Stochastic time series forecasting approach using Markov Chain for Robust Energy System Planning

Rohith Bala Krishnan, Sebastian Voswinckel, Theresa Reinhardt, Viktor Wesselak and Christoph Schmidt

Institute for Renewable Energy Technology, Nordhausen University of Applied Sciences, 99734 Nordhausen (Germany)

Abstract

Time series forecasting is essential for accurately modeling renewable energy system, particularly when dealing with the variability and uncertainty of generation from sources like Wind or Photovoltaics (PV). This paper addresses uncertainties in energy system modelling by proposing a Markov Chain integrated Machine Learning method for time series generation. Using photovoltaic feed-in profile as a case study, the stochastic-approach is validated against the historical data. Applied to a household PV system in Germany and optimized using the open energy modelling framework (oemof), the method generates multiple future scenarios. A comparative analysis reveals that this probabilistic approach provides a superior representation of uncertainty, resulting in more robust energy system outcomes compared to traditional methods.

Keywords: Time Series Forecasting, Renewable Energy, Photovoltaic, Markov Chain, Machine Learning, Uncertainty, Energy System Modelling, oemof, Germany, Stochastic Forecasting

1. Introduction

The increasing penetration of renewable energy sources has elevated the importance of energy system modelling as a critical methodology for optimizing power generation, distribution and consumption (REN21, 2021). However, the inherent variability and stochastic nature of renewable generation introduces significant challenges for traditional modeling paradigms, particularly from photovoltaic systems (PV). Conventional approaches frequently employ deterministic methods, such as use of Typical Meteorological year (TMY), which operates under the assumption of stationary future conditions based on historical patterns. This methodological limitation does not adequately capture the stochastic nature of renewable electricity generation and thus compromises the reliability of model results for strategic energy planning, especially in scenarios characterized by a high share of renewable energies.

To overcome these limitations, this research introduces a stochastic time series-forecasting framework based on Markov Chain modeling, computationally enhanced through integration with machine learning techniques. This hybrid methodology leverages multi-year historical data to probabilistically model temporal dependencies and state transitions within generation profiles. The Markov Chain formalism enables the generation of multiple, equiprobable future scenarios, thereby providing a more comprehensive representation of possible deviations compared to single-point forecasts. This approach facilitates robust energy system optimization by explicitly accounting for the inherent variability in renewable generation.

The practical efficacy of the proposed methodology is demonstrated through a case study focusing on PV feed-in profiles for a residential energy system in Germany. Utilizing the Open Energy Modelling Framework (oemof) for system optimization, a comparative analysis is performed between the stochastic Markov Chain approach and conventional forecasting methods. Empirical results indicate that the proposed methodology significantly enhances system robustness and operational reliability, providing a more resilient foundation for energy system design and decision-making in renewable-dominated energy landscapes.

10.18086/swc.2025.05.20 Available at http://proceedings.ises.org

2. Methodology: Stochastic Time-Series Modeling Framework

The forecasting method used in this research combines historical solar radiation data and Markov Chain-Machine learning hybrid approach to predict the feed-in power of a PV system over time. The methodology begins with acquiring high-resolution historical solar data from PVGIS-ERA5 database (Thomas Huld, 2012) for Nordhausen, Germany (51.505°N, 10.791°E). The dataset spans 19 years (2005 -2023) with hourly temporal resolution. The raw irradiation data undergoes normalization using Clear-Sky Index methodology (Ineichen & Peres, 2002). This normalization effectively decouples astronomical from meteorological effects, enabling more robust stochastic modeling.

$$CSI\ (t) = \frac{G_{\text{measured}}(t)}{G_{\text{clear-sky}}(t)} \qquad \text{(eq. 1)}$$

where $CSI\ (t) \in [0,1]$, with values clipped to handle numerical instabilities. The continuous CSI values are discretized and classified into five operational states using predetermined thresholds:

$$\text{State} = \begin{cases} \text{Night} & \text{if } CSI = 0 \\ \text{Overcast} & \text{if } 0 < CSI \ \leq 0.25 \\ \text{Cloudy} & \text{if } 0.25 < CSI \ \leq 0.5 \\ \text{Partly cloudy} & \text{if } 0.5 < CSI \ \leq 0.75 \\ \text{Clear} & \text{if } 0.75 < CSI \ \leq 1.0 \end{cases}$$

The methodology explicitly models each month separately to account for fundamental seasonal variations in weather patterns and solar geometry. This approach recognizes the seasonal transitions patterns, solar geometry effects and month-specific climate characteristics. The monthly segmentation approach also significantly enhances the physical realism and statistical accuracy of the generated scenarios compared to annual models that average out critical seasonal differences.

2.1. Conditional Entropy Analysis for Markov Order Optimization

A critical innovation involves the data-driven determination of the optimal Markov order (N) through conditional entropy analysis. The algorithm systematically computes conditional entropy values for temporal orders ranging from 1 to 48 hours, identifying the precise point of diminishing returns where additional historical context provides minimal predictive improvement.

