Characteristics of Environmental Data

This section provides an overview of various data classes and highlights the common characteristics of environmental data.

Types of Environmental Data

Environmental data can be categorized into the following classes based on the methods used for its collection:

1) In-situ Observation Data

In-situ observation data refers to data collected directly at the location of interest using various instruments and techniques. This means that either a person is present at the specific location to make measurements or perform sampling for subsequent analysis, or there is a fully or semi-automated instrument at that location conducting the measurements. The location of interest can vary widely, including the Earth's surface, underground accessed through wells, the surface, or depths of water bodies like lakes, seas, or rivers, or specific heights in the atmosphere.

In-situ observation data can be categorized into high-frequency and low-frequency data.

1-1) High-Frequency In-situ Observations

High-frequency in-situ data are continuously recorded, often at small intervals of seconds, minutes, or hours, to provide a detailed time series of environmental parameters. The data can be stored and retrieved in batches (offline data) or transmitted in real-time (online data).

Examples:

- **Weather Stations**: Continuous recording of temperature, humidity, wind speed, and precipitation.
- **River Gauges**: Continuous measurement of water levels and flow rates.
- **Air Quality Monitors**: Real-time monitoring of pollutants like PM2.5, ozone, and nitrogen dioxide.
- **Seismic Sensors**: Real-time monitoring of seismic activity to detect earthquakes.
- **Flood Warning Systems**: Real-time data on river levels and rainfall to provide early warnings for floods.
- **Ocean Buoys**: Real-time data on sea surface temperatures, wave heights, and salinity.

Measuring platforms in rivers for monitoring water quality.

1-2) Low-Frequency In-situ Observations

1-2-1) Periodic Data

Periodic sampling involves collecting data at regular (relatively larger) intervals to monitor changes over time, such as daily, weekly, monthly, or even seasonally, but not frequently enough to be considered continuous. This data is generally offline.

Examples:

- **Water Quality Sampling**: Monthly sampling of rivers and lakes to measure nutrient levels and contaminants. Samples are collected and then taken to a lab for analysis.
- **Soil Sampling**: Seasonal soil sampling to assess moisture and nutrient content. Samples are collected and analyzed in a lab.
- **Biodiversity Surveys**: Conducting periodic surveys of plant and animal species to monitor changes in biodiversity and ecosystem health.
- **Atmospheric Sampling**: Collecting air samples at specific times to analyze pollutants, greenhouse gases, and other atmospheric components.

1-2-2) Event-Based Sampling

Event-based sampling is triggered by specific events, such as storms, floods, or pollution incidents, to capture the impact of these events on the environment. This data is also generally offline.

Examples:

- **Storm water Sampling**: Collecting water samples during and after a storm to analyze runoff and pollution levels.

- **Post-Wildfire Surveys**: Sampling soil and water to assess the impact of wildfires on the environment.

Crowdsourced Data

Beyond using specialized equipment and trained professionals, in-situ environmental data can also be collected by ordinary citizens, often referred to as citizen scientists. These individuals use pure observation and personal devices such as smartphones to gather data. This method needs to be guided in a well-organized, objective-oriented way and is often facilitated by web or mobile applications.

Such apps play a crucial role in structuring the data collection process, ensuring data quality, and providing platforms for data submission. They often include guidelines, tutorials, and tools to help citizen scientists accurately record their observations. For example, the "CrowdWater" app enables users to contribute data on water levels and stream conditions, while other apps might focus on tracking wildlife sightings, air quality, or weather conditions. These applications harness the power of community participation to enhance environmental monitoring and research.

Such crowdsourced data can be very valuable because it allows for the collection of large quantities of data over extensive geographic areas and diverse environmental conditions. This breadth and depth of data collection can be challenging and costly to achieve through traditional methods. Additionally, crowdsourced data can provide real-time updates and localized insights, which are crucial for monitoring rapidly changing environmental conditions and for early detection of environmental issues.

