WEATHER CONDITIONS AND CLIMATE CHANGE WITH CLIMATEWINS

MACHINE LEARNING PREDICTIONS

SORIN CATIGHERA

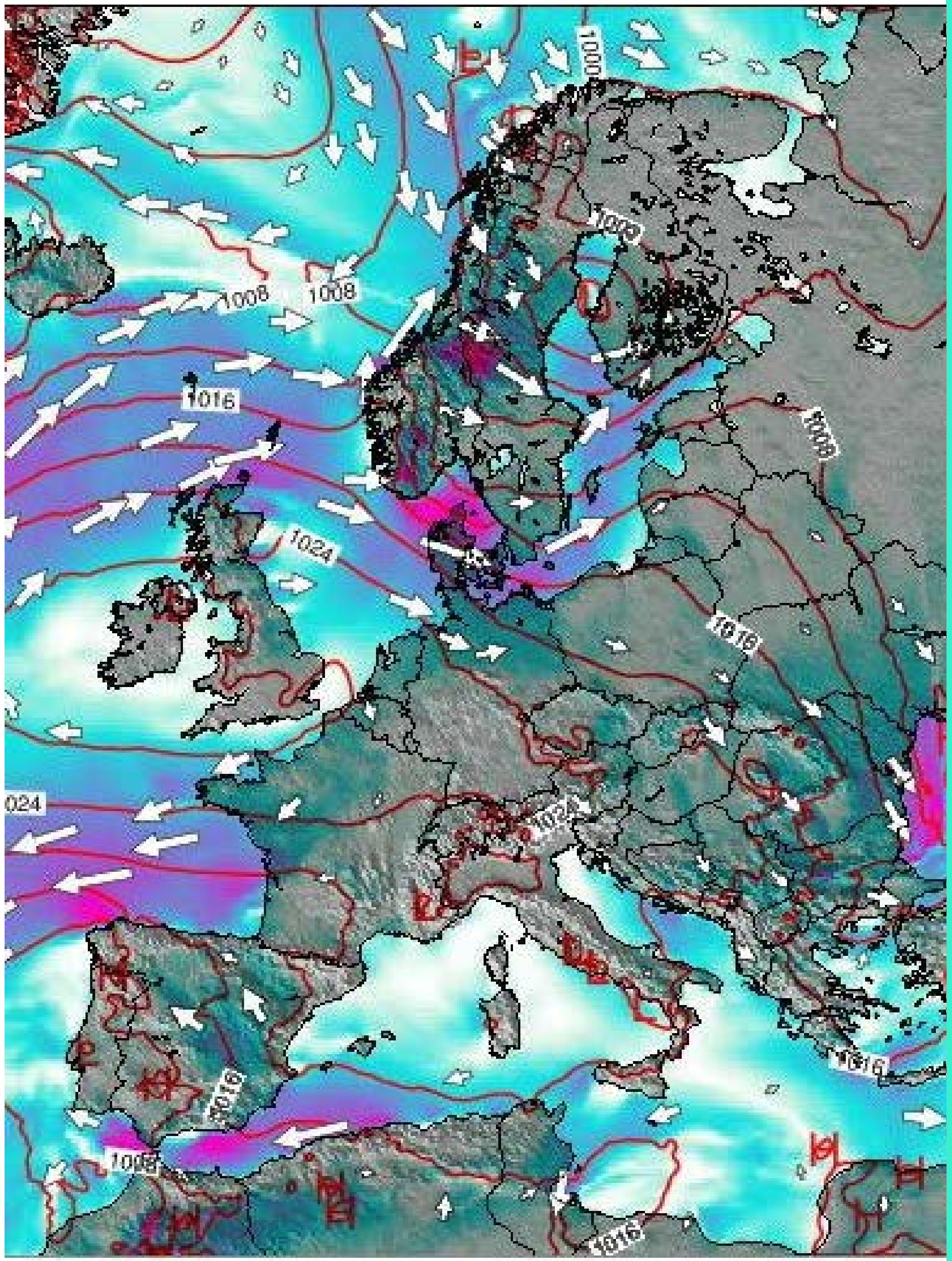
April 12, 2025

INTRODUCTION

As climate change accelerates, global weather patterns are becoming more erratic, leading to a rise in extreme weather events that put countless communities at risk.

