Scenario-Based Forecasting and Environmental Risk Mitigation

Learning Outcomes

After completing this topic, students will be able to:

- Explain the concept of scenario forecasting in the analysis of environmental and climate data.
- Describe the process of aggregation of environmental data time series into annual climate indicators and explain its role in the analysis of long-term risks.
- Apply the unsupervised machine learning (K-means) method to group climate years according to their overall characteristics.
- Interpret the resulting clusters as different climate scenarios
- Discuss how the results of scenario analysis can support decision-making

Introduction

Modern environmental monitoring systems increasingly depend on continuous data collection from various sources, including sensor networks, weather stations, satellite observations, and open ecological repositories. These systems generate large volumes of data regularly and form a basis for monitoring the environment's condition over long periods.

Unlike traditional analytical methods, which often rely on static or limited datasets, environmental monitoring is a dynamic process where data is continuously collected and updated. In this setting, machine learning involves not just developing individual models but also maintaining an ongoing analytical approach to handle time-structured data and perform long-term analyses.

Most important environmental variables, such as air pollutant levels, temperature, precipitation, or sea level, are naturally organized as time series. These data display seasonality, short-term fluctuations, and long-term trends, which require methods that analyze how conditions change over time rather than just observing current states.

In this context, machine learning offers methods for monitoring trends and detecting anomalies in environmental data. To be practically useful in operating systems, these methods need to be integrated into a broader ecological data management framework rather than used as isolated

analytical tools, supporting informed and responsible environmental decisions such as early detection of air quality issues, risk assessment of extreme weather events, or monitoring long-term sea level changes.

From Continuous Environmental Monitoring to Climate Scenarios

In long-term environmental monitoring, relying only on individual measurements or short time spans is often insufficient. To detect meaningful changes related to climate risks, data are generally summarized over extended periods, such as annually. This method allows individual years to serve as comprehensive snapshots of environmental conditions rather than just isolated measurements.

Besides average values, extreme events like heat waves or dry spells are crucial in environmental risk analysis. Their significance stems not only from their intensity but also from how often they occur and how long they last. Therefore, the analysis considers not just average temperatures and precipitation but also the frequency, duration, and combinations of harmful conditions.

When these indicators are viewed together, each year can be characterized by a set of features that collectively define its climate profile. In this context, machine learning techniques can then be used to identify similarities among years and group them based on overall traits.

K-Means Clustering Algorithm in the Context of Environmental Data

K-means [1], [2], [3] is a simple, widely used clustering algorithm for identifying similarities in large datasets. Its main idea is to split the data into a fixed number of groups (clusters), where each group contains records that are most similar to each other based on observed features. In this method, each year is represented by a set of numerical features, and the algorithm examines all these features at once, looking for overall similarity between records rather than matching in a single variable.

In practice, the algorithm begins with initial estimates of the group centers. It then repeatedly adjusts them to reduce the distance between each record and its group centroid. This process continues until the distribution of documents across groups stabilizes. The result is groups that represent common patterns in the data.

In environmental analyses, K-means groups different time periods, like individual years, based on their overall climate profile rather than just one variable. This method helps identify stable, transitional, and extreme climate conditions without relying on predefined thresholds or rules.

Let's imagine that each year, we track a few simple indicators like the average temperature, the number of very hot days, and the days without rain. Looking at these indicators individually, each year seems unique, making it hard to describe them clearly.

However, when all indicators are considered together, some years display very similar patterns, as shown in Figure 5.1. For example, some years may have moderate temperatures, rare heat waves, and sufficient precipitation. The second half of the year tends to have higher temperatures, more frequent heat waves, and longer dry periods.

In this case, the K-means algorithm does not specify the number of groups in advance; instead, it automatically detects groups of years that are most similar across all observed indicators. Each resulting group can be interpreted as a typical climate pattern.

One group may represent more stable, moderate years; the second may correspond to significant thermal stress; and the third could be linked to extreme heat and drought. When analyzing a new dataset, each year can be compared to these patterns and classified into the group it most closely resembles based on its characteristics.

The case study outlined below is based on this concept, utilizing long-term meteorological time series to classify actual climate years and analyze them in the context of climate risks.

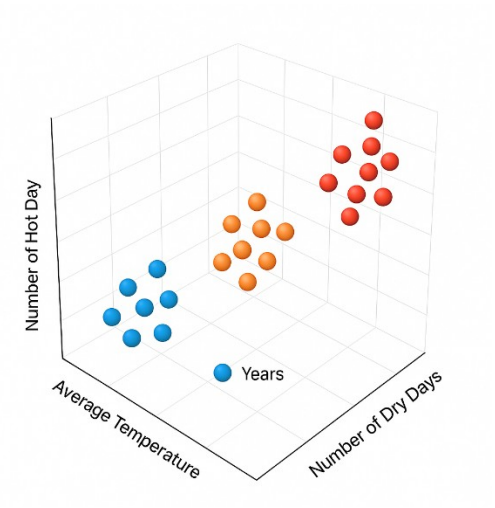

Figure 5.1 Illustrative Example of Clustering Annual Climate Profiles

Case study: Climate risk analysis based on meteorological weather series

Open-Meteo data source

For this case study, the Open-Meteo platform [4] is used as a reference for weather data. Open-Meteo provides free access to historical weather data for any location worldwide through a publicly available API, allowing users to choose a time range and a set of weather variables. Historical data date back to 1940, with a slight delay (up to 1 day) in accessing the most recent records.

While basic data access is free, Open-Meteo restricts the open application model by limiting API calls and advanced features. For professional and business users, extended API plans are available to offer higher limits, more datasets, and more reliable infrastructure.

The data from Open-Meteo is sourced from public resources, including national weather agencies and climate reanalysis. It is provided under the Creative Commons Attribution 4.0 (CC BY 4.0) license, meaning users must attribute the source. This license permits its use in educational and research work while promoting transparency and reproducibility of the analysis.

The download should be performed through a straightforward network interface, specifying the geographic location, time range (Figure 5.2), time resolution (hours or daily), and selected meteorological variables. This method ensures accurate, consistent data collection across different locations and analysis scenarios.

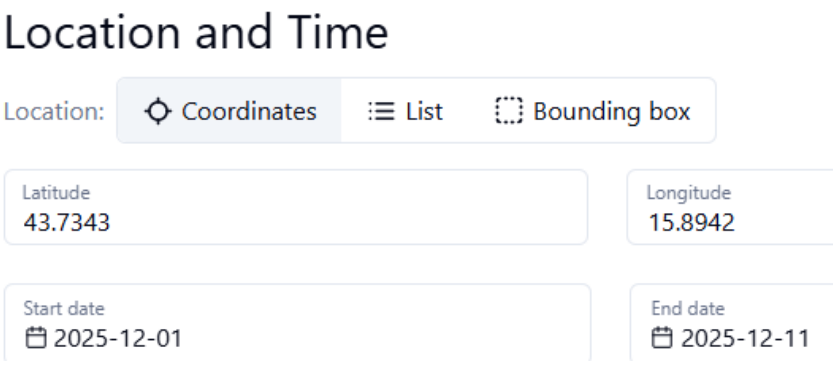

Figure 5.2 Selection of Data Location and Time Series on the Open-Meteo Platform

Example of Clustering Climate Years

For climate risk analysis, it is important to choose only variables that are directly related to long-term changes and environmental stress, and whose meaning is clearly understandable. In this context, the focus will be on variables describing thermal conditions, hydrological balance, and extreme weather events.

The basic indicators of thermal stress include the average daily air temperature at 2 meters (Mean Temperature (2 m) – temperature_2m_mean) and the highest daily air temperature (Maximum Temperature (2 m) – temperature_2m_max). These measurements will help monitor long-term warming trends and identify extreme heat periods. Additionally, the Mean Apparent Temperature (2 m) (apparent_temperature_mean), which accounts for the effects of moisture and wind, will be used to provide a more accurate assessment of the thermal burden on humans and the environment.