Despite its value, crowdsourced data collection faces several challenges. Ensuring the accuracy and reliability of data collected by non-experts can be difficult, as variations in observation methods and potential errors in data recording can affect data quality. Another challenge is the need for consistent and sustained participation from volunteers, which can fluctuate over time. Data privacy and security concerns also arise, as participants may be reluctant to share their location or other personal information. Lastly, the integration of crowdsourced data with data from traditional sources requires careful validation and harmonization to ensure compatibility and coherence in the overall dataset.

The CrowdWater app utilizes data crowdsourcing to collect water-related environmental data.

Unique Features of In-situ Observation Data

In situ observation data has several unique features compared to other types of data in environmental data analytics.

On the positive side:

1. In situ data usually has higher accuracy compared to, other types of environmental data. As a result, it is often used to calibrate and validate other kind of environmental data.
2. Collected at specific points, in situ data can provide high spatial resolution data for localized studies and detailed analysis.
3. High-frequency in situ data allows for the monitoring of short-term variations and trends over time, providing detailed temporal resolution.
4. In situ observations can include depth or vertical profiles (e.g., soil layers, water column profiles), offering detailed information about different layers of the environment.
5. In situ data often includes contextual information such as weather conditions, site characteristics, and specific events (e.g., storms, droughts) that may impact the data.

On the negative side:

Collecting in situ data often requires fieldwork for sampling, analysis or maintenance of measurement instruments, which can be labor-intensive and costly.

In situ data is limited to the locations where measurements are taken, potentially leading to gaps in spatial coverage.

Scaling up in situ observations to cover larger areas can be challenging compared to, e.g., remote sensing data.

2) Remote Sensing Data

Remote sensing (RS) data refers to the acquisition of information about the Earth's surface and atmosphere from a distance, typically using platform such as satellites, aircrafts, drones (Unmanned Aerial Vehicles, UAVs), or Balloons.

Satellite-based RS data is generally low-frequency and periodic because satellites follow fixed orbits and collect data when passing over specific areas at regular intervals. This makes them less flexible for real-time or need-based monitoring.

In contrast, data from aircrafts and UAVs (drones) is typically event-based or need-based, as they can be deployed specifically for targeted monitoring tasks or events.

Balloons can fall in between, being used either periodically or for specific events depending on the deployment.

Overall, RS data is rarely high-frequency, as most systems are constrained by their operational design and data collection methods.

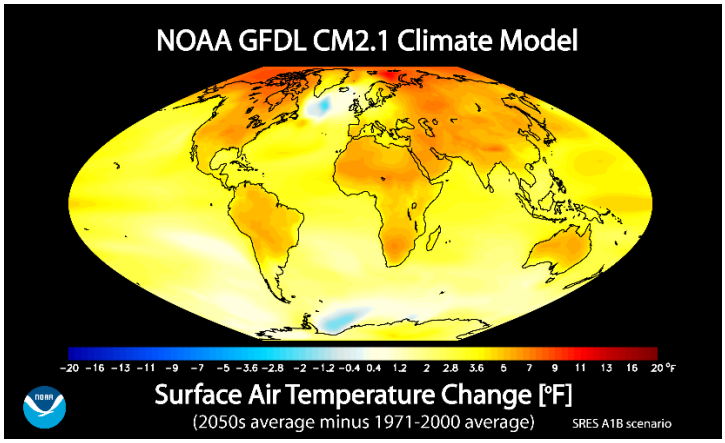
Remote sensing data primarily comes in the form of 2D images, though not necessarily the typical optical images captured by conventional cameras. In certain cases, it can also come in other formats, such as point clouds, especially when using technologies like LiDAR or radar that capture 3D spatial information. It is common for remote sensing (RS) data to be calibrated using in-situ observation data to improve accuracy and ensure that the remotely sensed measurements align with real-world conditions.

A Drone for Soil Moisture Monitoring.

3) Model-generated Data

Model-generated data in Environmental Data Analytics refers to data produced by computational models that simulate environmental processes, such as climate models, hydrological models, or air quality models. This data is an important source of information because it allows us to fill gaps where observational data is unavailable, or simulate scenarios that cannot be directly measured. Model-generated data can be tailored to specific needs, providing either high-frequency data or low-frequency data for long-term trends. Additionally, it can be event- or need-based, meaning the model can be run in response to specific events or requirements, making it highly adaptable for environmental analysis and decision-making.

