates artificial data; document and minimize where possible
4. Inconsistent methodology: Apply identical procedures across entire dataset and all stations
5. Data leakage in prediction: Ensure rigorous temporal separation between training and test data
6. Insufficient metadata: Poor documentation makes results irreproducible and errors difficult to trace
7. Ignoring spatial correlations: Neighboring stations have correlated measurements requiring appropriate treatment


Ice Cream Sales vs. Shark Attacks
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Ice Cream Sales
Shark Attacks

Data Sources and References

Description	Link
Gujarat Power Dashboard	https://climatedot.org/dashboard/power/
Emissions Data for Decarbonization	https://rmi.org/how-the-right-emissions-data-can-drive-decarbonization-in-high-emitting-industries/
High Res GHG Emissions Insights	https://www.cityclimateintelligence.com/high-resolution-ghg
Solar cities Uttar Pradesh	https://solarcitiesportal.upneda.org.in/
Women in Renewable Energy (WIRE)	
Climate change indicator	https://climatechangetracker.org/climate-change-progress
Energy Access Explorer	https://www.energyaccessexplorer.org/
World Resource Institute	https://datasets.wri.org/?_gl=1*1d97e5s_gcl_au*NDc4OTQzMzIwLjE3NDM3NDcyMjg
PDF Example (ASICS 2023 data book)	https://www.janaagraha.org/wp-content/uploads/2024/02/ASICS-2023-data-book.pdf

Data Sources and References

Dataset	Description	Official Link
WorldClim	Global high-resolution climate data (current, future, historical) up to 1 km² resolution atlas+1.	**WorldClim Data Portal**
PRISM	Topographically-aware, high-resolution spatial climate data for the United States (grids and station data) climatedataguide.ucar+1.	**PRISM Climate Group**
CRU Time-series	Global gridded time-series (CRU TS) at 0.5° resolution, covering month-by-month variations since 1901 catalogue.ceda+1.	**UEA CRU Data** *(Alt: CEDA Archive)*
ClimateNA	Software and data for generating scale-free point estimates and downscaled gridded data for North America sites.ualberta+1.	**ClimateNA Download (UAlberta)** *(Alt: ClimateNA.ca)*