To assess hydrological and dry stress, variables will be chosen to describe water availability and loss in the environment. The total daily precipitation (Precipitation Sum – precipitation_sum) will be used to estimate water input into the system, while reference evapotranspiration (Reference Evapotranspiration – et0_fao_evapotranspiration) will represent the potential water loss due to evaporation and transport. The average soil moisture in the 0-100 cm layer (Mean Soil Moisture (0-100 cm) – soil_moisture_0_100cm_mean) will serve as an indicator of actual water availability in the soil and prolonged dry stress.

As supplementary variables that provide a broader meteorological context and help interpret the results, maximum wind speed (Maximum Wind Speed (10 m) – wind_speed_10m_max), mean relative humidity (Mean Relative Humidity (2 m) – relative_humidity_2m_mean), and mean air pressure reduced to sea level (Mean Sea Level Pressure – pressure_msl_mean) will be used. Although these variables are not central to the analysis, they will improve understanding of the conditions under which climate risks develop and change.

Observation Period and Temporal Scope of the Analysis

Historical meteorological data from January 1, 1950, to December 31, 2024, will be used for climate risk analysis. The chosen period provides a sufficiently long time series to identify long-term changes in thermal and hydrological conditions while maintaining relatively high and stable data quality, which is essential for accurate analysis and interpretation.

The data will be downloaded daily, balancing between time resolution and the stability of the time series. Daily resolution helps track seasonal patterns and changes in adverse condition frequency while reducing noise and short-term fluctuations found in higher-frequency data.

Analytical Workflow and Use of Machine Learning Methods

In analyzing thermal and hydrological indicators, the focus will not be on simply noting the presence of seasonal patterns, which are expected and well known (e.g., summer dry periods), but rather on how their characteristics change over longer periods. Specifically, the analysis will examine shifts in the frequency, duration, and intensity of adverse conditions, along with possible changes in their timing within the year.

In the context of hydrological stress, the analysis will focus on periods when soil water availability decreases for several consecutive years, combined with increased evapotranspiration. It will also assess whether these periods occur earlier in the season, last longer, or show greater intensity than in previous years. These changes indicate an increased risk of drought, even though the seasonal pattern of the phenomenon remains unchanged.

Similarly, heat indicators will not only track summer maximums but also monitor changes in the number of hot days, the duration of heat periods, and the frequency of high temperatures outside typical seasonal ranges. This helps differentiate natural seasonal patterns from long-term shifts, with direct implications for climate risks.

Data Import to an Oracle Data Warehouse

The downloaded meteorological data is imported through the embedded Data Load interface in Oracle Autonomous Data Warehouse [5]. The interface allows loading data from local .xlsx files and automatically converting them into relational tables within the database, as shown in Figure 5.3.

During loading, the target table name is specified, and Excel file columns are mapped to their corresponding database data types. Date records are stored as time attributes, while numerical meteorological variables are loaded as numeric values suitable for further analysis, as shown in Figure 5.4.

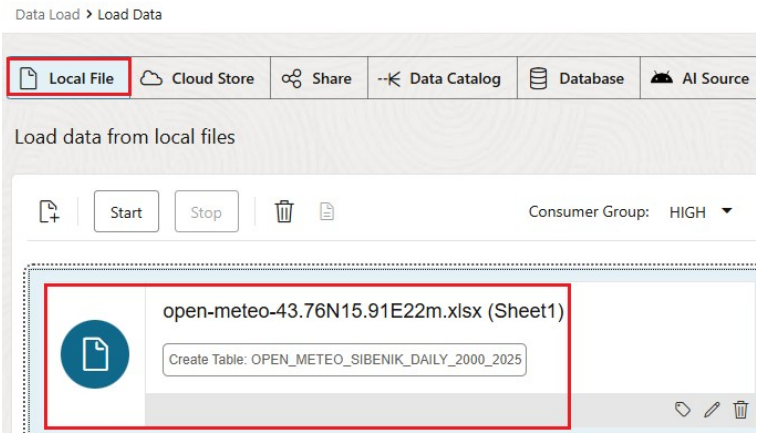

Figure 5.3 Data Load Interface for Importing Meteorological Data in XLSX Format

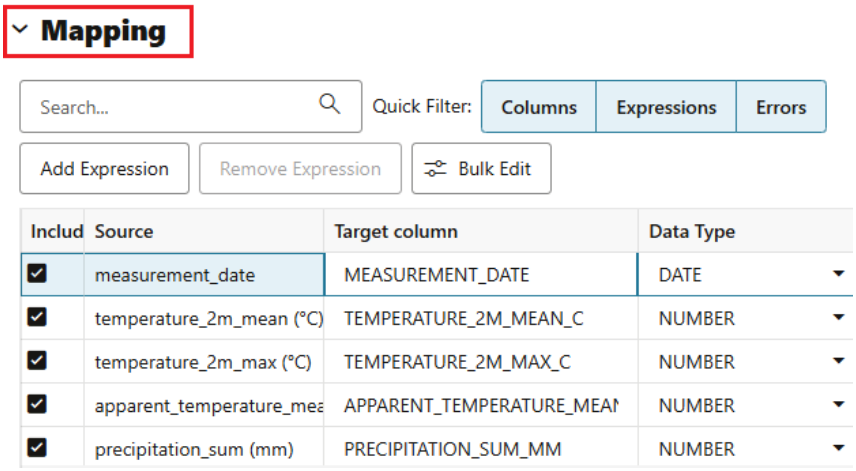

Figure 5.4 Mapping Columns from an Excel File to Database Data Types During Import

This procedure results in a structured table with daily time series of selected meteorological variables for the observed location (Šibenik) and the specified period (from 01/01/1940 to 12/31/2024), as shown in Figure 5.5.

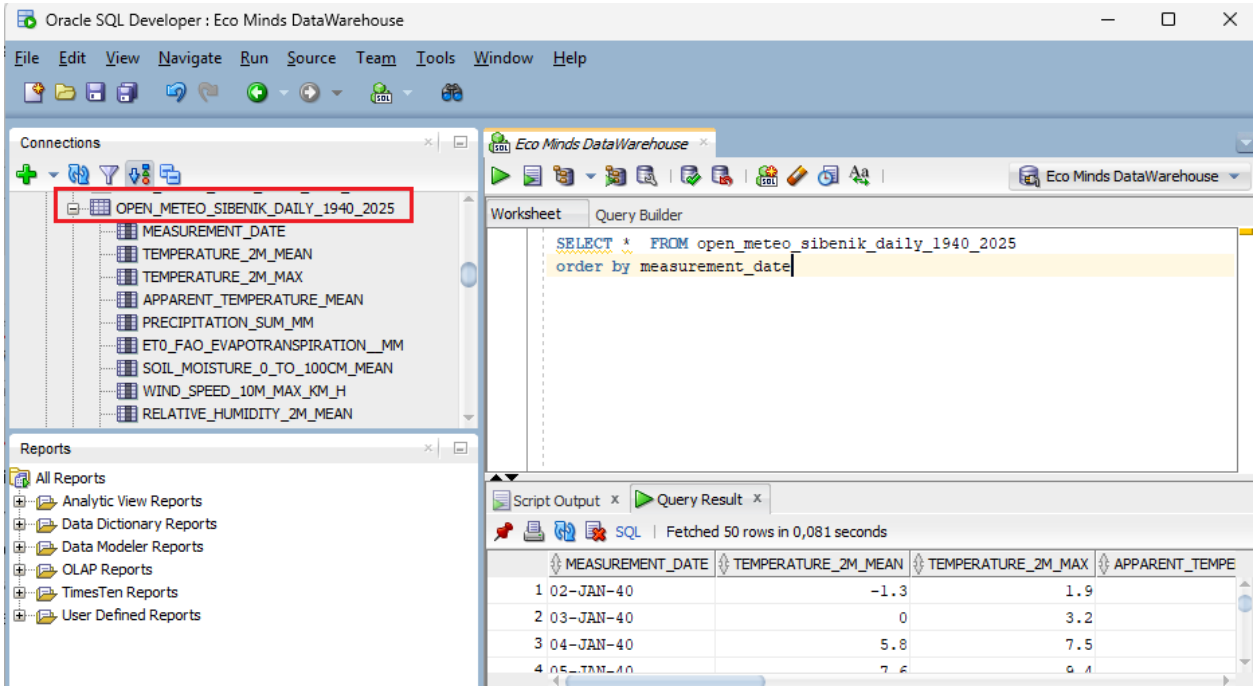

Figure 5.5 Structure of the OPEN_METEO_SIBENIK_DAILY_1940_2025 Table with Daily Meteorological Data

Data Aggregation and Feature Engineering

After importing daily meteorological measurements into the relational database, the next step is to aggregate the data and create annual input features (feature set). Since the goal of the analysis is to identify long-term patterns and climate year types, daily records are converted into yearly summaries that characterize the thermal, hydrological, and extreme conditions for each year as a single multidimensional record.