It is common for computational models that simulate environmental processes to be calibrated using actual observations, particularly in-situ observations, to ensure their accuracy before being used to generate data.



An example of model-generated data showing temperature increase predictions for 2025.

4) Narrative Data

Narrative environmental data refers to mostly textual or qualitative data collected from written or spoken sources that provide context, descriptions, and insights to complement quantitative data from in-situ, remote sensing (RS), or model-generated sources. This type of data is crucial for understanding environmental phenomena in a broader social, cultural, or historical context. It can come from surveys, previous scientific reports, academic papers, or social media, offering valuable perspectives on environmental issues, and human-environment interactions, that are not captured by purely quantitative data.

Collecting narrative environmental data often involves tasks such as text mining, web scraping, or document scraping to extract relevant qualitative information from various sources like websites, reports, and social media.

Concluding remark on Environmental Data Types

The four categories of environmental data—In-situ Observation Data, Remote Sensing Data, Model-generated Data, and Narrative Data—each can come in various formats depending on the collection method and the specific use case. In the following table, we summarize the most common data formats for each category. However, it is important to note that this represents typical instances of data and may not cover every possible format or instance in these categories.

Data Category	Common Data Formats
In-situ Observation	Time series data, scalar values, and sometimes in the form of text or images (such as images from trap cameras).
Remote Sensing	2D images, 3D point clouds.
Model-generated Data	Time series, 2D maps (such as images or gridded data), 3D voxel or point cloud data, scalar values.
Narrative Data	Mostly text (occasionally in the form of scalar values or audio).

Characteristics Common To Environmental Data

Environmental data, regardless of the specific phenomenon it represents, typically shares a set of common characteristics:

1) Spatial Nature:

The spatial component of environmental data—typically represented by latitude, longitude, and sometimes altitude—is crucial because environmental processes are highly dependent on geography. Factors such as topography, land use, vegetation, and proximity to natural features like water bodies or urban centers play a significant role in shaping environmental variables. Accurate modeling and analysis of variations in environmental variables require location-specific data, as it enables a more precise understanding of how environmental phenomena behave across different landscapes.

2) Temporal Nature:

The temporal aspect of environmental data—represented by timestamps or time intervals—is equally important, as many environmental processes evolve over time. Variables such as temperature, precipitation, air quality, and ecosystem dynamics often exhibit daily, seasonal, or long-term trends and cycles. Time-series analysis and temporal modeling are therefore key

techniques in environmental data analytics, enabling the study of how environmental variables change and interact over various time scales.

3) Seasonality and Cyclic Patterns:

Environmental data often exhibits seasonal trends (e.g., rainfall, temperature) that repeat on a yearly or other cyclic basis.

4) Spatial and Temporal Correlation:

Environmental data usually has spatial and temporal dependencies, meaning nearby locations or time points tend to be correlated.

5) Non-linearity:

Many environmental processes, such as climate dynamics or pollution dispersion, exhibit non-linear relationships between variables. This means that small changes in one variable can lead to disproportionate effects in others, complicating the task of predicting or modeling these systems accurately.

6) Scale Dependency:

Environmental data is influenced by the spatial and temporal scales at which it is collected or analyzed. Patterns or relationships that are evident at one scale (e.g., local, regional, or global) may not be visible or may behave differently at another scale. For example, a trend in temperature change observed over a city might not reflect broader regional climate patterns. Choosing the appropriate scale is critical, as environmental phenomena can behave differently across scales, impacting the accuracy and relevance of any analysis or model. When environmental data is reported, it should be accompanied by context such as the scale and resolution at which it was collected to ensure proper interpretation and application in analyses.

7) Multi-modality:

Environmental data is often multi-modal, meaning it is collected in different forms or *modalities*, such as images (satellite or aerial), time series (sensor data), scalar values (e.g., temperature or humidity readings), and text (reports or observations). Integrating and analyzing these different modalities together can provide richer insights but also presents challenges in harmonizing diverse data types.