The conditional entropy H(Y|X) quantifies the uncertainty remaining in predicting future irradiance values Y given historical observations X of length N (Shalizi & Crutchfield, 2001). As N increases, conditional entropy typically decreases, reflecting improved predictive capability through expanded historical context. The conditional entropy is defined as:

$$H\ (S_{t+1} | S_t, S_{t-1}, \dots, S_{t-N+1}) = -\sum_{s_{t+1}, s_t, \dots, s_{t-N+1}} P\ (s_{t+1}, s_t, \dots, s_{t-N+1}) \log P(s_{t+1} | s_t, \dots, s_{t-N+1}) \qquad \text{(eq. 2)}$$

The optimal order selection employs a strength method with double differences to precisely identify the strongest point, indicating that further increases in order provide negligible predictive benefit relative to model complexity. This data-driven approach automatically prevents overfitting while capturing essential temporal dependencies inherent in solar irradiance patterns. Notably, the optimal order varies monthly to reflect changing meteorological patterns: shorter orders (12-24 hours) typically suffice during stable conditions, while longer orders (24-36 hours) may be necessary during dynamically changing weather, particularly in regions experiencing frequent frontal passages. The conditional entropy framework provides a rigorous mathematical foundation for temporal dependency analysis, ensuring that the Markov model incorporates sufficient historical context without unnecessary complexity.

Figure 1 and Figure 2 represents the entropy curve and the strength analysis for the month of July respectively. Figure 3 represents the entropy analysis with optimum Markov for each month. As previously mentioned it is clearly observed in figure 3 that the optimum order is dynamic for each month depending on the weather conditions, which vary from month to month.

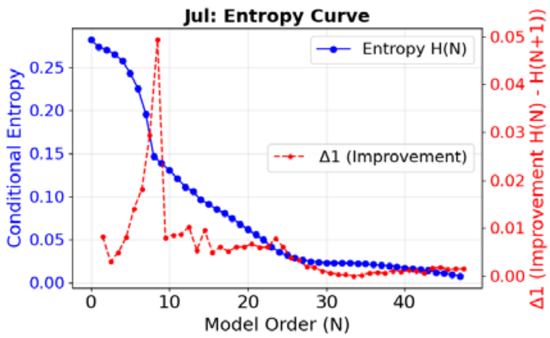

Figure 1: Entropy curve and first difference for the month of July

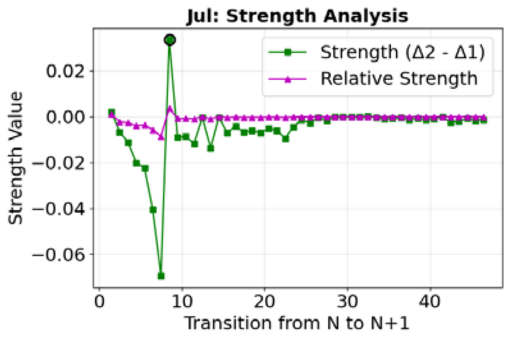

Figure 2: Markov Order Optimization (highlighted with black circle) through Entropy Strength Analysis