Aggregation is carried out using a SQL query that employs multiple CTEs (common table expressions). Each part computes a specific set of indicators, and the results are merged into a single annual feature table.

In the block yearly_temps, the annual thermal image is generated using statistical measures of air temperature at 2 meters and skin temperature. Each year counts::

- average and median mean daily temperature (AVG, PERCENTILE_CONT(0.5)), simultaneously covering a "typical" temperature and a robust measure less sensitive to extremes,
- minimum and maximum values (MIN, MAX), where the maximum is taken from temperature_2m_max to capture extreme heat load better,

- analogous sensing temperature measures (apparent_temperature_mean) as an indicator of actual thermal stress, which includes the impact of moisture and wind,
- in addition, the average relative humidity is calculated as a contextual variable that helps to interpret the differences between "dry heat" and periods with increased thermal load due to moisture.

Extreme heat events are evaluated using heatwave criteria. In the heatwave section, days are first identified when temperature_2m_max exceeds 30 °C, and these days are consecutive. Continuity is established by grouping consecutive dates based on the difference between the date and the sequence number. Only episodes lasting at least 5 days (HAVING COUNT(*) >= 5) are classified as heatwaves. Using these episodes, heatwave_stats calculates the annual totals.

- the number of heat waves in the year,
- average, minimum and maximum heat wave duration.

Besides the temperature level itself, the structure of extreme events (such as how often they occur and how long they last) is captured, which is vital for understanding climate risk.

To characterize dry risk as integrated (rather than solely through precipitation), a composite "drought risk day" indicator was introduced in the **yearly_drought** dataset. Such a day is defined as the simultaneous fulfilment of three conditions:

- very low precipitation (precipitation_sum_mm<1),
- high reference evapotranspiration (et0_fao_evapotranspiration__mm>4),
- low soil humidity (soil_moisture_0_to_100cm_mean <0,20).

Additionally, specific days with high evapotranspiration and low soil moisture are identified, allowing for differentiation between years where "atmospheric demand for water" dominates and years where the primary issue is a long-term water deficit in the soil. This way, dry stress is not oversimplified into a single signal. Instead, it is modeled as a function of inputs (precipitation), losses (ET0), and actual water availability (soil moisture).

The **yearly_wind** block adds a component to extreme wind conditions through:

- maximum recorded wind speed in a year,
- number of days with strong wind (threshold 50 km/h).

These indicators are not the main focus of thermal-hydrological risk, but they help in understanding extreme events relevant to operational monitoring and risk management.

The final SELECT combines all calculated annual feature sets into one result, using LEFT JOIN for indicators that do not need to exist every year (e.g., heatwaves), with missing values filled with

zero (COALESCE). The output is an annual dataset of input features, with each year as a record and multiple environmental stress factors: heat, precipitation, drought, and wind. This feature matrix represents the input data space for subsequent use of unsupervised machine learning methods (e.g., K-means).

```
WITH yearly_temps AS (
  SELECT
    EXTRACT(YEAR FROM measurement_date) AS year,

    -- Air temperature statistics
    ROUND(AVG(temperature_2m_mean), 1) AS avg_year_temp,
    ROUND(
      PERCENTILE_CONT(0.5)
      WITHIN GROUP (ORDER BY temperature_2m_mean),
      1
    ) AS median_year_temp,
    MIN(temperature_2m_mean) AS min_year_temp,
    MAX(temperature_2m_max) AS max_year_temp,

    -- Apparent temperature (thermal stress)
    ROUND(AVG(apparent_temperature_mean), 1) AS avg_apparent_temp,
    ROUND(
      PERCENTILE_CONT(0.5)
      WITHIN GROUP (ORDER BY apparent_temperature_mean),
      1
    ) AS median_apparent_temp,
    MIN(apparent_temperature_mean) AS min_apparent_temp,
    MAX(apparent_temperature_mean) AS max_apparent_temp,

    -- Relative humidity
    ROUND(AVG(relative_humidity_2m_mean), 1) AS avg_relative_humidity

  FROM open_meteo_sibenik_daily_1940_2025
  WHERE measurement_date BETWEEN
        DATE '1950-01-01' AND DATE '2024-12-31'
  GROUP BY EXTRACT(YEAR FROM measurement_date)
),
yearly_hydrology AS (
  SELECT
    EXTRACT(YEAR FROM measurement_date) AS year,

    -- Hydrological indicators
    COUNT(CASE WHEN precipitation_sum_mm < 1 THEN 1 END) AS dry_days,
    COUNT(CASE WHEN precipitation_sum_mm >= 1 THEN 1 END) AS rainy_days,
    MAX(precipitation_sum_mm) AS max_daily_precipitation_mm

  FROM open_meteo_sibenik_daily_1940_2025
  WHERE measurement_date BETWEEN
        DATE '1950-01-01' AND DATE '2024-12-31'
  GROUP BY EXTRACT(YEAR FROM measurement_date)
```

```
),
heatwaves AS (
  SELECT
    grp,
    MIN(measurement_date) AS start_date,
    COUNT(*) AS duration_days
  FROM (
    SELECT
      measurement_date,
      measurement_date
        - ROW_NUMBER() OVER (ORDER BY measurement_date) AS grp
    FROM open_meteo_sibenik_daily_1940_2025
    WHERE measurement_date BETWEEN
          DATE '1950-01-01' AND DATE '2024-12-31'
      AND temperature_2m_max > 30
  )
  GROUP BY grp
  HAVING COUNT(*) >= 5
),
heatwave_stats AS (
  SELECT
    EXTRACT(YEAR FROM start_date) AS year,

    -- Heatwave indicators
    COUNT(*) AS heatwave_count,
    ROUND(AVG(duration_days), 1) AS heatwave_avg_duration_days,
    MIN(duration_days) AS heatwave_min_duration_days,
    MAX(duration_days) AS heatwave_max_duration_days

  FROM heatwaves
  GROUP BY EXTRACT(YEAR FROM start_date)
),
yearly_drought AS (
  SELECT
    EXTRACT(YEAR FROM measurement_date) AS year,

    -- Combined drought risk (all conditions met)
    COUNT(
      CASE
        WHEN precipitation_sum_mm < 1
         AND et0_fao_evapotranspiration__mm > 4
         AND soil_moisture_0_to_100cm_mean < 0.20
        THEN 1
      END
    ) AS drought_risk_days,

    -- Individual drought-related conditions
    COUNT(
      CASE
        WHEN et0_fao_evapotranspiration__mm > 4
        THEN 1
      END
    ) AS high_et0_days,
```

```
    COUNT(
      CASE
        WHEN soil_moisture_0_to_100cm_mean < 0.20
        THEN 1
      END
    ) AS low_soil_moisture_days

  FROM open_meteo_sibenik_daily_1940_2025
  WHERE measurement_date BETWEEN
        DATE '1950-01-01' AND DATE '2024-12-31'
  GROUP BY EXTRACT(YEAR FROM measurement_date)
),
yearly_wind AS (
  SELECT
    EXTRACT(YEAR FROM measurement_date) AS year,

    -- Wind extremes
    MAX(wind_speed_10m_max_km_h) AS max_wind_speed_km_h,
    COUNT(
      CASE
        WHEN wind_speed_10m_max_km_h > 50
        THEN 1
      END
    ) AS strong_wind_days_50kmh

  FROM open_meteo_sibenik_daily_1940_2025
  WHERE measurement_date BETWEEN
        DATE '1950-01-01' AND DATE '2024-12-31'
  GROUP BY EXTRACT(YEAR FROM measurement_date)
)
SELECT
  t.year,

  -- Heatwave indicators
  COALESCE(h.heatwave_count, 0) AS heatwave_count,
  COALESCE(h.heatwave_avg_duration_days, 0) AS heatwave_avg_duration_days,
  COALESCE(h.heatwave_min_duration_days, 0) AS heatwave_min_duration_days,
  COALESCE(h.heatwave_max_duration_days, 0) AS heatwave_max_duration_days,

  -- Air temperature
  t.avg_year_temp,
  t.median_year_temp,
  t.min_year_temp,
  t.max_year_temp,

  -- Apparent temperature
  t.avg_apparent_temp,
  t.median_apparent_temp,
  t.min_apparent_temp,
  t.max_apparent_temp,

  -- Relative humidity
```

```
  t.avg_relative_humidity,

  -- Hydrology
  y.dry_days,
  y.rainy_days,
  y.max_daily_precipitation_mm,

  -- Drought stress
  COALESCE(d.drought_risk_days, 0) AS drought_risk_days,
  COALESCE(d.high_et0_days, 0) AS high_et0_days,
  COALESCE(d.low_soil_moisture_days, 0) AS low_soil_moisture_days,

  -- Wind
  w.max_wind_speed_km_h,
  w.strong_wind_days_50kmh

FROM yearly_temps t
LEFT JOIN heatwave_stats h ON t.year = h.year
JOIN yearly_hydrology y ON t.year = y.year
LEFT JOIN yearly_drought d ON t.year = d.year
LEFT JOIN yearly_wind w ON t.year = w.year
ORDER BY t.year;
```

