Weather-data driven models for regional energy forecasting

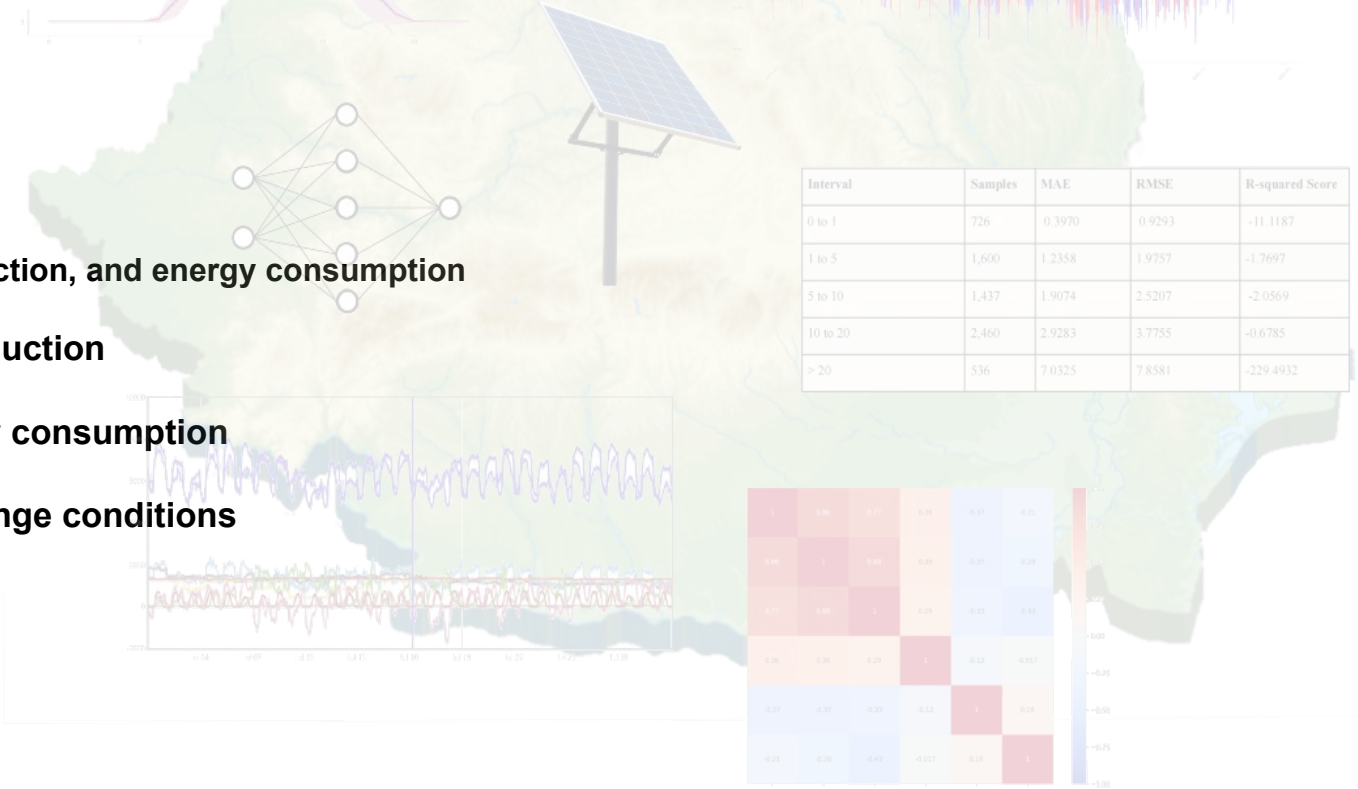
Eugen Mihuleț

WeADL 2025 Workshop

The workshop is organized under the umbrella of WinDMiL, project funded by CCCDI-UEFISCDI, project number PN-IV-P7-7.1-PED-2024-0121, within PNCDI IV

The project aims to apply forecasting models for PV energy production and energy consumption in Romania, based on meteorological data and using advanced machine learning techniques.

The main goal is to forecast photovoltaic energy production and energy consumption, considering climate variability and the impact of climate change, in order to support effective planning and management of energy resources at national and regional levels.



Interval	Samples	MAE	RMSE	R-squared Score
0 to 1	726	0.3970	0.9293	-11.1187
1 to 5	1,600	1.2358	1.9757	-1.7697
5 to 10	1,437	1.9074	2.5207	-2.0569
10 to 20	2,460	2.9283	3.7755	-0.6785
> 20	536	7.0325	7.8581	-229.4932

Specific objectives::

- **Analyze the correlations between weather, PV production, and energy consumption**
- **Apply ML models for forecasting PV energy production**
- **Apply ML models for forecasting national energy consumption**
- **Evaluate future energy trends under climate change conditions**

Project relevance:

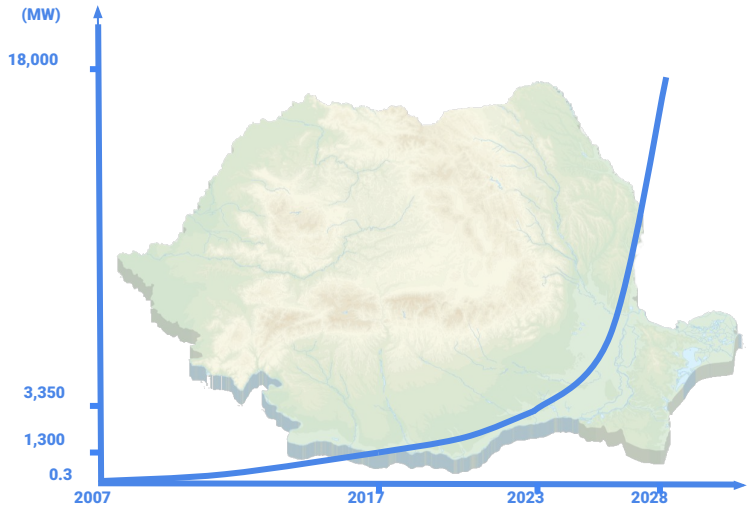
- Contributes to optimizing energy systems amid rising PV capacity and climate change
- Supports the transition to a smarter and more efficient electricity grid