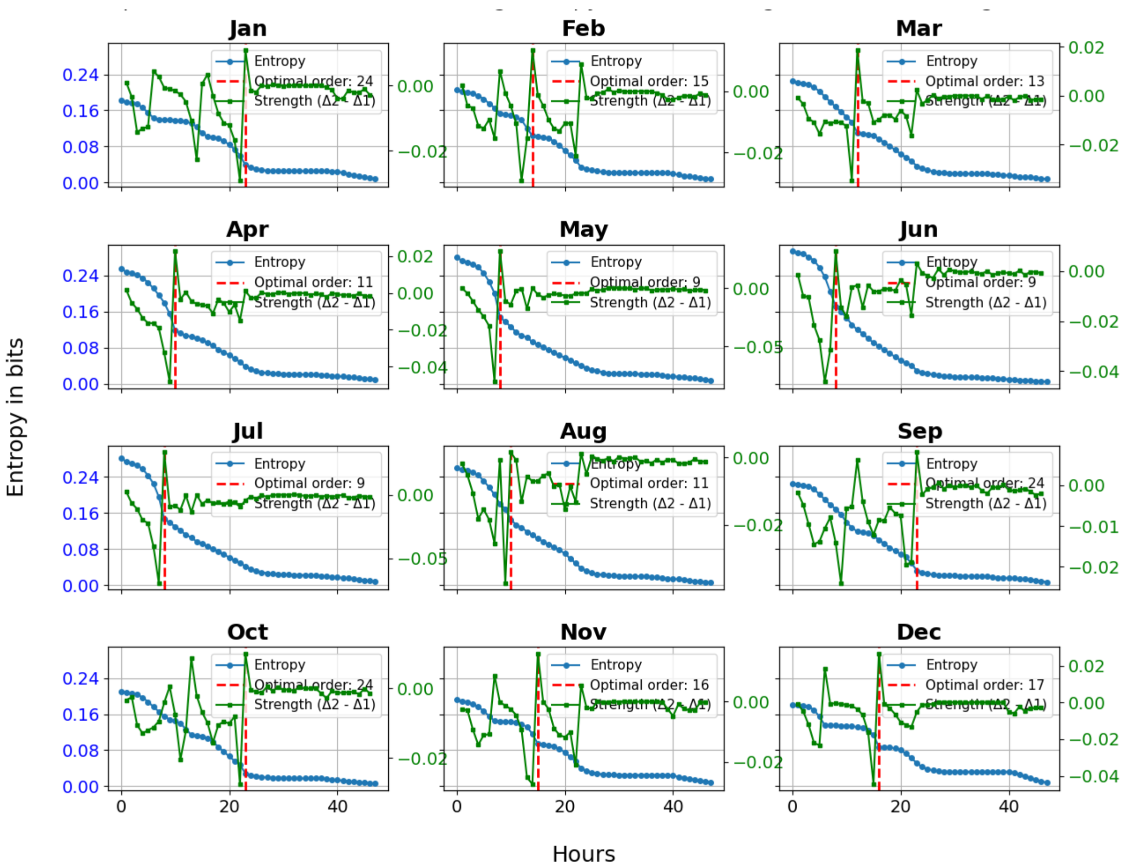

Figure 3: Entropy-Strength analysis with optimum Markov's order for each month

2.2. Novel Markov Chain-Machine Learning Hybrid Approach

To overcome the limitations of traditional Markov chains, which consider only immediate state transitions and ignore critical patterns such as diurnal cycles, seasonal variations, and complex multi-scale dependencies, a hybrid feature engineering approach was implemented. This enhancement incorporates a rich set of features including recent state history, statistical trends (mean and standard deviation), cyclical time encodings (hourly sine and cosine), and physical indicators (daylight masks), enabling the Random Forest model (L.Breiman, 2001) to learn non-linear relationships and contextual dependencies. This integration ensures that diurnal patterns adhere to physical reality, accounts for the differing behavior of meteorological states at various times of day, and significantly improves predictive accuracy by capturing the nuanced interactions between temporal context and irradiance dynamics. The random forest classifier utilizes comprehensive feature vectors:

$$X_t = \left[s_{t-N}, s_{t-N+1}, \dots, s_{t-1}, \mu_{\mathrm{window}}, \sigma_{\mathrm{window}}, \sin\left(\frac{2\pi h}{24}\right), \cos\left(\frac{2\pi h}{24}\right), 1_{\mathrm{daytime}}\right] \quad \text{(eq. 3)}$$

Where,

s_{t-i} : Previous states (encoded numerically)

μ_{window} : Rolling mean of previous states

σ_{window} : Rolling standard deviation

h : Hour of day

1_{daytime} : Daytime indicator function

This hybrid approach is coupled with a probabilistic scenario generation system using Monte Carlo simulations to quantify forecast uncertainty, producing an ensemble of possible futures through controlled random sampling. This method also incorporates a noise injection mechanism that selects subsequent states based on class probabilities, introducing variability while maintaining physical and statistical constraints. The critical importance of this approach for energy systems lies in its ability to quantify uncertainty, providing a distribution of potential outcomes rather than a single deterministic forecast. This enables robust optimization that performs reliably across diverse scenarios and facilitates rigors risk assessment by calculation probabilities for extreme events, such as prolonged cloudy periods.

To ensure unwavering physical realism, the model enforces astronomical validation through solar position calculations and state transition rules, guaranteeing that all generated sequences adhere to meteorologically plausible dynamics and historically observed irradiation bounds, thereby producing synthetic data that is both statistically sound and physically consistent. Furthermore, to comprehensively characterize the inherent variability and uncertainty in solar generation profiles, 1,000 independent iterations were simulated for each calendar month, with the resultant data aggregated and visualized in the form of heat maps. Figure 4 represents the state prediction heat map for the month of July with a mean prediction accuracy of 90.2 %.