Analysis and Visualization of Aggregated Annual Indicators

Based on the yearly totals calculated by the SQL query described in the previous sub-chapter, a visual analysis of key thermal and hydrological indicators was conducted. This acts as a supplementary analytical tool for understanding long-term patterns and monitoring changes in climate conditions over time.

Figure 5.6 displays the average yearly air temperature in Šibenik from 1950 to 2024, calculated as the annual mean of the daily average air temperature at 2 m (temperature_2m_mean).

The visualization confirms a long-term trend of rising temperatures, with marked year-to-year variability. Notably, since the early 2000s, average temperatures have been higher than previous multi-year averages, and recent years have also experienced periods of significantly elevated temperatures.

This trend not only suggests gradual warming but also a shift in the distribution of annual thermal conditions, which is crucial for subsequent heat-wave analysis and for clustering climate years using machine-learning methods.

Figure 5.7 shows the annual number of heat waves, defined as periods of at least five consecutive days with a maximum daily temperature over 30 °C, along with their average

duration. Early in the series, heat waves are rare and generally short. However, in recent years, their frequency has increased, and the length of individual events has grown significantly.

This pattern shows a shift in how thermal stress impacts us: instead of occasional extreme events, longer periods of high temperatures have become more frequent in recent years, significantly affecting human health, ecosystems, and resource use.

Figure 5.8 shows the annual count of dry and rainy days, with dry days defined as days with less than 1 mm of precipitation. The graph reveals considerable year-to-year variation, with a recent trend toward more dry days and fewer rainy days.

This change in the distribution of precipitation days doesn't necessarily mean a decrease in total annual rainfall. Instead, it shows a higher concentration of rain events and longer dry periods, both of which are important for dry stress development.

Figure 5.9 illustrates the annual number of days meeting the criteria for pronounced dry risk, characterized by very low precipitation, increased reference evapotranspiration, and reduced soil moisture. During most of the observed period, such conditions are not met simultaneously; however, in recent years, there have been notably more dry days.

The emergence of multi-decade periods with combined arid stress indicates a shift in how atmospheric and hydrological factors interact, with drought no longer just a result of precipitation but also increased water loss and decreased soil moisture.

It is important to recognize that adverse features don't necessarily happen each year; their combinations and levels of expression vary annually. The multidimensional nature of yearly climate features emphasizes the need for a method that can analyze multiple indicators simultaneously and detect natural patterns in their occurrence.

Following this, the next step in the analysis is to use unsupervised machine learning methods that treat annual records as complete climate profiles. This approach allows grouping years by their overall similarities, helping to identify different types of climate years and interpret them in terms of climate risks.

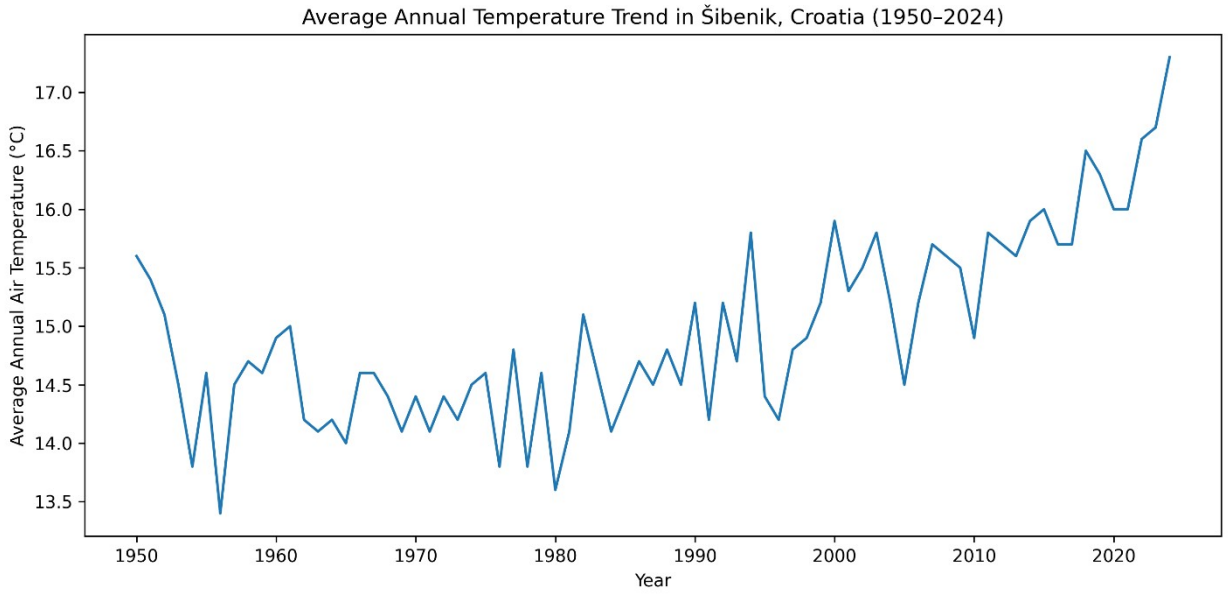

Figure 5.6 Average Annual Air Temperature in Šibenik (1950–2024)