Global trend:

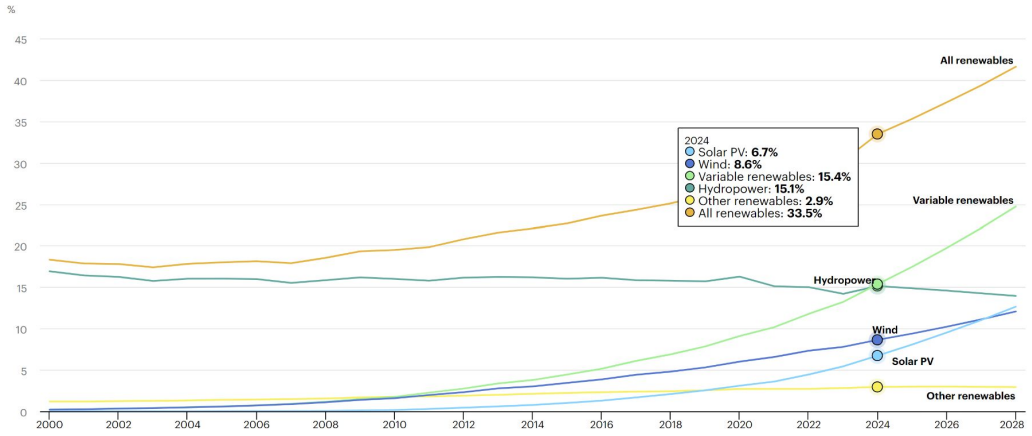
- Significant increase in the share of renewable energy (4.5x in Romania by 2028, according to SolarPower Europe)
- 151 countries have net zero emission targets, including Romania (by 2050)

Forecasting challenges:

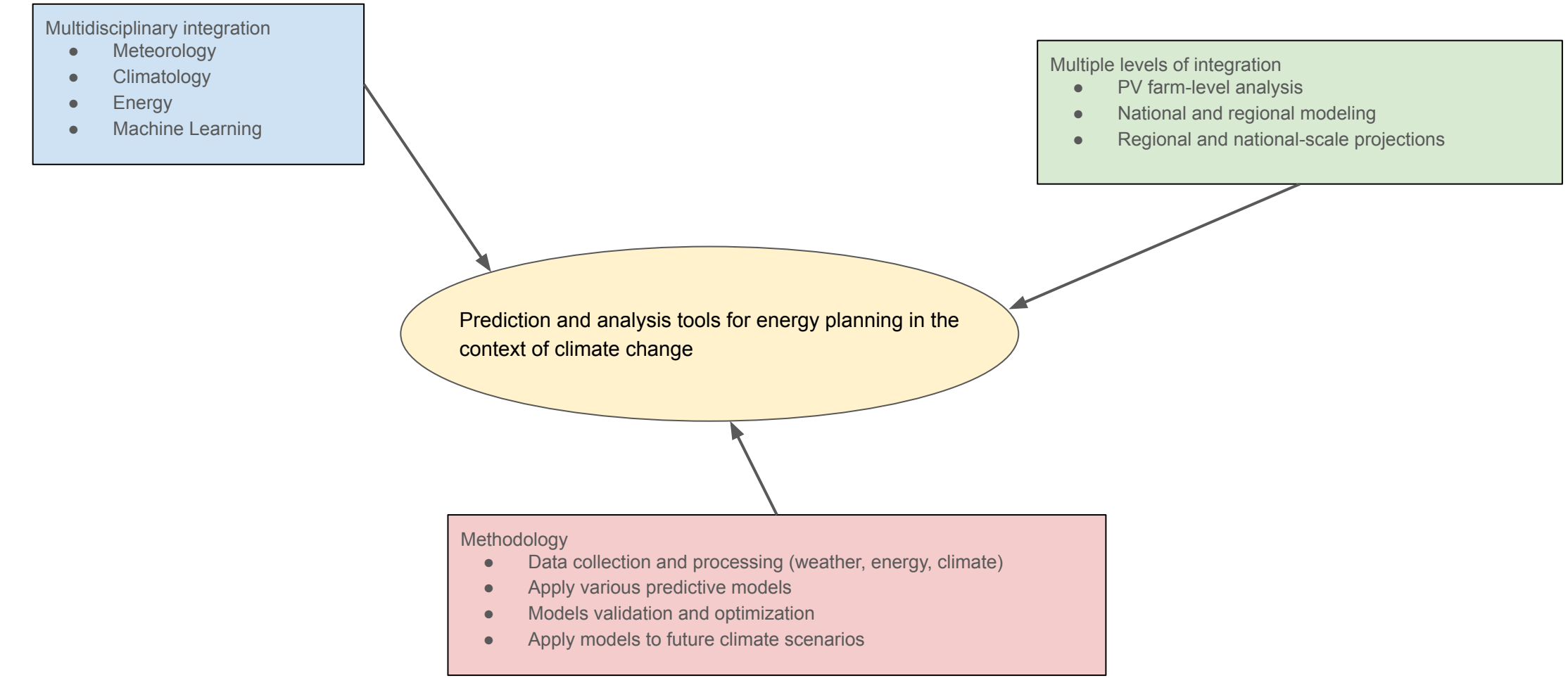
- Variability of renewable energy sources due to weather conditions
- Complex interactions between weather, production, and consumption
- Need for advanced models to balance supply and demand