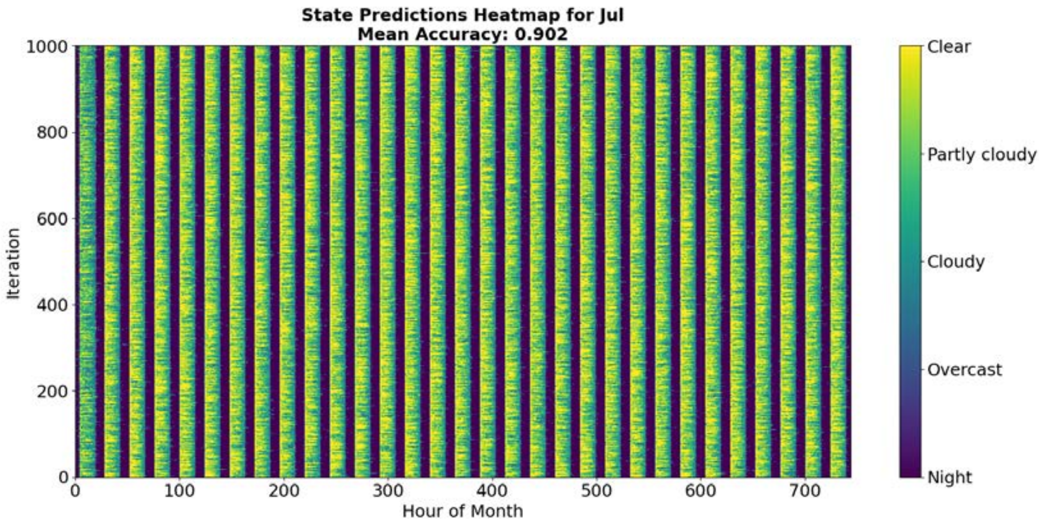

Figure 4: Predicted state heat map over N iterations for the month of July

To provide a granular perspective on the variability between individual simulation runs, a zoomed-in view of a representative subset of these iterations is additionally presented in figure 5, illustrating the diverse range of potential daily generation trajectories and underscoring the critical importance of modeling this temporal stochasticity for robust energy system planning. The mean accuracy for each month μ_{acc}^{m} quantifies the Markov chain-ML hybrid model's average state prediction reliability. It also demonstrates significant variation driven by seasonal weather patterns and the inherent dynamism present in the historical data, where lower accuracy values correspond to month characterized by higher meteorological variability and less predictable state transition. As quantified in Table 1, months with diminished accuracy, such as April or June, are not indicative of model deficiency but rather reflect periods where the historical data itself was highly dynamic, featuring rapid and less persistent shifts between weather states like Clear and Cloudy. Conversely, higher accuracy scores during August or November signify periods of greater atmospheric stability and more predictable patterns in the historical record. This relationship confirms that the model accurately captures the intrinsic predictability of each season, with the accuracy metric μ_{acc}^{m} thereby serving as a direct indicator of historical data variability and forecasting difficulty for different times of the year.

$$\mu_{acc}^{m} = \frac{1}{\mathrm{K}}\sum_{\mathrm{K}=1}^{\mathrm{K}}\left(\frac{1}{N_{\text{test}}^{\text{m}}}\sum_{t=1}^{N_{\text{test}}^{\text{m}}} 1(\hat{s}_t^k = s_t^m\right) \qquad \text{(eq. 4)}$$

Where,

K – Monte Carlo iterations

$N_{\text{test}}^{\text{m}}$ – test hours in month m

$1(\hat{s}_t^k = s_t^m)$ – indicates correct state classification

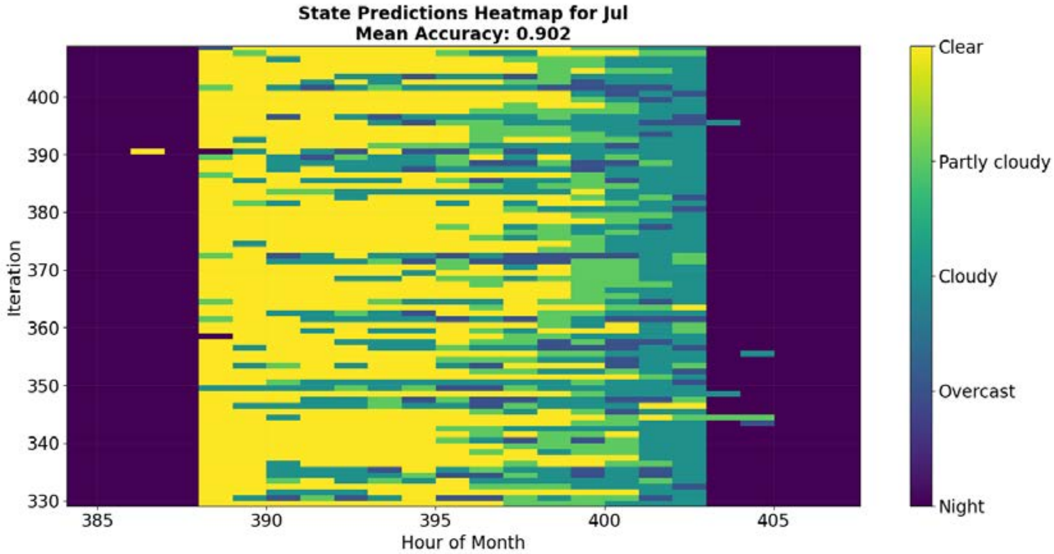

Figure 5: Predicted state heat map (cropped view) for one day in July to show variability in the predicted profiles

Table 1: Monthly forecasting accuracy and variability metrics

Month	Minimum Accuracy (%)	Maximum Accuracy (%)	Mean Accuracy (μ_{acc}^{m}) (%)	Std. Deviation
January	81	86	84	0.009
February	86	93	90	0.011
March	78	85	82	0.013
April	66	73	68	0.009
May	84	87	85	0.006
June	78	80	79	0.006
July	88	91	90	0.004
August	95	98	97	0.01
September	79	84	82	0,008
October	98	99	98	0.003
November	92	95	94	0,005
December	88	93	91	0.008

2.3. Irradiation Reconstruction from Predicted States

The conversion from meteorological states to quantitative irradiation values employs a state-dependent sampling approach where, for each predicted state at hour h, irradiation values are drawn from truncated normal distributions specific to both the state and hour of day:

$$G_{\text{predicted}}(t) \sim \text{TN}\left(\mu_{\text{state,h}}, \sigma_{\text{state,h}}, a_{\text{state,h}}, b_{\text{state,h}}\right) \quad \text{(eq. 5)}$$