To safeguard human safety and well-being, the company is leveraging machine learning to pursue the following objectives:

- Identify weather patterns outside the regional norm in Europe;
- Determine if unusual weather patterns are increasing;
- Generate possibilities for future weather conditions over the next 20 to 50 years based on current trends;
- Determine the safest places for people to live in Europe over the next 25 to 50 years.



MACHINE LEARNING OPTIONS

- **Convolutional Neural Networks (CNNs**) are a specialized type of deep learning model designed to process data with a grid-like topology, such as images. They're particularly well-known for powering breakthroughs in computer vision, but they're also used in audio, video, and even some types of text data.

  Benefits:
  - Automatic Feature Extraction.
  - Excellent for Image and Video Processing.
  - Versatile Applications Speech recognition like time series, document classification, autonomous driving.
  - Highly Scalable - NNs can be stacked into very deep networks (like ResNet, VGG, etc.) for handling complex tasks with high accuracy.

  We applied this model in various configurations to identify different weather stations across Europe. However, prediction accuracy remained low (around 15%), and loss increased with each additional epoch. The model was also used to recognize images representing different weather types (e.g., rain, sunshine, etc.). Prediction accuracy was 71%.

- **Random Forest** is a popular and powerful supervised machine learning algorithm used for both classification and regression tasks. It's an ensemble method, meaning it combines multiple models to produce more accurate and stable predictions.

  Benefits:
  - High Accuracy & Robustness (Handles both linear and non-linear relationships well).
  - Resistant to Overfitting.
  - Handles Missing Data & Outliers.
  - Works with Large Datasets.

  In our analysis, we plotted two different trees using Random Forest to determine:
  - Which weather stations are most influential (Maastricht, Basel, and Debilt).
  - Which indicators from which weather stations are most important in determining whether a day will be pleasant or not (for Maastricht are: precipitation, maximum and mean temperature).

Random Forest was considered in Thought Experiment 3.

MACHINE LEARNING OPTIONS

- **Hierarchical clustering** is a powerful unsupervised machine learning technique used to group similar data points into clusters based on their distance or similarity. It's particularly useful for exploratory data analysis when the number of clusters is not known in advance.

  Benefits:
  - No Need to Predefine the Number of Clusters.
  - Interpretable Visual Output (dendogram).
  - Handles Various Cluster Shapes.
  - Good for Small to Medium Datasets.
  - Distance-Based Clustering.

  This model was used in our analysis to find actionable weather categorizations which help detect whether unusual weather patterns are occurring.

- **Generative Adversarial Networks (GAN)** is an innovative class of deep learning models that fall under unsupervised learning. A GAN is composed of two neural networks that play a game-theoretic scenario: a generator and a discriminator which are trained simultaneously. If trained well, the generator becomes very good at producing data that looks indistinguishable from the real thing.

  Benefits:
  - High-Quality Data Generation.
  - No Need for Labeled Data.
  - Creative Applications.

  Potential use of GANs in weather prediction:
  - *Precipitation Forecasting*: GANs can analyze regional sky imagery to predict the likelihood of precipitation.
  - *Weather Anomaly Detection*: By learning normal weather data patterns, GANs can identify anomalies potentially signaling extreme events, enabling timely preventative measures by authorities.
  - *Climate Scenario Simulation*: GANs can generate synthetic weather data consistent with various climate change models, facilitating research into the potential impacts of climate change on weather patterns.

  GAN was considered in Thought Experiment 2.

MACHINE LEARNING OPTIONS

- **Isolation Forest** is an efficient and powerful unsupervised machine learning technique used primarily for anomaly detection. It's especially well-suited for detecting rare or unusual data points in large datasets.

  Benefits:
  - Can handle large, high-dimensional datasets with speed and low memory usage.
  - Suitable for real-time or near-real-time anomaly detection.
  - No need for labeling.
  - Designed to find rare patterns.
  - Minimal tuning required.

  Thought Experiment 1 considers this ML technique.

- **Autoencoders** are a type of unsupervised neural network primarily used for dimensionality reduction, feature learning, and anomaly detection. Their main function is to learn a compressed representation of input data and then reconstruct it as closely as possible. They're especially useful when learning patterns in data without labeled outputs.

  Benefits:
  - Are great at spotting anomalies.
  - Common in fraud detection, cybersecurity, manufacturing, and climate anomaly analysis.
  - Captures more complex relationships.
  - Autoencoders can learn compact, meaningful representations.