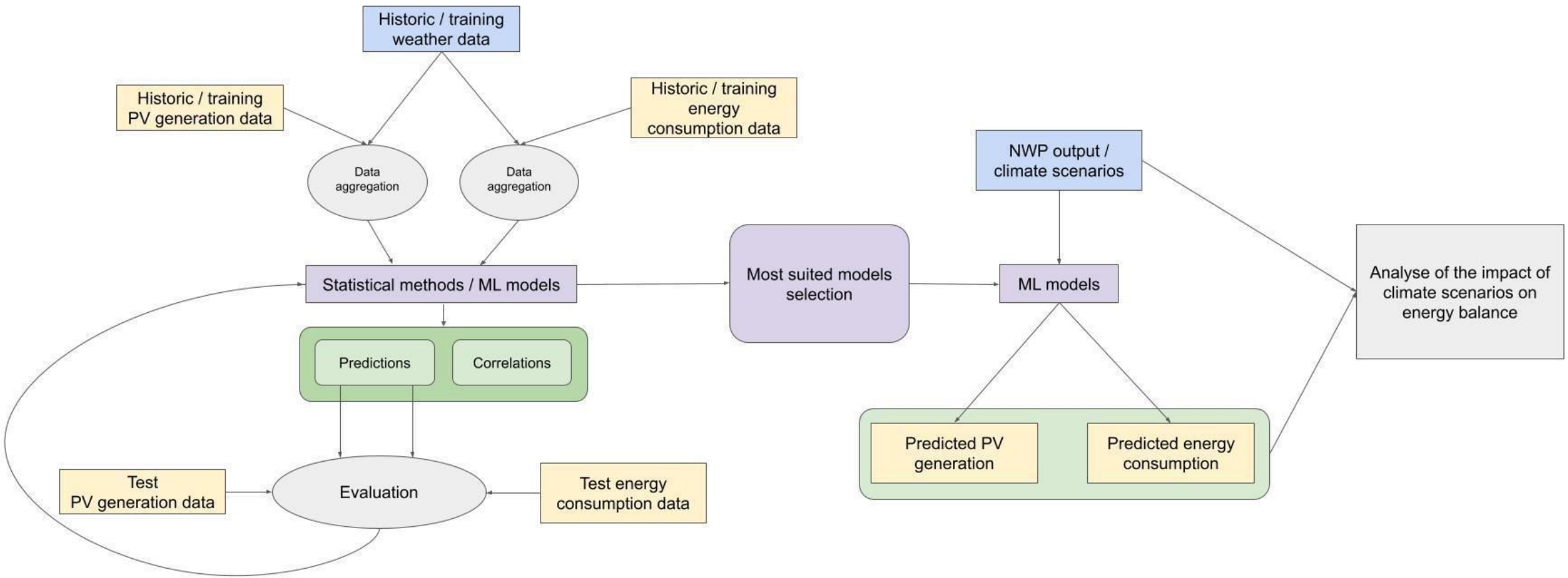


(Installed - also estimated - PV power in Romania Solar Power Europe)



(source: IEA (2024), Renewables 2023, IEA, Paris https://www.iea.org/reports/renewables-2023, Licence: CC BY 4.0)


Multidisciplinary integration
Meteorology
Climatology
Energy
Machine Learning
Multiple levels of integration
PV farm-level analysis
National and regional modeling
Regional and national-scale projections
Prediction and analysis tools for energy planning in the context of climate change
Methodology
Data collection and processing (weather, energy, climate)
Apply various predictive models
Models validation and optimization
Apply models to future climate scenarios


Historic / training weather data
Historic / training PV generation data
Historic / training energy consumption data
Data aggregation
Data aggregation
Statistical methods / ML models
Predictions
Correlations
Evaluation
Test PV generation data
Test energy consumption data
Most suited models selection
NWP output / climate scenarios
ML models
Predicted PV generation
Predicted energy consumption
Analyse of the impact of climate scenarios on energy balance

Employed data sources

Meteorological data:

- 135 ANM stations (<800 m elevation)
  - parameters: solar radiation, air temperature at 2m, wind speed/direction at 10m, humidity (2020–2024; temperature from 2008)
  - interval: January 2020 - July 2024 (January 2008 - 31 July 2024 pentru $T_{avrg\ daily}$)
- Solcast 15-min data: global, direct, and diffuse radiation, cloud opacity, air temperature, sun angles, wind, humidity

PV Data:

- UNISOLAR data, collected from PV farms from 5 campuses of La Trobe University, Victoria (Australia)
  - parameters: site metadata, 15-min solar generation, monthly summaries
  - interval: January 2020 - Decembre 2021

Energy consumption:

- Transelectrica national electricity data (2007–present) every 15 min
- Consumption and production by source (PV, wind, coal, hydro, nuclear, biomass, hydrocarbons)

Climate indices:

- HDD & CDD (Heating/Cooling Degree Days) daily since 2008 (source: CMCC Weather for Energy Tracker - daily resolution, 2008 - present)
- Used indices: CDD10, 16, 18, 21, 23, 26 and HDD14, 16, 18, 20; national average temperature T2m