The distribution parameter are carefully calibrated using historical data, with the mean $\mu_{\text{state,h}}$ defined as the midpoint between the 10^{th} and 90^{th} percentiles of historical observation for that specific state and hour, and standard deviation $\sigma_{\text{state,h}}$ derived as one-fourth of the range between these percentiles, ensuring that the sampled values remain within physically realistic bounds while reflecting observed variability.

$$\mu_{\text{state,h}} = \frac{G_{\text{low,state,h}} + G_{\text{high,state,h}}}{2} \quad \text{(eq. 6)}$$

$$\sigma_{\text{state,h}} = \frac{G_{\text{high,state,h}} + G_{\text{low,state,h}}}{4} \quad \text{(eq. 7)}$$

The boundary parameters $a_{\text{state,h}}$ and $b_{\text{state,h}}$ corresponds directly to these percentile derived limits: $G_{low,\text{state,h}}$ (10^{th} percentile) and $G_{\text{high,state,h}}$ (90^{th} percentile) effectively constraining the generated irradiation within historically observed ranges for each unique state-hour combination.

$$G_{\text{low,state,h}} = percentile\left(G_{\text{historical,state,h}}, 10\right) \quad \text{(eq. 8)}$$

$$G_{\text{high,state,h}} = percentile\left(G_{\text{historical,state,h}}, 90\right) \quad \text{(eq. 9)}$$

Figure 6 represents a heat map with the average solar irradiation patterns across different months and hours of the day, categorized by predefined cloud conditions. While this visualization displays average irradiation values for clarity, the optimization algorithm utilizes the boundary cases (upper/lower bounds of each cloud condition category) rather than mean values. This bounds-based approach ensures robust system design against worst-case scenarios while maintaining computational efficiency.

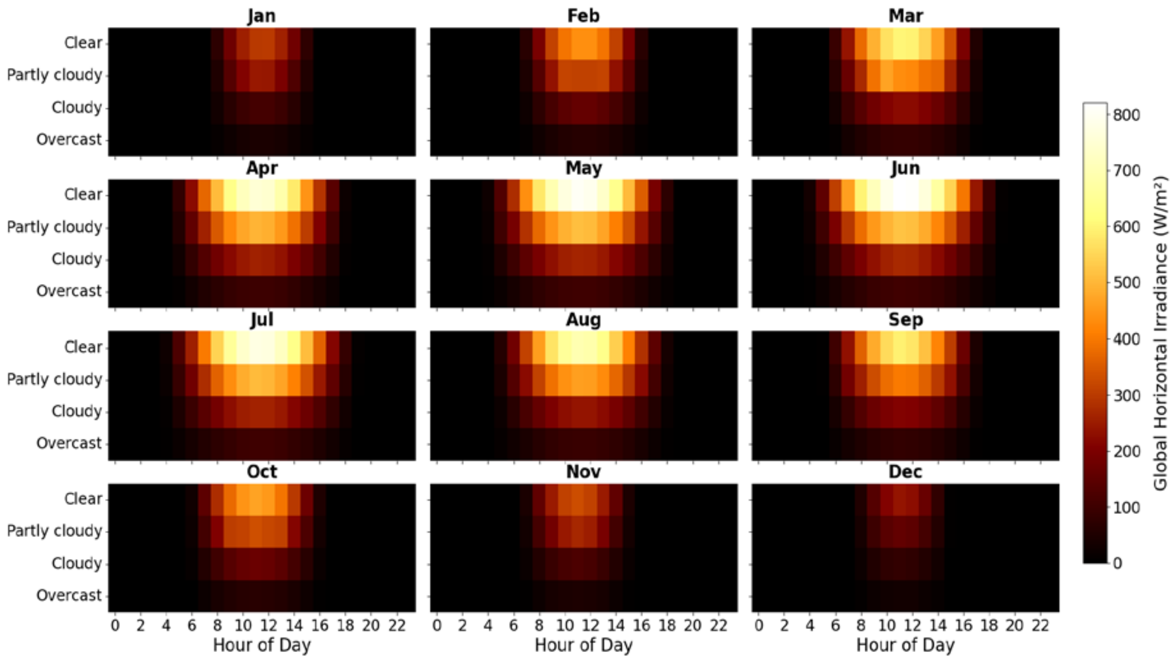

Figure 6: Monthly Average Solar Irradiation Profiles by Hour and Cloud Condition

In figure 5, one can also observe that certain algorithmic iteration generate false hour predictions during nocturnal hours, which contravenes physical solar geometry constraints. These unphysical outputs are systematically rectified through implementation of boundary conditions that enforce zero irradiation values during periods of astronomical night, thereby ensuring physical realism and model consistency with fundamental energy resource availability.