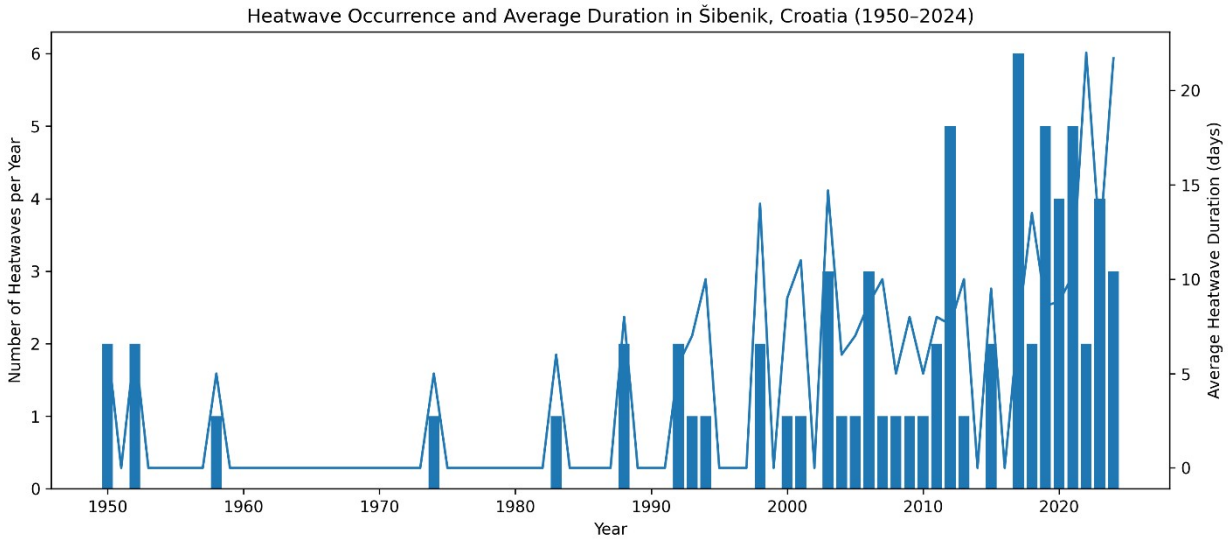

Figure 5.7 Frequency and Average Duration of Heat Waves in Šibenik (1950–2024)

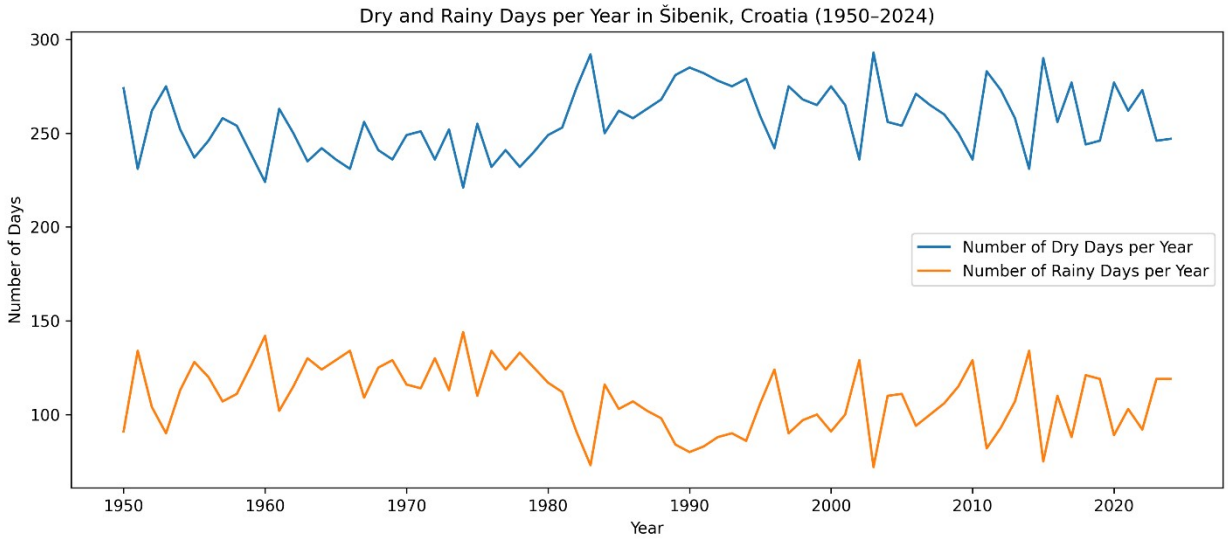

Figure 5.8 Number of Dry and Rainy Days per Year in Šibenik (1950–2024)

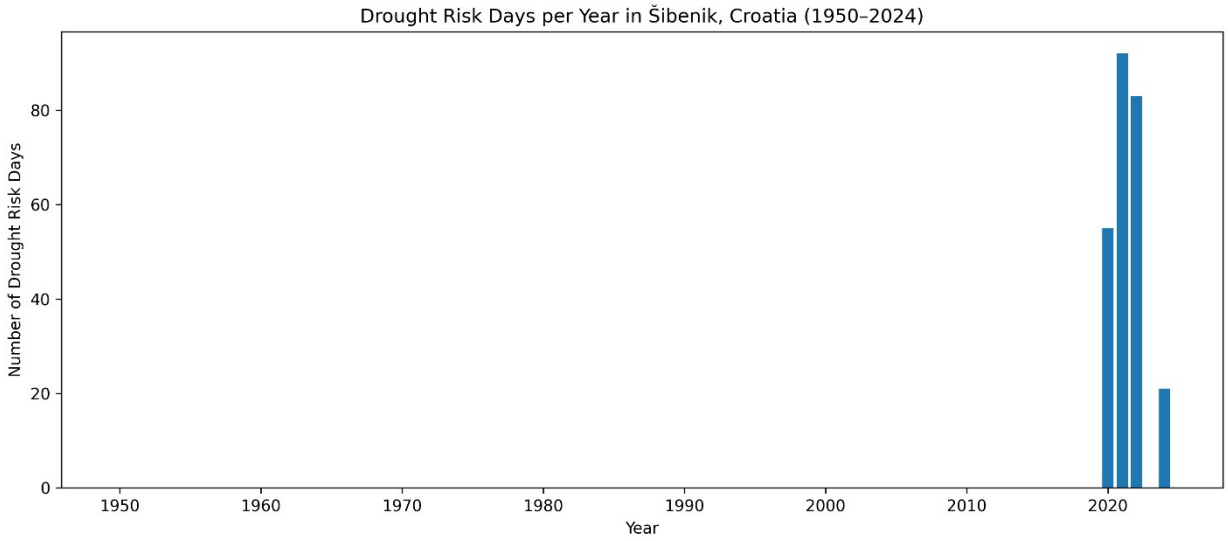

Figure 5.9 Number of Days Meeting the Criterion of Pronounced Drought Risk in Šibenik (1950–2024)

Unsupervised Machine Learning for Grouping Annual Climate Profiles

After aggregating daily meteorological measurements into yearly datasets of characteristics, the K-means clustering method was applied to the data space. The analysis was performed directly within Oracle's data warehouse using the built-in DBMS_DATA_MINING package [6], which allows the integration of data preparation, modeling, and result interpretation within a single relational framework.

A grouping model is defined as a cluster model (CLUSTERING) applied to the table CLIMATE_YEARLY_FEATURES_SIBENIK, which is the direct result of the SQL aggregate query described in the previous sub-chapter. It contains one record for each year with its corresponding set of annual aggregated features, including thermal, hydrological, dry, and wind indicators.

The number of clusters is set to 3 in the model parameters table (CLUS_NUM_CLUSTERS = 3). This choice enables balanced segmentation of annual climate profiles into fewer, more understandable types without excessively splitting the results.

The K-means model has been trained on all available annual features, with input variables automatically normalized and standardized within the Oracle Data Mining framework. This ensures that variables with larger numeric ranges (such as number of days or extreme values) do not dominate the clustering process.

After building the model, each year is assigned to a specific cluster using the built-in CLUSTER_ID function. The results show that the years are grouped into three distinct clusters, each with different time series and climate features, as illustrated in Table 1.

A single cluster (ID = 4), which can be seen as a **stable climate scenario** (marked in green), includes the largest number of years and is characterized by the lowest average temperatures, almost no heat waves or dry stress, and the highest number of rainy days. This cluster consists of years with relatively stable, moderate climates. Figure 5.10 illustrates that this cluster dominates most of the observed period, especially in the second half of the 20th century.

Another cluster (ID = 5), representing a **transitional climate scenario** (highlighted in blue), features higher average temperatures, more frequent heat waves, and more dry days with fewer rainy days. This combination suggests increased thermal and hydrological stress, although it does not reach the extreme values of the dry indicator. Figure 5.10 shows that years in this cluster appear sporadically in earlier periods but become much more common after 2000.