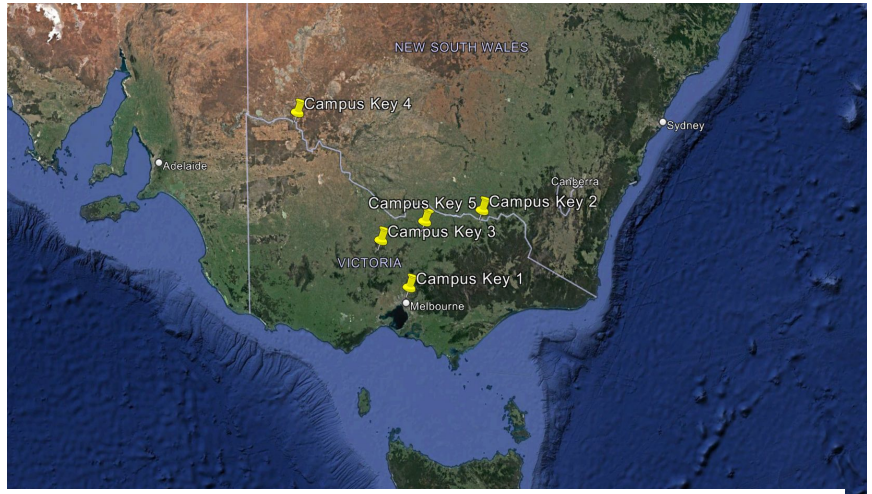
Methods for correlation and prediction

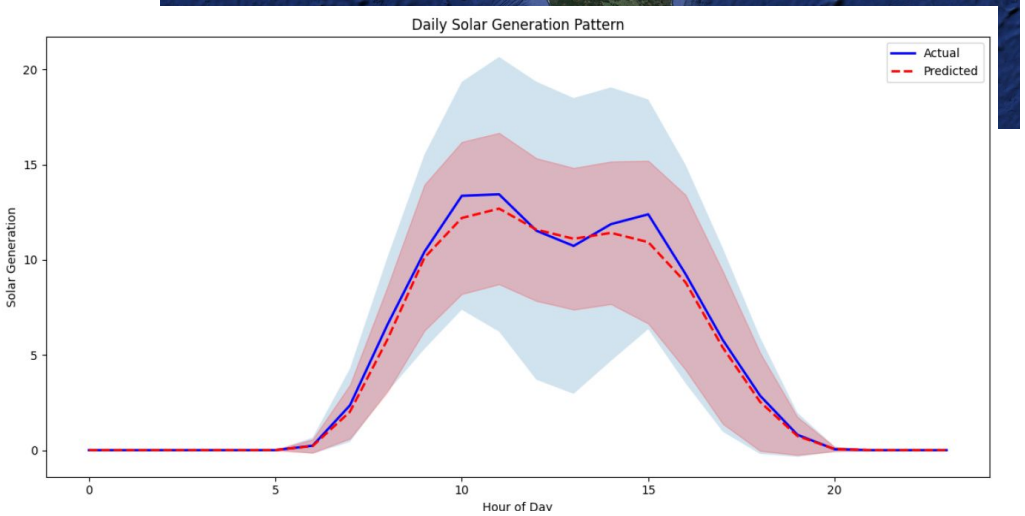
Machine Learning and Statistical Algorithms:

- Random Forest Regressor
- XGBoost
- Artificial Neural Networks (ANN)
- Linear Regression
- Pearson Correlation Coefficient (PCC)
- Principal Component Analysis (PCA)
- Performance metrics: MAE, RMSE, R^2

PV Generation Forecast Results

- 68,000 entries were used, spanning 2 years at 15-minute intervals, with 80% training data and 20% testing data - electrical energy generation and meteorological data (global radiation, air temperature, diffuse radiation, wind direction and speed, etc.)
- Several ML and statistical models were tested, with the best results achieved by Random Forest Regressor
- Global horizontal irradiance (GHI) was the most important component for prediction
- The best performance was obtained with Random Forest Regressor (R-squared score 0.7546, MAE 1.55, RMSE 3.3142 kW)
- Higher errors occurred at very low and very high generation values