2.4. Predictive Validation Using Historical Consistency Metrics

To comprehensively validate the physical plausibility of the generated scenarios, a historical consistency analysis that quantifies the alignment between predicted irradiation profiles and observed historical patterns was employed. The core of this validation is the overlap percentage metric, which calculates the proportion of the predicted mean profile that falls within the 10th to 90th percentile range of the historical data for each corresponding hour. A high overlap percentage (typically >70 %) is a strong indicator that the forecasts are physically realistic and do not extrapolate beyond historically observed meteorological conditions, thereby ensuring the generated scenarios are both novel in their temporal sequencing yet grounded in empirical reality. This metric also serves as a crucial complement to conventional accuracy scores by evaluation distributional realism rather than just state-by-state correctness.

$$Overlap\ \% = \frac{100}{N}\sum_{i=1}^{N} 1\left(\hat{G}_i \in [G_{10\%}(i), G_{90\%}(i)]\right) \qquad \text{(eq. 10)}$$

Where $\hat{G}_i$ represents the predicted mean irradiation at time step I, $G_{10\%}(i)$ and $G_{90\%}(i)$ are the 10th and 90th percentiles of historical observation at the corresponding temporal position. The comparative plot figure 7 visually contrasts the predicted ensemble of irradiation profiles against the historical variability envelope for a month in winter and summer. Figure 7 also represents the forecasted mean, its standard deviation and how they align with the observed historical percentiles, mean and extremes. The overlap percentages, which quantify the physical plausibility of these predictions, are summarized in table 2 for each month, providing a clear metric of how well the stochastic forecasts replicate the bounds of historically observed patterns.

Table 2: Monthly overlap percentages quantifying the physical plausibility of the prediction

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Overlap percentage	88.3	94.5	89.4	85.7	83.8	84.6	86.4	80.1	83.9	92.9	94.3	91.7

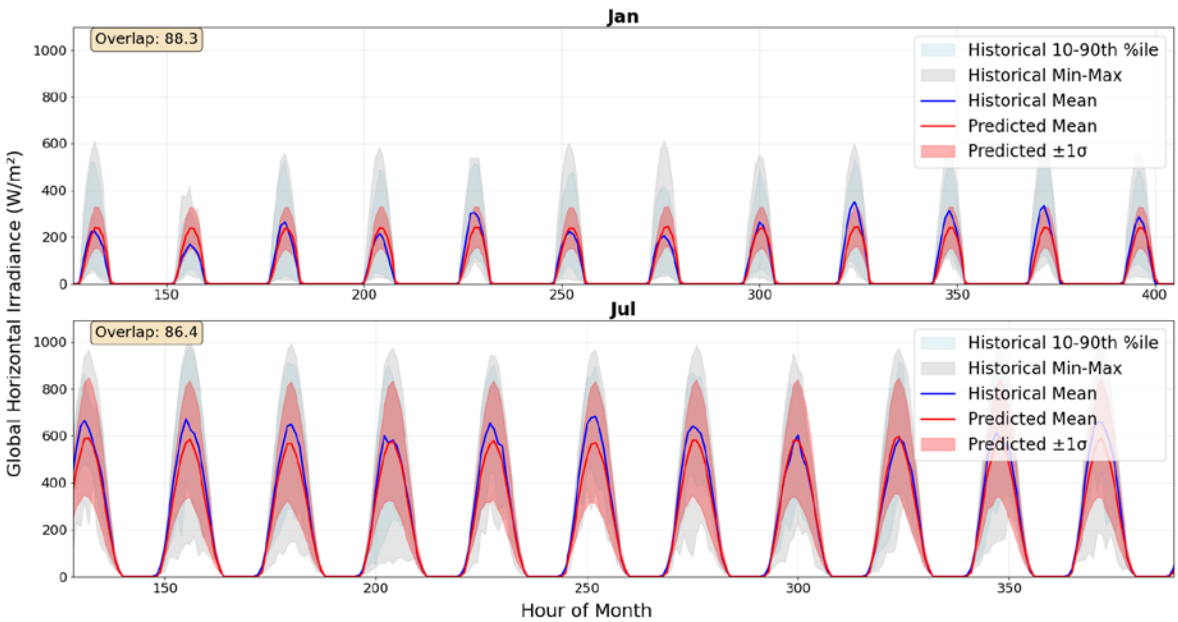

Figure 7: Validation of predicted profile against historical variability envelope plotted for the month of January (winter) and July (summer)

2.5. Predictive Validation Using Historical Consistency Metrics

This section details the final conversion of the stochastically generated irradiation data into photovoltaic (PV) power output, a critical step for energy system modelling. For each forecast iteration, the physically consistent solar irradiance components (direct normal irradiance and diffuse horizontal irradiance) from the Markov-Chain generated global horizontal irradiation is reconstructed using the Erbs decomposition model, which correlates the diffuse fraction with the clear-sky index (Erbs et al., 1982). These derived irradiance components, along with ambient temperature and wind speed from Typical Meteorological Year data, serve as inputs to a detailed PV system model within the PVlib Python library (Kevin S. Anderson, 2023). The model simulates the performance of a specific 6.6 kW_p rooftop system using Canadian Solar modules and a Power-One inverter, accounting for factors like solar geometry, module temperature, and system losses, ultimately producing normalized AC power feed-in profiles for each scenario that are physically credible and directly usable for energy system optimization in tools like oemof.