8) Uncertainty:

Environmental data is often prone to uncertainty.

Data uncertainty refers to the discrepancies between the quantities that describe a system (such as measurements or model predictions) and the actual state of the system, which cannot be fully quantified or precisely known. If these discrepancies can be quantified in a relatively straightforward way, they are considered **errors** rather than uncertainty.

In environmental data, uncertainty arises from various sources, including **measurement uncertainty**, which results from limitations or issues with instruments, **representativity**

uncertainty, occurring when data collected at a specific location or time does not fully capture the broader area or time period, and **interpolation uncertainty**, which arises when observational data is used to estimate values at unmeasured locations, leading to potential inaccuracies. Quantifying these uncertainties and understanding their impact on analytics and machine learning tasks is a critical aspect of environmental data analysis (EDA).

9) Heterogeneity:

Environmental data is often highly heterogeneous, as it may originate from various collection methods, instruments, or even the same method and instrument used at different times by different people. As a result, the scale, resolution, and uncertainty of different subsets of the data can vary significantly. This diversity creates challenges for data integration and analysis, requiring careful handling to ensure consistency and comparability across datasets.

10) Large Volume:

Environmental datasets, especially those derived from remote sensing technologies or large-scale simulations, tend to be massive. High spatial and temporal resolution means that vast amounts of data are generated, which can strain storage and processing capabilities, requiring sophisticated data management and analysis tools.

11) Non-Normality:

Environmental data often does not follow a normal distribution, which is a key assumption in many statistical methods. Applying techniques that rely on normality without verifying the data's distribution can lead to misleading results. Therefore, it is essential to assess the data distribution before using parametric statistics and, if necessary, consider alternative non-parametric methods that do not assume normality.

12) Missing Data:

Environmental data often has missing values due to sensor failures, data collection issues, or other limitations. Handling missing data through techniques like interpolation, imputation, or specialized models is essential for accurate analysis.

Importance of Meta-data in Environmental Data Analytics

Metadata refers to information that describes and provides context about the actual data, such as its source, collection methods, accuracy, resolution, and format.

Metadata is always essential in data analytics and machine learning for ensuring transparency, interoperability, reproducibility, and comparability across datasets. However, it is especially critical in Environmental Data Analytics (EDA) for several reasons.

- First, environmental data requires substantial context for accurate analysis, as it is often not as self-descriptive as data in other fields, such as financial data (though this is a relative comparison).

- Given that environmental data is heterogeneous and scale-dependent, integrating and analyzing it without detailed metadata would be extremely challenging, if not impossible. Metadata provides vital information on how and where the data was collected, including instruments, platforms, and protocols, which ensures compatibility or allows for the appropriate data conversion methods needed for the fusion of heterogeneous data sources.
- Furthermore, environmental data inherently carries uncertainty, and without detailed metadata—such as measurement precision, sensor accuracy, model assumptions, and calibration details—uncertainty quantification and proper error analysis cannot be effectively performed, compromising the reliability and validity of the analysis.

Common Metadata for the 4 Classes of Environmental Data:

1. **In-situ Observation Data**:
    - Location (latitude, longitude, and sometimes altitude).
    - Time of collection.
    - Instrument type and settings.
    - Calibration details.
2. **Remote Sensing Data**:
    - Satellite or sensor type and orbit.
    - Spatial resolution (e.g., pixel size).
    - Temporal resolution (e.g., revisit frequency).
    - Spectral bands used.
    - Processing level (raw, processed, etc.).
    - Georeferencing information (coordinate system, projection).
3. **Model-generated Data**:
    - Model type and version.
    - Input data sources.
    - Model assumptions and parameters.
    - Spatial and temporal resolution.
    - Uncertainty estimates.
    - Calibration and validation methods.
4. **Narrative Data**:
    - Source of the text or audio (e.g., survey, report, social media).
    - Date of collection or publication.
    - Author or contributor details.
    - Method of data extraction (e.g., text mining, web scraping).
    - Language and format.
    - Any categorization or tags related to the content.