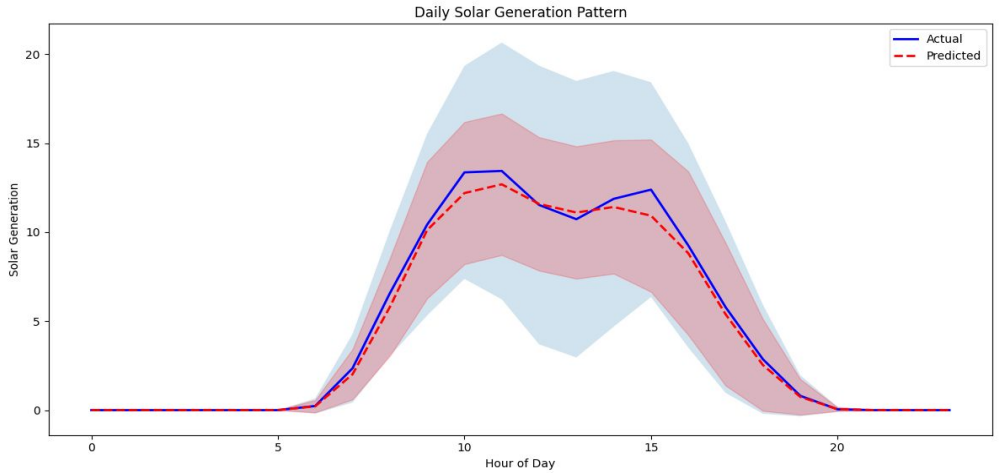


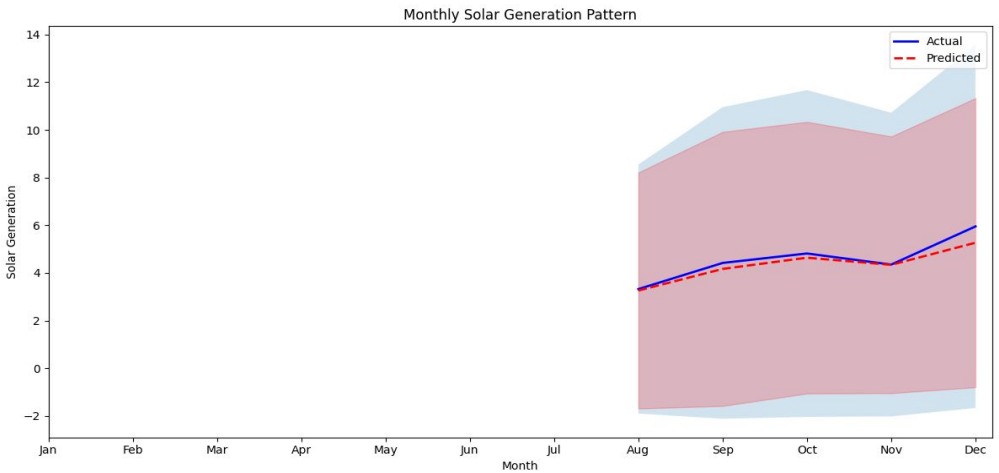


PV Generation Forecast Results

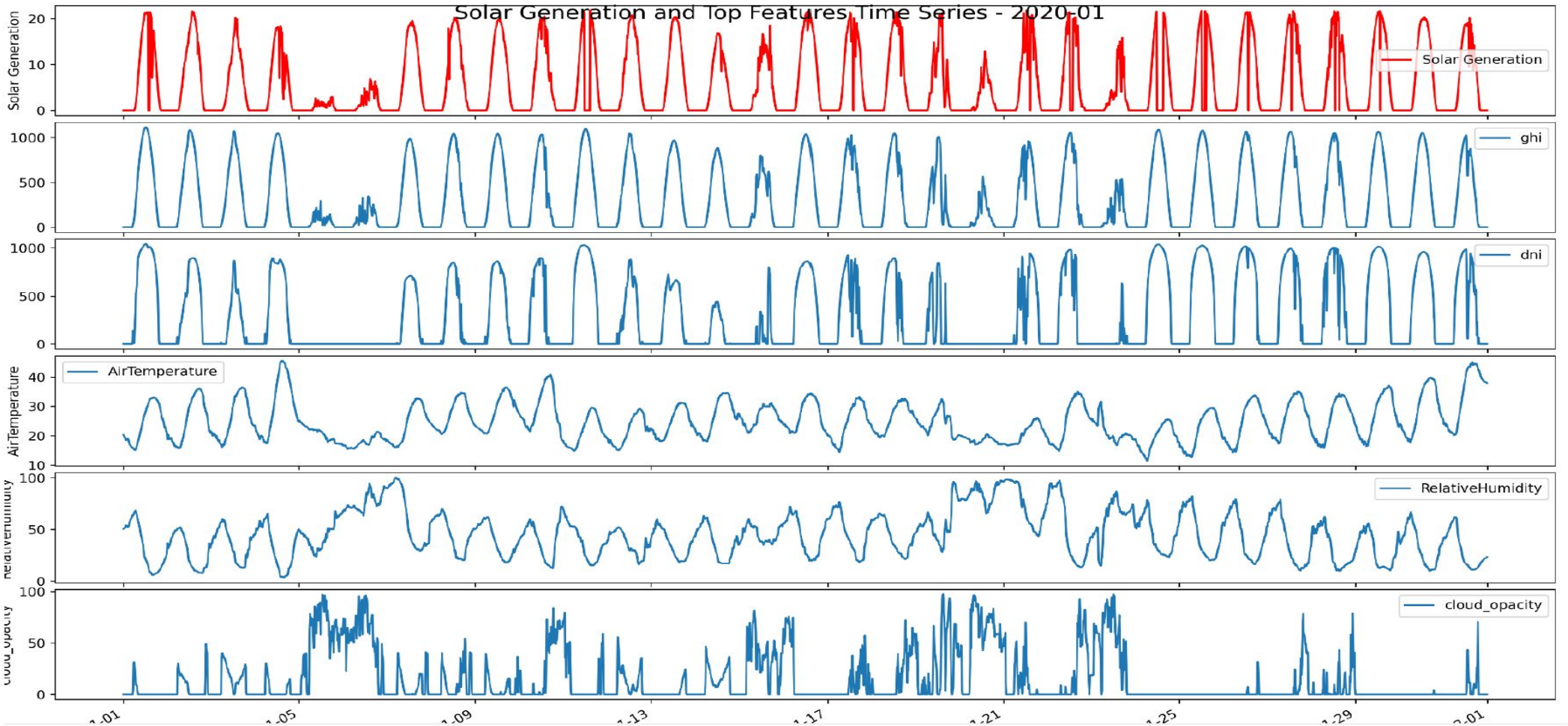
Metric	Overall	Zero Values	Non-Zero Values	Sensitivity Test
Samples	13,728	6,969	6,759	N/A
Mean Absolute Error	1.5504	0.7611	2.3642	1.19
Root Mean Squared Error	3.3142	3.0918	3.5288	2.71
R-squared Score	0.7546	0.0000	0.7268	0.85

Interval	Samples	MAE	RMSE
0 to 1	726	0.3970	0.9293
1 to 5	1,600	1.2358	1.9757
5 to 10	1,437	1.9074	2.5207
10 to 20	2,460	2.9283	3.7755
> 20	536	7.0325	7.8581





PV Generation Forecast Results



PV Generation Forecast Results

Metric	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Features Used	13 features*	6 features**	1 feature	9 features***	5 features****
MAE	**1.55**	1.57	1.99	1.62	1.59
RMSE	**3.31**	3.38	3.84	3.38	3.40
R-squared score	**0.75**	0.74	0.67	0.74	0.74

Metric	Morning (All)	Morning (Non-zero)	Noon (All)	Noon (Non-zero)	Afternoon (All)	Afternoon (Non-zero)
Samples	3,442	2,956	1,716	1,393	3,417	2,410
MAE	2.0570	1.9510	5.5176	4.2055	**1.3804**	1.8066
RMSE	3.4945	2.8790	7.1497	5.5129	**2.4808**	2.6815
R-squared Score	0.7456	0.8137	0.1173	0.1154	**0.8348**	0.8044