3. Enhancing model robustness

Oemof is a Python toolbox for energy system modelling and optimization (Hilper, S. et al., 2018). The inbuilt package oemof.solph is used to build and optimize the energy system with the help of a commercial (e.g. Gurobi or Cplex) or open-source solvers (e.g. CBC (Coin-or branch and cut) and GLPK) (Krien, U. et al., 2020). This case study models an energy system for a single-family household to analyze whether a dynamic time series (PV feed-in profile), generated through probabilistic forecasting, significantly influences the system performance. Figure 8 represents the basic structure of the energy system model to understand the flow of the system. To create such energy systems, either code-based scripts or browser-based drag & drop tools such as openPlan (RLI, 2025) or EnSys (in.RET, 2025) can be used. The energy system model optimizes a household configuration with 1000 iterations using key parameters: 3,500 kWh electrical demand, 16,000 kWh heat demand, 1,500 kWh e-mobility demand, 4 % interest rate, and 0.4 €/kWh electricity price, with technology investments determined through annuity calculations. This case study reveals how stochastic optimization captures capacity trade-offs across PV, heat pumps, batteries, and thermal storage under varying generation patterns.

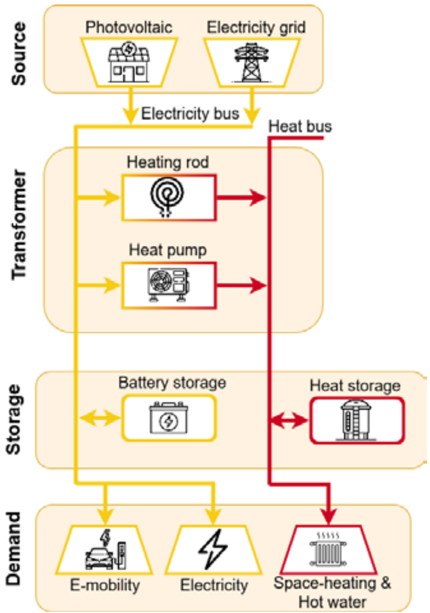

Figure 8: Basic structure of energy system model for a single-family household in Germany using block diagram

Figure 9 quantifies the impact of meteorological uncertainty on energy system design by examining the distribution of optimal PV capacities identified through multiple optimization runs, each using a unique stochastically generated PV generation profile. The box plot reveals the central tendency and variability in capacity decisions, while the swarm plot specifically highlights how individual iterations, each representing a plausible weather year, lead to distinct optimization outcomes. The points clustered closely together indicate weather scenarios that result in similar capacity choices, whereas outliers represent years with exceptional meteorological conditions that drive significantly different system sizing. This visualization demonstrates that relying on a single typical meteorological year fails to capture the substantial design variance induced by inter-annual weather variability, potentially leading to either under-sized systems vulnerable to energy shortfalls or over-sized systems with poor economic returns.

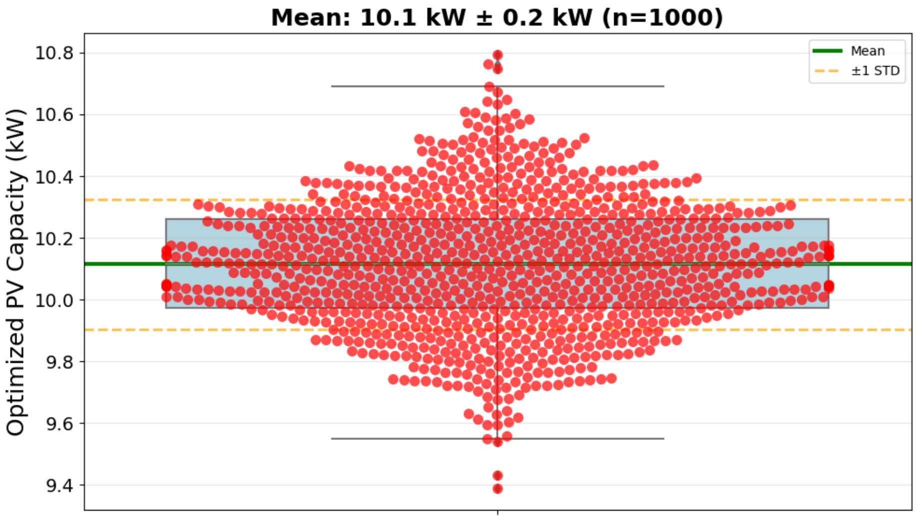

Figure 9: Distribution of Optimal PV Capacity across different PV profiles

3.1. Analysis of Capacity Drivers

To determine whether optimal PV capacity variations stem from dynamic temporal patterns or simply difference in annual energy yield, a correlation analysis between optimized capacities and PV profile full-load hours (a direct measure of annual productivity) was conducted. The analysis revealed a statistically insignificant correlation ($p \geq 0.05$) with a near-zero coefficient, demonstrating that capacity decisions are driven primarily by profile dynamics rather than total annual generation. The full load hours varies from 864 h to 968 h with a mean value of 921 h for the simulated location. Observing scenarios in figure 9 reinforce this finding where identical full-load hours produced different optimal capacities, and conversely, where similar capacities emerged from profiles with substantially different annual yields. Thus indicating that the timing and variability of generation (captured through Markov chain - ML approach) are the dominant factors influencing system sizing rather than simple annual sum metrics.