  Thought Experiment 1 considers this ML technique.

MACHINE LEARNING OPTIONS

- **A Variational Autoencoder** is a type of neural network that, like standard autoencoders, compresses input data into a lower-dimensional space and then reconstructs it. However, VAEs add a probabilistic twist. This makes VAEs capable of generating new data samples by sampling from the learned latent space, making them generative models.

  Benefits:
  - Unlike standard autoencoders, VAEs can generate new, realistic data by sampling from the latent space.
  - Used in tasks like image synthesis, synthetic data creation, and even music or text generation.
  - Because VAEs are based on Bayesian principles, they can model uncertainty—useful in fields like medical diagnosis or scientific simulation.
  - VAEs can learn meaningful, compressed representations of complex input data—great for unsupervised learning.
  - Useful in filling missing values or detecting unusual data that doesn't fit the learned distribution.

  VAE was considered in Thought Experiment 2.

- **Extreme Gradient Boosting (XGBoost)** is one of the most powerful and widely used machine learning techniques—especially popular in structured data problems like classification, regression, and ranking. XGBoost is a gradient boosting framework that builds an ensemble of decision trees, where each new tree corrects the errors of the previous ones.

  Benefits:
  - High Accuracy
  - Built-in Regularization
  - Handles Missing Data
  - Fast and Scalable
  - Feature Importance
  - Supports Custom Loss Functions
  - Cross-validation & Early Stopping

  XGBoost was considered in Thought Experiment 3.

THREE THOUGHT EXPERIMENTS

1. Identify weather patterns outside the regional norm in Europe

ClimateWins employs unsupervised machine learning techniques—Isolation Forest and Autoencoders—to identify anomalies by analyzing real-time weather data (such as temperature and precipitation) against 25 years of historical climate patterns.

2. Generating Possibilities for Future Weather Conditions Over 20-50 Years

ClimateWins uses Generative Adversarial Networks (GANs) and Variable Autoencoders (VAEs) to generate weather scenarios for Europe from 2045-2075:

- Hot & Dry (Paris at 40°C, 30% less rain, 40% odds);
- Wet & Wild (Paris with 25% more rain, Oslo's blizzards double, 35%);
- Mixed Chaos (erratic shifts, 25%).

3. Determining the Safest Places to Live in Europe Over the Next 25-50 Years

To identify Europe's safest places to live in the next 25-50 years, ClimateWins leverages Random Forest and Extreme Gradient Boosting (XGBoost) models, balancing climate risks (floods, heatwaves) and socio-economic factors (healthcare, GDP).

THOUGHT EXPERIMENT 1

- **Premise**

  Let’s assume that ClimateWins uses a system called the "European Weather Watch" (EWW), which is a network of advanced weather drones—AI-driven units hovering over cities, forests, and coastlines from Athens to Oslo. These drones collect continuous data: temperature, precipitation, humidity, wind speed, and atmospheric pressure, beaming it to a cloud-based hub every 10 minutes. ClimateWorks, who wants to identify weather patterns outside the regional norm, uses machine learning to sift through this data. The aim is to pinpoint anomalies—like a sudden 22°C day in Berlin in February or an unprecedented dry spell in Scotland—before they catch communities off guard.

- **Setup**
  - *Data*: The EWW holds 25 years of historical data from sources like Copernicus Climate Data Store, tailored to Europe’s varied climates—continental in Germany, oceanic in the UK, Mediterranean in Italy. This establishes the “norm” for each region: monthly averages, seasonal shifts, and typical daily fluctuations.
  - *Anomaly Criteria*: A pattern is “outside the norm” if it exceeds two standard deviations from the regional mean over a day or week—say, a temperature 10°C above Berlin’s February average or rainfall 40 mm below Scotland’s norm.
  - *ML Tool*: ClimateWorks will use Isolation Forest, a machine learning algorithm that isolates anomalies by randomly partitioning data. The benefit of using this algorithm is that it flags outliers fast. Supplement with Autoencoders because of the Isolation Forest’s focus on sharp anomalies—slow shifts (e.g., humidity drops).

- **Scenario**

  On one day in the month of February, the EWW drones hum over Europe, and the Isolation Forest processes the latest 48-hour data:
  - Berlin, Germany: 22°C, clear skies, light wind—unseasonably warm.
  - Edinburgh, Scotland: 5°C, 1 mm rain—strikingly dry for its wet reputation.
  - Rome, Italy: 12°C, 10 mm rain—standard winter day.