Metric	Data Subset	Random Forest	Linear Regression	ANN	XGBoost
Samples	Overall	13,728			
	Zero Values	6,969			
	Non-Zero Values	6,759			
MAE	Overall	**1.5504**	2.3801	1.7072	1.6154
	Zero Values	**0.7611**	1.7336	0.9655	0.9210
	Non-Zero Values	2.3642	3.0466	2.4720	**2.3313**
RMSE	Overall	**3.3142**	4.0029	3.5172	3.3521
	Zero Values	**3.0918**	4.0349	3.5224	3.1539
	Non-Zero Values	3.5288	3.9697	**3.5118**	3.5449
R-squared Score	Overall	**0.7546**	0.6419	0.7236	0.7489

Factors Affecting PV Generation Prediction

- Global horizontal irradiance (GHI):
  - Top predictor across all models
- Cyclic temporal characteristics:
  - time of day significantly improves model performance
- Complexity of relationships between parameters:
  - Random Forest and XGBoost indicate nonlinear relationships between meteorological variables and PV production
- Extreme values:
  - difficulties in predicting low production (0-1 kW) and high production (>20 kW) possibly caused by underrepresentation of these values in the training dataset
- Temporal dependencies:
  - including production from the previous hour improves predictions
  - highlights the importance of time series in PV modeling
- Temperature interacts with panel efficiency

Weather impact on National Energy Diagram

- Experiments using XGBoost Regressor for predicting energy consumption and production at national level
- Comparative analysis of predictions with and without including air temperature as a prediction factor
- Study of correlations between HDD (Heating Degree Days) and CDD (Cooling Degree Days) with energy consumption
- Investigating the impact of different temperature thresholds for HDD and CDD on prediction accuracy
- Implementation and testing of a model for predicting PV production at national level using global radiation data

Weather impact on National Energy Diagram

- Complex relationship between meteorological conditions and energy dynamics at national level
- A limitation of national-level analysis is the masking of local weather impact on consumption
- Disproportionate influence of meteorological events in densely populated areas
- Romania's diverse geography is a crucial factor in energy production and consumption models
- Demographic distribution: 54% urban vs. 46% rural, requires differentiated models
- Urban centers have a significant role in national energy consumption
- Future directions - using clustering techniques and NUTS3 data for granular analysis

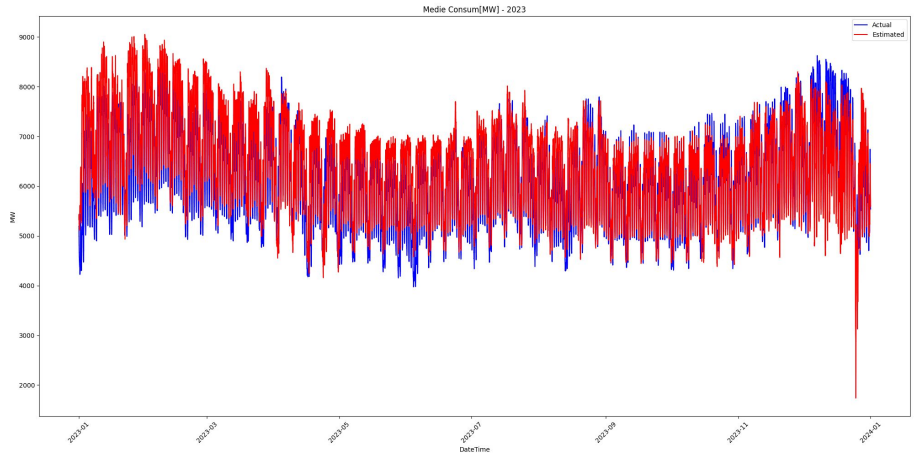
XGBoost Model for Forecasting Consumption and Production

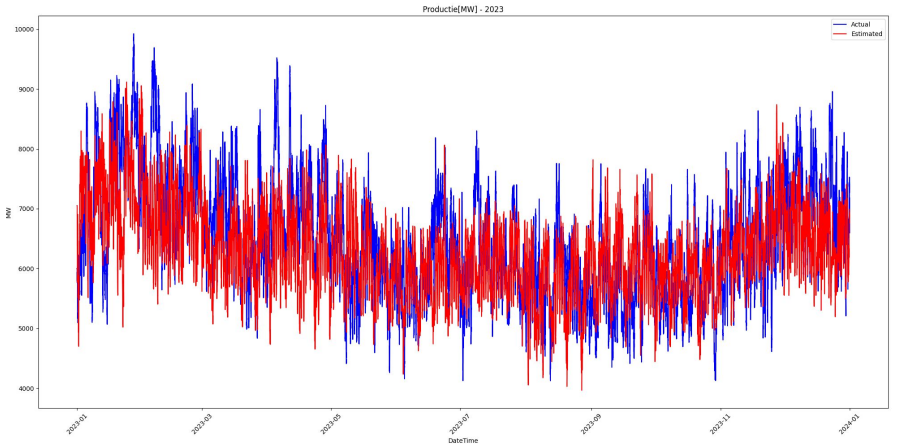
- Trained on 2022 data to predict 2023
- Model performance:

  Consumption forecast: RMSE = 8.93%, MAE = 6.86%, R^2 = 0.66

  Average consumption: RMSE = 8.87%, MAE = 6.76%, R^2 = 0.66

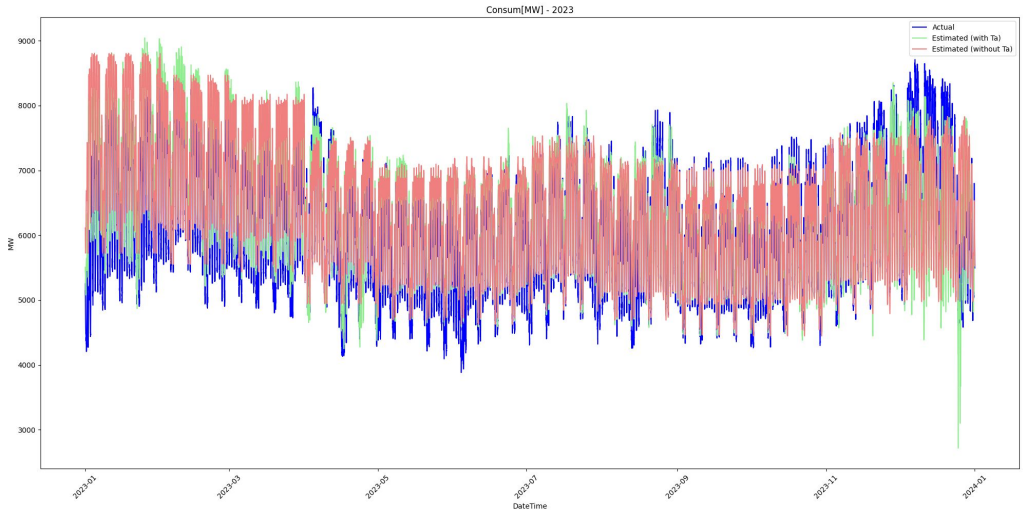
  Production forecast: RMSE = 13.50%, MAE = 10.67%, R^2 = 0.21

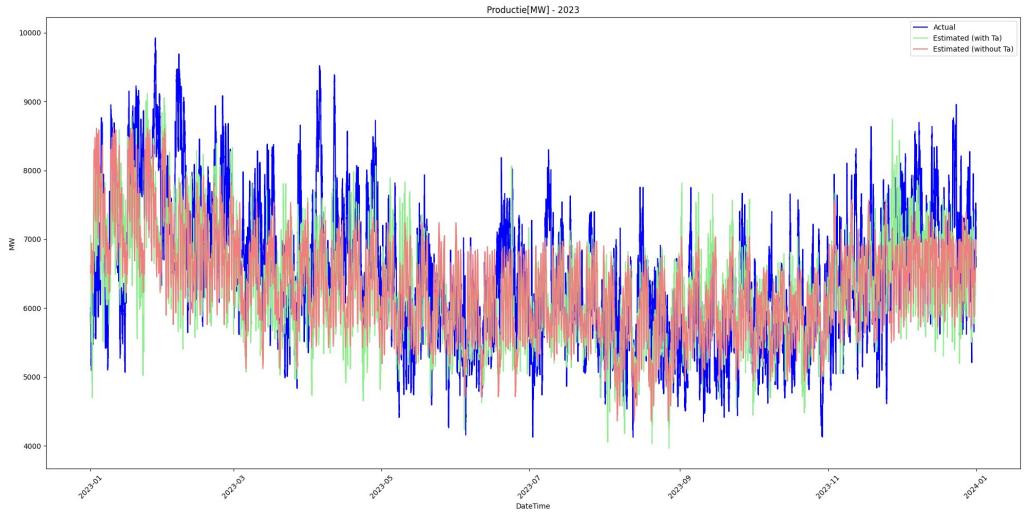




XGBoost with and without Temperature input

- 10-min Consumption:
  with Ta: RMSE = 8.93%, MAE = 6.86%, R2 = 0.66
  without Ta: RMSE = 10.23%, MAE = 7.85%, R2 = 0.56
- Production:
  with Ta: RMSE = 13.50%, MAE = 10.67%, R2 = 0.21
  without Ta: RMSE = 12.86%, MAE = 10.14%, R2 = 0.29
- Hourly Avg Consumption:
  with Ta: RMSE = 8.87%, MAE = 6.76%, R2 = 0.66
  without Ta: RMSE = 10.13%, MAE = 7.75%, R2 = 0.56