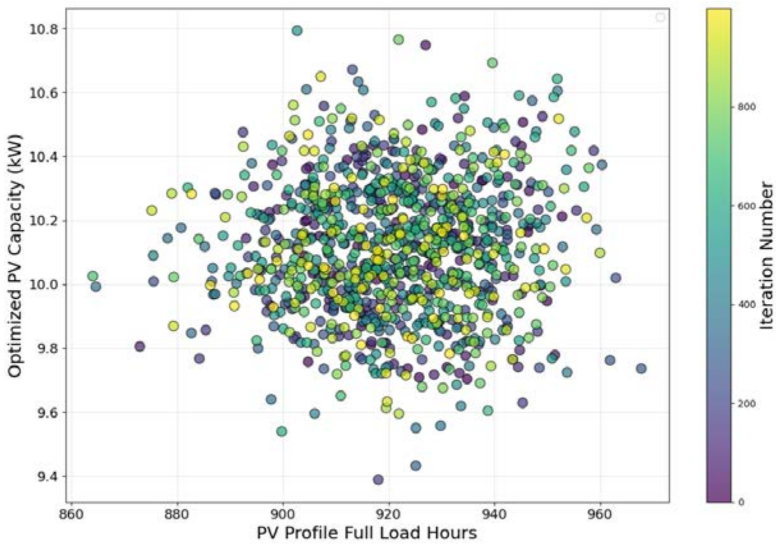

Figure 10: Optimal PV capacity vs PV profile productivity (full load hours)

The analysis reveals remarkably stable optimal PV capacity sizing with a mean of 10.1 kW and minimal variability (standard deviation: 0.21 kW, coefficient of variation: 2.09 %) across 1000 meteorological scenarios, indicating for this small household system, inter-annual weather variability has limited impact on optimal capacity decisions. However, this apparent stability masks critical implications for larger energy systems, where such modest relative variations translate to substantial absolute capacity differences. A 2.09 % coefficient of variation in a 100 MW community solar farm represents 2.09 MW of capacity uncertainty, while in a 10 GW regional system it equates to 209 MW of potential over- or under-investment. Furthermore, the observed 1.4 kW range (9.39 – 10.79 kW) demonstrates that even rare meteorological extremes can drive capacity requirements 7 % above or below the mean, highlighting how stochastic optimization captures low-probability high-impact events that traditional single-scenario approaches miss, ensuring system resilience across the full spectrum of possible weather conditions rather than just average patterns.

4. Summary

This paper addresses a critical limitation in conventional energy system modeling (the reliance on deterministic Typical Meteorological Year (TMY) data) by developing a novel Markov Chain - Machine Learning hybrid methodology for generating stochastic photovoltaic (PV) generation scenarios. The approach processes 19 years of historical irradiation data through a Clear-Sky Index normalization and state discretization framework, then employs monthly-specific Markov models with entropy-optimized orders to capture season-dependent transition probabilities. Enhanced with Random Forest classification and physical constraint enforcement, the method generates multiple equiprobable, physically realistic PV generation profiles that maintain historical consistency while exploring plausible meteorological variations. Applied to a German case study and optimized using the Open Energy Modeling Framework (oemof), the results demonstrate that optimal PV capacity is driven by temporal generation patterns, not just annual energy yield. The findings highlight the importance of incorporating uncertainty through dynamic time series for robust renewable energy system design, ensuring reliable performance across diverse meteorological conditions.

5. Acknowledgments

This work is a part of our main project Zero Carbon Cross Emission (ZORRO II) and supported by the Thuringian Ministry of Environment, Energy, Nature Conservation and Forestry.

6. References

Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32. https://doi.org/10.1023/A:1010933404324

Erbs, D. G., Klein, S. A., & Duffie, J. A. (1982). Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. Solar Energy, 28(4), 293–302. https://doi.org/10.1016/0038-092X(82)90302-4

Hilpert, S. et al., (2018). The Open Energy Modelling Framework (oemof) - A new approach to facilitate open science in energy system modelling, Energy Strategy Reviews, Volume 22, Pages 16-25, https://doi.org/10.1016/j.esr.2018.07.001.

Ineichen, P., Perez, R. (2002). A new airmass independent formulation for the linke turbidity coefficient, Solar Energy Vol. 73, No. 3, pp. 151-157, https://doi.org/10.1016/S0038-092X(02)00045-2

Institut für Regenerativ Energietechnik (in.RET) (2025). EnSys, https://ensys.hs-nordhausen.de/, accessed on: 16.10.2025

Kevin S. Anderson, C. W. (2023). pvlib python: 2023 project update. The Journal of Open Source Software

Krien, U. et al., (2020). oemof.solph—A model generator for linear and mixed-integer linear optimisation of energy systems, Software Impacts, Volume 6, https://doi.org/10.1016/j.simpa.2020.100028.

REN21. (2021). Renewables 2021 Gloabl Status Report. Paris: REN21 Secretariat.

Reiner Lemoine Institut (RLI) et al., (2025). OpenPlan, https://open-plan-tool.org/de/, accessed on: 16.10.2025.

Shalizi, C.R., Crutchfield, J.P. (2001). Computational Mechanics: Pattern and Prediction, Structure and Simplicity. Journal of Statistical Physics 104, 817–879. https://doi.org/10.1023/A:1010388907793

Thomas Huld, r. M. (2012). A new solar radiation database for estimation PV performance in Europe and Africa. Solar Energy, Volume 86, Issue 6. doi:https://doi.org/10.1016/j.solener.2012.0