  The Isolation Forest outputs anomaly scores:
  - Berlin: High score (22°C vs. 2°C February norm).
  - Edinburgh: Elevated score (1 mm vs. 15 mm daily norm).
  - Rome: Low score (matches norm).

THOUGHT EXPERIMENT 1

- **Thought Experiment Steps**

1. *Training:* Will train the Isolation Forest on Berlin's historical data—winters averaging 0-4°C, occasional snow. It learns to "expect" cold; 22°C stands out like a flare, isolated in fewer partitions than typical points.
2. *Instant Alert:* The ClimateWins notification system communicates: "Berlin Anomaly—Temperature +20°C Above Norm." Edinburgh flags too: "Potential Dry Spell—Rainfall -14 mm Below Norm." Rome stays quiet.
3. *Validation:* A data analyst checks—Berlin's heat real, not a glitch? Edinburgh's dryness holds—satellites show no clouds. Both are anomalies, not errors.
4. *What If Scenarios:* - If 30 drones over Germany hit 20°C+, it's a regional heatwave—ClimateWorks warns of energy spikes.

   - If the threshold is set too high (3 sigma), then Edinburgh's drought slips—water reserves dip unnoticed. If it's set too low (1 sigma), every drizzle triggers noise.

- **Questions to Consider**

  - *Precision*: Is 22°C in Berlin a fluke or a signal? How fine-tuned should the norm be—monthly or weekly?
  - *Coverage*: What if drones miss rural Spain—does patchy data warp the norm?
  - *Action*: Berlin's heat flagged—then what? Alert power grids or farmers? Edinburgh's dryness—ration water or wait?
  - *Edge Cases*: Can Isolation Forest catch subtle anomalies (e.g., a slow humidity drop) or just sharp ones like Berlin's?

Outcome

The EWW catches Berlin's heatwave—ClimateWins briefs authorities, averting blackouts. Edinburgh's dry spell bumps the region into "monitor" mode as data trends begin to stack. Isolation Forest does its job well, but with fine-tuning and pairing it with Autoencoders helps catch gradual shifts and compensate for drone data gaps. Long-term forecasting still needs refinement, but when it comes to flagging outliers, it's a reliable sentinel worth watching.

Next Steps

Refine the Machine Learning models to sharpen accuracy.
Validate and stress-test outputs (what if a drone misreads 22°C as 12°C).
Expand data coverage and features (cover rural gaps, obtain more variables, like soil moisture).
Operationalize Alerts (Edinburgh's dryness needs a playbook to avoid water shortages).

THOUGHT EXPERIMENT 2

- **Premise**

  Imagine that ClimateWins possesses an advanced AI system specifically designed to forecast Europe's future weather patterns. This system is a network of supercomputers running machine learning models, fed by decades of climate data and real-time feeds from satellites, ocean buoys, and urban sensors stretching from Reykjavik to Valletta. ClimateWins goal is to generate plausible weather scenarios 20 to 50 years ahead, based on current trends. Will Paris face annual monsoons? Could Oslo see snow-free winters? ClimateWins aims to map these possibilities.

- **Setup**
  - *Data*: ClimateWins pulls from 50 years of historical data via Copernicus and IPCC AR6 projections, plus live inputs: CO2 levels, sea temperatures, deforestation rates. It's tuned to Europe's patchwork climates: arctic in Scandinavia, temperate in France, arid in southern Spain.
  - *Scenario Goal*: Generate multiple "weather futures"—not one prediction, but a range of possibilities (wetter, hotter, stormier) over 20-50 years, each with probabilities, for example, Paris with 20% more rain by 2070, or Oslo 5°C warmer by 2070.
  - *ML Tool*: ClimateWorks will use Generative Adversarial Networks (GANs)—a "generator" crafts weather scenarios, a "discriminator" checks them against physics and trends. It's like an artist sketching futures, refined by a critic. Augment with Variational Autoencoders (VAEs) because of the GANs risk overfitting (Paris's 40°C too sharp). VAEs add uncertainty bands (e.g., 38-42°C).