The third cluster (ID = 3), representing the **unfavorable (extreme) climate scenario** marked in orange, although covering fewer years, is clearly characterized by the highest average temperatures, the most heat waves, and an exceptionally high number of days with combined arid stress. As shown in Figure 5.10, this cluster occurs almost exclusively in the most recent observation period, indicating a qualitative shift in climate characteristics rather than a gradual change in individual variables.

Table 1. Results of K-means Clustering of Climate Indicators for Šibenik

Cluster ID	Number of years in the cluster	Average annual temperature (°C)	Average temperature sensing (°C)	Average number of heat waves per year	Average Days at Risk of Drought	Average dry days per year	Average number of rainy days per year	Average maximum wind speed (km/h)
3	**4**	**16,5**	**15**	**3,5**	**62,8**	**264,8**	**100,8**	**44,5**
4	46	14,5	12,7	0,1	0	249,4	115,8	53,7
5	25	15,5	13,9	2,1	0	268,1	97,2	51,8

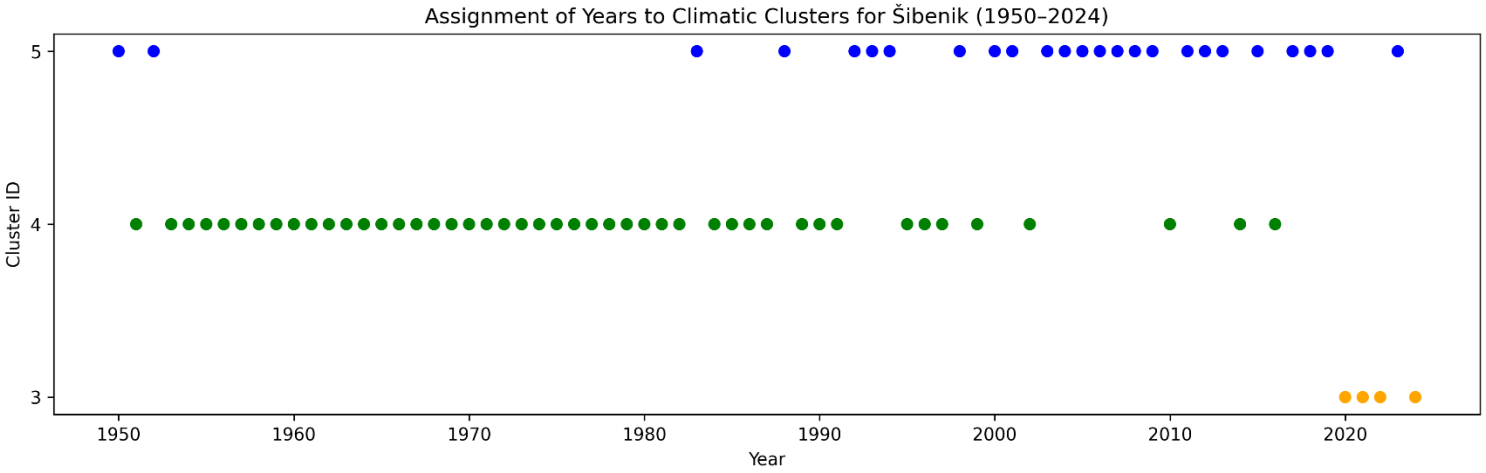

Figure 5.10 Assignment of Years to Climate Clusters in Šibenik (1950–2024)

```
-- ======================================================
-- K-means clustering of yearly climate profiles
-- ------------------------------------------------------
-- This script implements unsupervised machine learning
-- directly inside Oracle Autonomous Data Warehouse.
-- The input table CLIMATE_YEARLY_FEATURES_SIBENIK
-- contains one row per year with aggregated climatic
-- indicators derived from daily meteorological data.
--
-- The goal of the analysis is to group years with
-- similar climatic characteristics using the K-means
-- clustering algorithm.
-- ======================================================

-- ------------------------------------------------------
-- Step 1: Define K-means model settings
-- ------------------------------------------------------
-- Create a settings table used by Oracle Data Mining
-- to configure the clustering algorithm.

CREATE TABLE KMEANS_SETTINGS (
  setting_name  VARCHAR2(30),
  setting_value VARCHAR2(30)
```

```
);

-- Specify the number of clusters (K = 3)
INSERT INTO KMEANS_SETTINGS
VALUES ('CLUS_NUM_CLUSTERS', '3');


-- ------------------------------------------------------
-- Step 2: Create the K-means clustering model
-- ------------------------------------------------------
-- The model is trained directly inside the database
-- using the DBMS_DATA_MINING package.
-- Each row represents one year (case),
-- and all remaining columns are used as input features.

BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'CLIMATE_KMEANS_SIBENIK',
    mining_function     => DBMS_DATA_MINING.CLUSTERING,
    data_table_name     => 'CLIMATE_YEARLY_FEATURES_SIBENIK',
    case_id_column_name => 'YEAR',
    settings_table_name => 'KMEANS_SETTINGS'
  );
END;


-- ------------------------------------------------------
-- Step 3: Verify that the model was successfully created
-- ------------------------------------------------------

SELECT model_name, mining_function
FROM user_mining_models
WHERE model_name = 'CLIMATE_KMEANS_SIBENIK';


-- ------------------------------------------------------
-- Step 4: Assign each year to a cluster
-- ------------------------------------------------------
-- The CLUSTER_ID function returns the cluster
-- to which each yearly climate profile belongs.

SELECT
  year,
  CLUSTER_ID(CLIMATE_KMEANS_SIBENIK USING *) AS cluster_id
FROM CLIMATE_YEARLY_FEATURES_SIBENIK
ORDER BY year;


-- ------------------------------------------------------
-- Step 5: Profile clusters using average feature values
-- ------------------------------------------------------
-- This query summarizes the climatic characteristics
-- of each cluster by calculating average values
```

```
-- of selected indicators.

SELECT
  CLUSTER_ID(CLIMATE_KMEANS_SIBENIK USING *) AS cluster_id,
  COUNT(*) AS years_in_cluster,

  -- Temperature characteristics
  ROUND(AVG(avg_year_temp), 1) AS avg_temp,
  ROUND(AVG(avg_apparent_temp), 1) AS avg_apparent_temp,

  -- Heat stress indicators
  ROUND(AVG(heatwave_count), 1) AS avg_heatwaves,

  -- Drought-related indicators
  ROUND(AVG(drought_risk_days), 1) AS avg_drought_days,
  ROUND(AVG(dry_days), 1) AS avg_dry_days,

  -- Precipitation characteristics
  ROUND(AVG(rainy_days), 1) AS avg_rainy_days,

  -- Wind extremes
  ROUND(AVG(max_wind_speed_km_h), 1) AS avg_max_wind

FROM CLIMATE_YEARLY_FEATURES_SIBENIK
GROUP BY CLUSTER_ID(CLIMATE_KMEANS_SIBENIK USING *)
ORDER BY cluster_id;
```

Confirmation of Identified Climate Patterns Using Recent Data

The results of the case study in the Šibenik area (Croatia) show that climate change is clear not only through gradual temperature rises but also through shifts in the structure of climatic years and more frequent adverse combinations of thermal and hydrological conditions. Visual analysis of yearly indicators reveals an increase in the frequency and duration of heat waves, as well as the occurrence of years with significant combined dry stress.

Cluster analysis results using K-means indicated that, since 2020, a distinct type of climate year has emerged in the observed area, characterized by the most pronounced thermal and dry stress. This cluster is distinguished by high average and maximum temperatures, more frequent and longer heat waves, and extended periods of combined dry stress; over time, it appears almost exclusively in the most recent period of analysis.