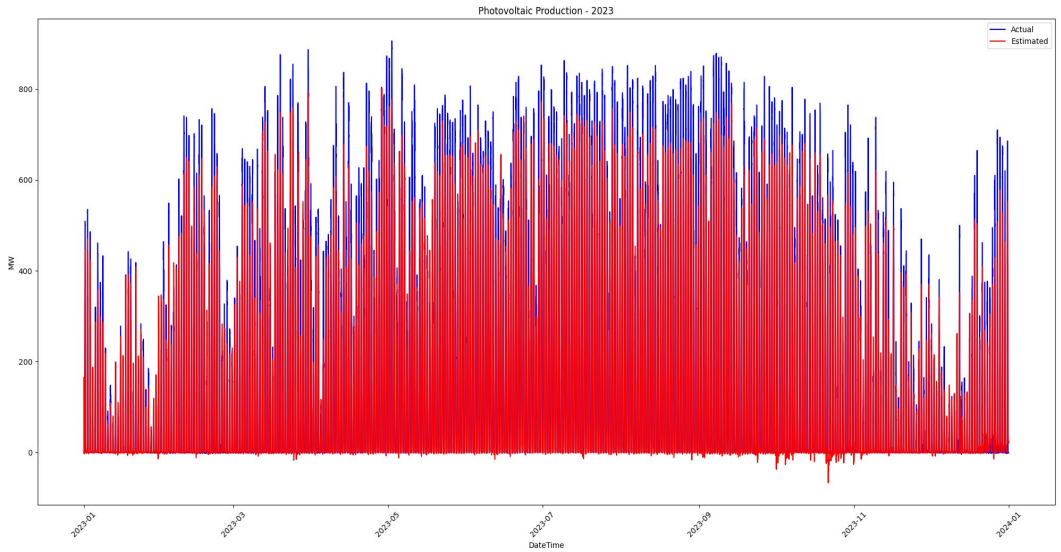


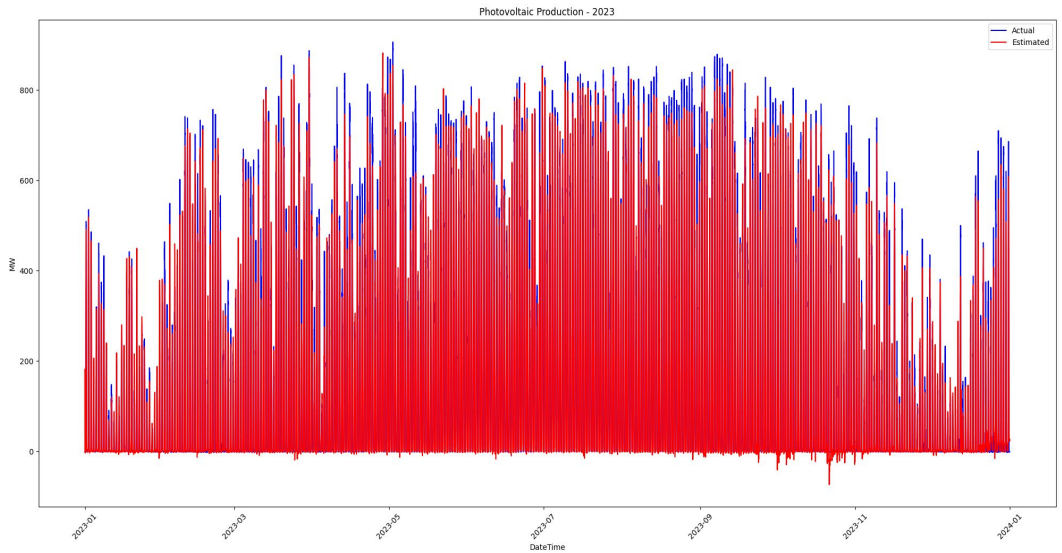


PV Production Prediction with XGBoost

- Performance metrics:

  RMSE (%) = 33.51 MAE (%) = 18.41 R2 Score = 0.95



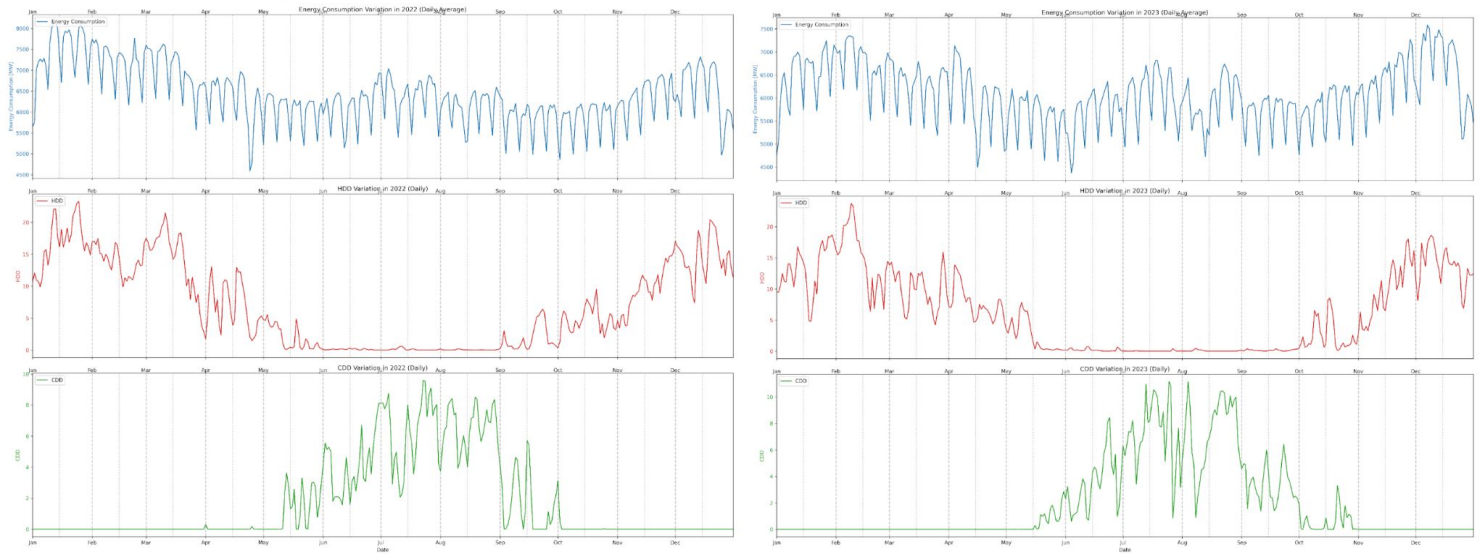


Correlation between HDD/CDD and Energy Consumption

Pearson Correlation Coefficients (CDD/ HDD vs energy consumption):

CDD 10C: **-0.3674** / CDD 16C: -0.2116 / CDD 18C: -0.1518 / CDD 21C: -0.0571/ CDD 23C: -0.0043 /CDD 26C: 0.0388

HDD 14C: **0.6371** / HDD 16C: 0.6277 / HDD 18C: 0.6115 / HDD 20C: 0.5909



Corelație HDD / CDD vs energy consumption in 2022 (left side) and 2023 (right side)

Factors affecting energy consumption prediction

- Limitations of national-level data - masking regional variations
- Air temperature - crucial factor for prediction accuracy
- Importance of temporal characteristics: time of day (most important factor), day of the week, legal holidays
- More accurate prediction for consumption than for production
- Weak and unexpected negative correlation with CDD18 (Cooling Degree Days)
- Different performance of HDD and CDD correlations: strong seasonal influences
- National PV production prediction - high R^2 score, but relatively large RMSE (likely due to sustained growth in national PV capacity)

Conclusions

- ML models appear promising for predicting PV generation and national energy consumption
- Identified limitations:
  - Using national-level data leads to masking of local weather characteristics
  - Production and consumption depend on many factors that cannot be integrated into models
  - Annual increase in production capacity and changes in consumption profile
  - Limited prediction capacity for extreme meteorological events
- Next research activities:
  - Implementation of data clustering techniques
  - Use of NUTS3 energy data for granular analysis
  - Exploration of urban-rural differences and zonal population density characteristics in energy consumption
  - Testing of other ML models (e.g., LSTM)
  - Integration with climate scenarios (e.g., CMIP5)