- **Scenario**

  It's June 1, 2025. ClimateWins AI system boots up, tasked with simulating Europe's weather from 2045-2075. It churns through 2025's data nd comes up with three scenarios:
    - *Scenario A (Hot & Dry):* Paris hits 40°C summers by 2070, 30% less rain—Med-like shift. Oslo's winters average 2°C, snow rare by 2070.
    - *Scenario B (Wet & Wild):* Paris sees 25% more rain, flooding every five years by 2075. Oslo's winters stay cold but wetter—blizzards double.
    - *Scenario C (Mixed Chaos):* Paris oscillates—dry decades, then wet spikes. Oslo warms to 0°C winters, storms up 50% by 2075.

  Each comes with a likelihood: A (40%), B (35%), C (25%), based on how well they fit trends and physics.

THOUGHT EXPERIMENT 2

- **Thought Experiment Steps**
  1. *<u>Training</u>:* The GAN's generator learns from 1975-2025 Copernicus data—Paris's 15°C June norm creeps to 18°C, Oslo's -5°C January to -2°C. It mixes this with IPCC's warming curves, predicting up 2045-2075 patterns. The discriminator rejects nonsense (e.g., 50°C Oslo winters).
  2. *<u>Scenario Spin</u>:* The AI system outputs Scenario A—Paris's rainfall drops as heat spikes. The discriminator greenlights it: aligns with drying trends in southern Europe. Scenario B's floods pass too—Atlantic shifts support it.
  3. *<u>Validation:</u>* A climate scientist vets Scenario C—mixed chaos fits erratic jet stream models but stretches data. The GAN tweaks it, boosting storm odds over Oslo.
  4. *<u>What If Scenarios</u>:* - CO_2 hits 500 ppm by 2050—Scenario A jumps to 60% odds, Paris a desert by 2075.

     - GAN overfits 2025's heatwave—predicts all-dry futures, ignoring wetter possibilities.
- **Questions to Consider**
  - *Plausibility*: Does Scenario A's 40°C Paris hold—too extreme, or climate's new norm? How many scenarios are useful—three or thirty?
  - *Uncertainty*: 50 years out—can GANs capture tipping points (e.g., ice cap melt) or just extrapolate?
  - *Use*: Paris planners see Scenario B—build floodwalls? Oslo preps for Scenario C—stormproof homes?
  - *Limits:* What if 2025's data misses a trend (e.g., methane spikes)—does the model hallucinate past it?

Outcome

The ClimateWins AI system delivers: Scenario A warns southern Europe of drought, B flags northern flood risks, C preps for chaos. Paris bets on flood defenses by 2055; Oslo stockpiles storm gear by 2075. The GAN is very good at painting futures—40% odds aren't certain, but they're actionable. This can be improved by blending in Variational Autoencoders (VAEs) for uncertainty bands, and feeding fresher data. It's not a lock for 2075's weather, but for "possibilities," it's a window worth opening.

Next Steps

Enhance the GAN Model for Robustness.
Validate Scenarios Against Climate Models.
Expand Data Inputs for Richer Scenarios.
Link to Broader Company Objectives (spotting anomalies, tracking trends, finding safe places).

THOUGHT EXPERIMENT 3

- **Premise**

  ClimateWins, in order to rank the safest places to live from 2050 to 2075 (25-50 years out), developed a "Haven Index"—an AI-driven platform. Safety means resilience against climate threats (floods, heatwaves, storms) and strong socio-economic buffers (infrastructure, healthcare, economy). The Index isn't a map of utopias; it's a data engine crunching weather forecasts, sea-level models, and demographic trends to score cities and regions. Will Copenhagen outshine Lisbon? Could rural Latvia trump urban Milan? The Haven Index aims to guide families, planners, and governments.

- **Setup**
  - *Data*: The Index pulls 50 years of climate data (1975-2025) from Copernicus and IPCC AR7, plus live 2025 feeds—temperature trends, rainfall shifts, population density. Socio-economic stats come from Eurostat: hospital beds, GDP per capita, flood defenses.
  - *Safety Metric:* A "safe" place scores low on climate risks (e.g., <10% flood chance by 2075) and high on livability (e.g., >80% access to clean water, stable governance). Scores range 0-100—higher is safer.
  - *ML Tool*: ClimateWins deploys Random Forest, an ensemble model that weighs dozens of factors—heatwave frequency, infrastructure quality to rank locations. Enhance with XGBoost which will lift score accuracy for edge cases, and factor in wildfires.