Insight into nearly complete meteorological data for 2025, presented in Table 2 (concluding on December 12, 2025), shows that this year's characteristics closely align with the previously identified pattern. Four thermal waves with an average duration of over 13 days have been recorded, along with an extremely high maximum daily temperature of 38.4 °C and more than 100

days experiencing combined dry stress, while the number of rainy days has decreased. This combination of thermal and hydrological indicators matches the profile of the most extreme cluster identified in historical data analysis.

This confirms that the applied K-means analysis is not just a retrospective description of past conditions but also a tool for detecting structural changes in the system's climate behavior. Therefore, the year 2025 is not an isolated extreme but part of a pattern that began in the early 2020s, emphasizing the importance of this approach for early identification of adverse scenarios and environmental risk assessment.

Table 2 Key climate indicators for 2025 (Šibenik)

Indicator	Value
Number of Heatwaves	4
Average thermal wave duration (days)	13,5
Maximum daily temperature (°C)	38,4
Average annual air temperature (°C)	17
Average temperature sensing (°C)	15,3
Dry Days	252
Number of rainy days	94
Number of days with combined dry stress	101
Number of days with increased evapotranspiration	128
Maximum wind speed (km/h)	48,4

Implications for Scenario-Based Environmental Risk Mitigation

In the Šibenik area, the identified climate clusters serve as evidence-based scenarios illustrating different levels of environmental risk. Years with relatively stable climate conditions act as the baseline scenario. However, one cluster is characterized by significant heat and drought stress, representing an unfavorable worst-case scenario. The increasing frequency of such years in recent decades, along with nearly complete data for 2025, indicates that these severe conditions are becoming a regular feature of the local climate rather than a rare event.

Practically, the increased frequency of years with unfavorable climate traits emphasizes the need for better water resource management, especially during long dry periods and the tourist season when water demand rises. The growing frequency and duration of heat waves highlight the importance of preventive actions in public health, urban adaptation, and energy use, particularly in crowded areas and summer months. Additionally, understanding the shift from more

stable to riskier climate types can help develop local early-warning systems and climate change adaptation plans.

Extending the Analysis to Additional Cities

To verify the applicability of the proposed approach, additional analyses were conducted for three cities in different geographic and climate regions: Rome (Italy), Skopje (North Macedonia), and Kyiv (Ukraine).

For each city, the same analytical process as for Šibenik was used, including selecting meteorological variables, aggregating daily measurements to annual data, and applying the K-means algorithm. However, the model was trained separately on location-specific data. This ensures that identified clusters represent local climate patterns and the relationships between indicators within each city.

Cluster analysis results for Rome, Skopje, and Kiev are presented in the related tables, which show the averages of key climate indicators per cluster (Tables 3-5), as well as in figures that illustrate the distribution of climate years per cluster for the period 1950-2024 (Figures 5.11-5.13). Although the methodology is consistent across all cases (as previously for Šibenik), the resulting clusters differ in their characteristics and temporal trends, confirming that climate patterns and risks vary across regional contexts.

Table 3 Results of K-means Clustering of Climate Indicators for Rome

Cluster ID	Number of years in the cluster	Average annual temperature (°C)	Average temperature sensing (°C)	Average number of heat waves per year	Average Days at Risk of Drought	Average dry days per year	Average number of rainy days per year	Average maximum wind speed (km/h)
3	20	16	15,3	3	40,3	268,7	96,6	35,3
4	21	15,4	14,8	2,8	15,1	255,5	109,8	34,4
5	34	14,7	14	1,2	5	250,1	115,1	34,4

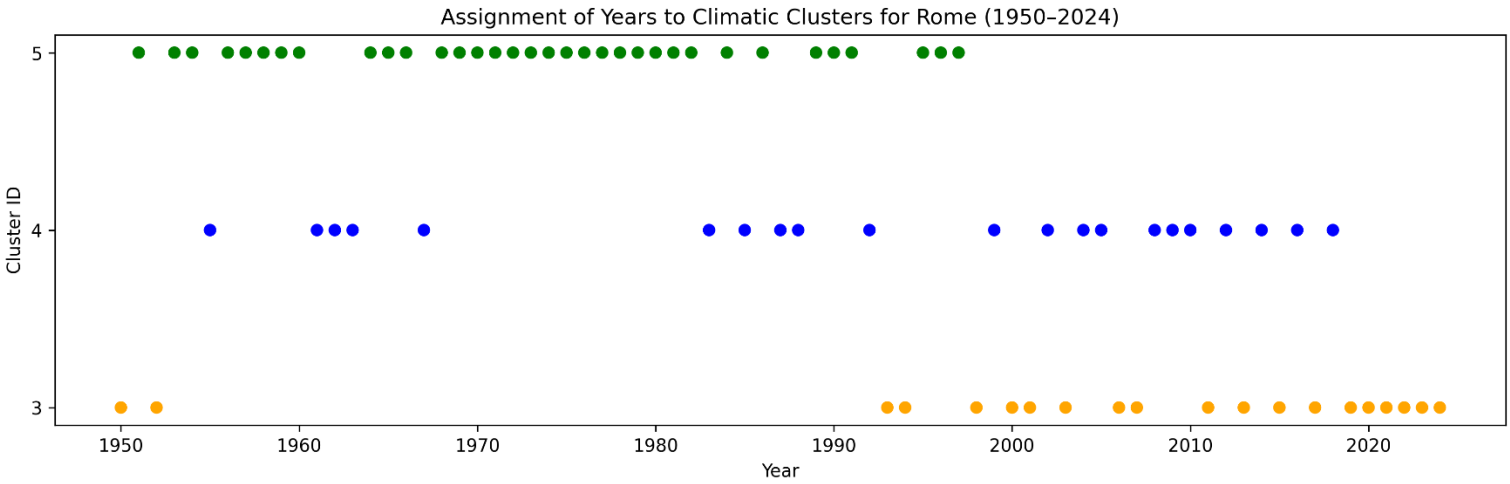

Figure 5.10. Assignment of Years to Climate Clusters in Rome (1950–2024)

Table 4 Results of K-means Clustering of Climate Indicators for Skopje

Cluster ID	Number of years in the cluster	Average annual temperature (°C)	Average temperature sensing (°C)	Average number of heat waves per year	Average Days at Risk of Drought	Average dry days per year	Average number of rainy days per year	Average maximum wind speed (km/h)
2	45	12,6	11,1	2,6	36,6	269,8	95,4	26,7
4	3	13,1	11,4	2,3	66,7	301,3	64	31
5	27	14,1	12,6	4,6	78,2	285,1	80,1	26,2

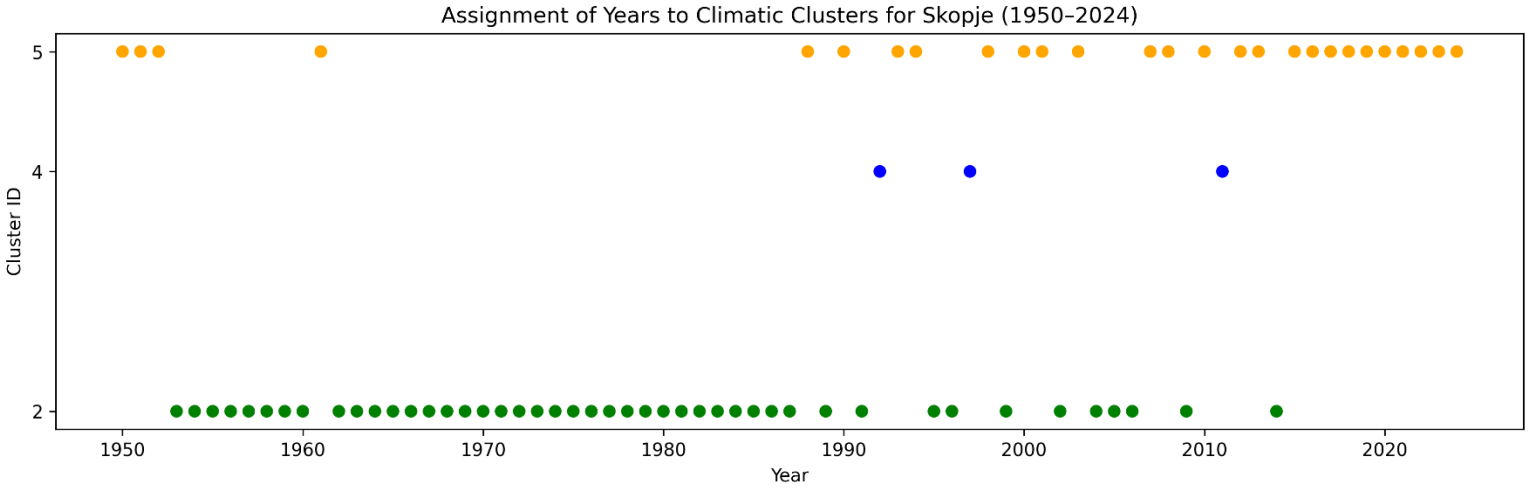

Figure 5.11 Assignment of Years to Climate Clusters in Skopje (1950–2024)