- **Scenario**

  It's October 1, 2025. The Haven Index comes online, tasked with ranking Europe's safest spots for 2050-2075. It processes 2025's reality—global warming at 1.1°C, sea levels up 12 cm (from 1975), migration reshaping cities—and outputs top contenders:

  - *Copenhagen, Denmark:* Score 92—low flood risk (dykes upgraded), stable economy, 95% renewable energy.
  - *Inland Latvia (e.g., Cēsis):* Score 87—minimal heatwaves, cheap land, growing healthcare.
  - *Lisbon, Portugal:* Score 68—rising droughts, strained water, but robust ports.
  - *Milan, Italy:* Score 55—heatwaves hit 40°C, urban density taxes grids.

  Copenhagen takes the lead, with Latvia emerging as an unexpected contender. Lisbon falls behind, while Milan's urban vulnerabilities drag down its performance.

THOUGHT EXPERIMENT 3

- **Thought Experiment Steps**

1. *<u>Training the Index</u>:* The Random Forest learns from 1975-2025. It weighs 30 features—flood probability (10%), hospital beds (15%), GDP (10%)—to predict 2050-2075 safety.
2. *<u>Ranking Run</u>:* Copenhagen's low climate risk (5% flood odds, per IPCC) and high livability (99% water access) score it 92. Latvia's rural buffer (2°C cooler than coasts) hits 87—Milan's 40°C summers drag it to 55.
3. *<u>Validation:</u>* A planner questions Milan's low score—heatwaves real, but new solar grids help. The model re-runs, factoring 2025's grid upgrades—Milan climbs to 62, still behind Latvia..
4. *<u>What If Scenarios</u>:* - Warming hits 3°C by 2040—Copenhagen's flood defenses face pressure (score drops to 85), Latvia holds (90).
   - Lisbon invests in desalination by 2030—jumps to 75, overtakes Milan.
   - Model misses Latvia's soil erosion—rural charm fades if farms fail.

- **Questions to Consider**

  - *Plausibility*: Is Copenhagen's 92 score overly hopeful—flood defenses versus intensifying storms?
  - *Scale*: Cities like Lisbon shine, but what about villages—does Cēsis skew rural bias?
  - *Action*: Families pick Copenhagen—overcrowding by 2075? Governments fund Latvia—gentrification risk?
  - *Blind Spots:* What if Milan's tech boom flips its score—or a new climate threat (e.g., wildfires) hits Latvia?

Outcome

The Haven Index crowns Copenhagen and Latvia for 2025-2075—families relocate, governments fund Cēsis's clinics. Lisbon plans water reforms, Milan bets on grids. Random Forest does a good job giving us rankings within a wide range (92 vs 55), but refinement helps. Integrating XGBoost improves performance on edge cases and incorporates wildfire variables.

Next Steps

Refine the Random Forest Model for Precision.
Validate Rankings Against Real-World Trends.
Enrich Data for Comprehensive Coverage.

RECOMMENDATIONS

In order to achieve its objectives, ClimateWins needs to employ different machine learning techniques, depending on the specific goal:

- Identifying Weather Patterns Outside the Regional Norm in Europe:
  - Use Isolation Forest as Primary Tool: Its strength lies in fast, unsupervised anomaly detection across high-dimensional weather data (temperature, precipitation, wind), ideal for Europe's diverse climates.
  - Supplement with Autoencoders: The experiment noted Isolation Forest's focus on sharp anomalies—slow shifts (e.g., humidity drops).
- Generating Possibilities for Future Weather Conditions Over 20-50 Years:
  - Use GANs as Primary Tool: GANs generate diverse, non-linear scenarios (hot/dry, wet/wild), critical for 20-50-year horizons where physics models falter.
  - Augment with Variational Autoencoders (VAEs): GANs risk overfitting (Paris's 40°C too sharp). VAEs add uncertainty bands (e.g., 38-42°C).
- Determining the Safest Places to Live in Europe Over 25-50 Years:
  - Use Random Forest as Core Tool: Their strength in handling diverse features (climate risks, socio-economic metrics) and interpretability suits safety rankings.
  - Enhance with XGBoost: Random Forest miss complex interactions (Milan's grid vs. heat). XGBoost's gradient boosting lifts Milan's score accuracy.

Check out my Github for the Python scripts and dataset used in this analysis here.