Table 5 Results of K-means Clustering of Climate Indicators for Kyiv

Cluster ID	Number of years in the cluster	Average annual temperature (°C)	Average temperature sensing (°C)	Average number of heat waves per year	Average Days at Risk of Drought	Average dry days per year	Average number of rainy days per year	Average maximum wind speed (km/h)
3	2	7	4,1	0	0	248	117,5	52,5
4	63	7,7	4,8	0	0	243,2	122	39,6
5	10	8,9	6,1	1,2	0	255,6	109,6	37,1

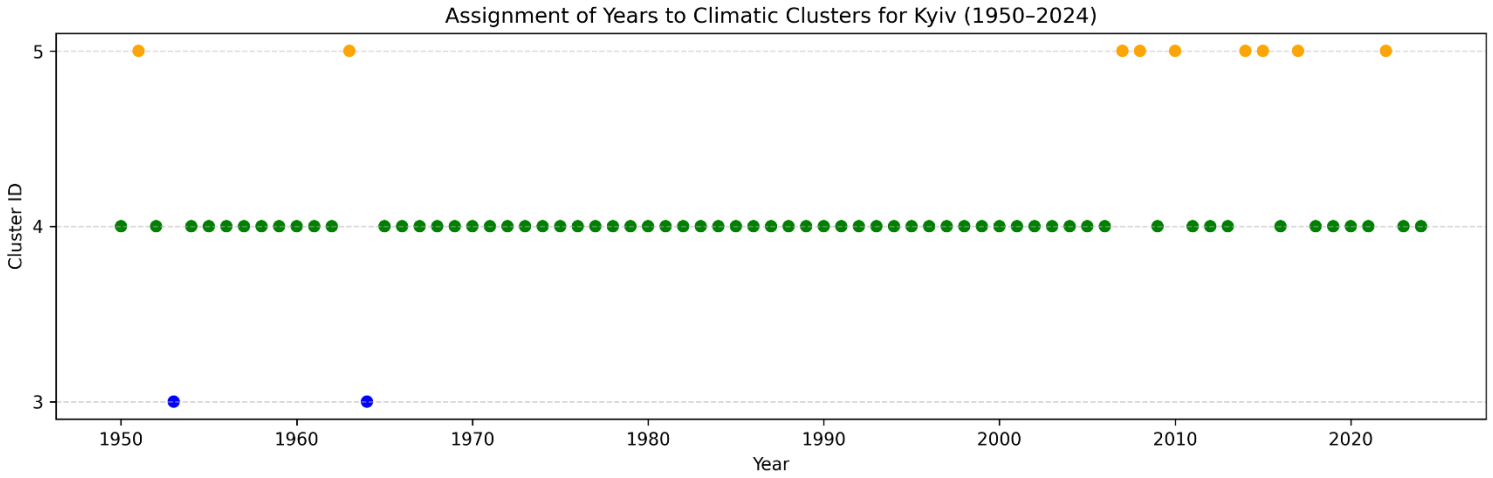

Figure 5.11 Assignment of Years to Climate Clusters in Kyiv (1950–2024)

The analysis of additional cities shows that in Rome, Skopje, and Kyiv, there has been a recent shift toward years with more severe extreme climate features. However, these patterns differ. In all three cities, the frequency of years with higher temperatures and greater thermal stress has risen. Conversely, more stable climate patterns occur more often in the older part of the data series.

It is especially notable that, since the early 2000s and even more after 2010, more extreme clusters have become more common in continental cities like Skopje and Kyiv, indicating the spread of thermal stress beyond typical Mediterranean regions. Although the intensity and combinations of extreme conditions differ from those in coastal cities, the findings suggest a structural shift in climate patterns in these areas as well.

These findings confirm that the identified patterns are not just local anomalies but are part of a broader trend toward climate change, with extreme scenarios gradually emerging across different

climate zones, highlighting the need for locally adapted yet consistently applied analytical methods.

Exercise Tasks

- Download historical daily weather data for the selected location using (e.g., Open-Meteo API)
- Aggregate daily meteorological data into annual climate indicators (e.g., average temperature, number of very hot days, number of dry days).
- Establish annual climate profiles based on selected indicators.
- Apply a K-means algorithm to group climate years according to their characteristics.
- Assign each year to its cluster and display the results in tabular or graphical format.
- Interpret the resulting clusters as different climate scenarios (stable, transient, unfavorable).
- Compare newer data to previously identified clusters and assess which type of climate year it belongs to.

Assessment

What is the primary goal of using the K-means algorithm in the presented case study?

A. Future Temperature Numeric Forecast
B. Grouping of climate years according to their similarities
C. Calculation of seasonal averages of meteorological variables
D. Detection of individual extreme days

What is the purpose of aggregating daily meteorological data to an annual level?

A. Reduce data volume without losing information
B. Easier to spot long-term patterns and climate types
C. Remove seasonality from data
D. Prepare monitored learning data

Which set of indicators best describes the adverse (extreme) climate scenario identified in the case study?

A. High temperatures, high rainy days, low wind
B. Moderate temperatures and rare heat waves
C. High temperatures, frequent heat waves and a large number of days with combined dry stress
D. Large amount of precipitation and short-term extremes

Why is the cluster identified after 2020 in the case of the city of Šibenik considered an indicator of structural change in climate conditions?

A. Because it is based on only one extreme year
B. Because it includes new combinations of thermal and hydrological stress
C. Because it appears almost exclusively in the recent period
D. Because thresholds predefine it

What claim best describes the role of the K-means algorithm in the context of scenario forecasting environmental risks?

A. K-means directly predicts future climate conditions as scenarios
B. K-means identifies typical patterns that can be interpreted as scenarios
C. K-means uses classical meteorological analysis as scenarios
D. K-means is used for short-term forecasts

Additional Sources

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, 2nd ed. New York, NY, USA: Springer, 2021 [Online]. Available: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf, Accessed: Dec. 15, 2025

[2] Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, Machine Learning – A First Course for Engineers and Scientists. Cambridge, U.K.: Cambridge University Press, 2022 [Online]. Available: http://smlbook.org/book/sml-book-draft-latest.pdf, Accessed: Dec. 15, 2025

[3] O. Simeone, A brief introduction to machine learning for engineers, arXiv preprint, arXiv:1709.02840, 2017. [Online]. Available: https://arxiv.org/pdf/1709.02840.pdf, Accessed: Dec. 15, 2025

[4] Historical Weather API Documentation, 2024. Available: https://open-meteo.com/en/docs/historical-weather-api, Accessed: Dec. 15, 2025.

[5] Oracle, Autonomous AI Lakehouse. [Online]. Available: https://www.oracle.com/autonomous-database/autonomous-ai-lakehouse, Accessed: Dec. 15, 2025.

[6] Oracle, DBMS_DATA_MINING Package Reference, Oracle Database 19c Documentation. [Online]. Available: https://docs.oracle.com/en/database/oracle/oracle-database/19/arpls/DBMS_DATA_MINING.html, Accessed: Dec. 15, 2025.