NOVA IMS Information Management School

MDSAA

Master of Science

Data Science and Advanced Analytics

Effective Use of Numerical Weather Predictions for Solar Energy Forecasting

Local vs. Global Models

Felix Sami Gaber

Project Work

presented as a partial requirement for obtaining the Master's degree in
Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

EFFECTIVE USE OF NUMERICAL WEATHER PREDICTIONS FOR SOLAR FORECASTING: LOCAL VS. GLOBAL MODELS

by

Felix Sami Gaber

Project Work presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervisor: Fernando Bação, PhD

July, 2024

DECLARATION OF INTEGRITY

I declare that I have carried out this academic work with integrity. I confirm that I have not resorted to plagiarism or any other form of misuse of information or falsification of results during the process of preparing this work. I also declare that I am aware of NOVA Information Management School's Rules of Conduct and Code of Honor.

Stuttgart, July 14, 2024

ACKNOWLEDGEMENTS (OPTIONAL)

I would like to express my deepest gratitude to Daniel Lassahn, CTO and Founder of Alitiq, for providing me the opportunity to write this thesis in collaboration with Alitiq. Working together to brainstorm and solve problems has been immensely enjoyable and rewarding.

Furthermore, I am sincerely thankful to Professor Fernando Bação, PhD, from the NOVA Information Management School in Lisbon, for his academic guidance and supervision. His valuable insights and feedback have greatly contributed to the depth and quality of this thesis.

ABSTRACT

Accurate solar power forecasting is crucial for the integration of renewable energy into the grid, addressing challenges of data scarcity in new installations and the scalability of forecasting models. This study examines the effectiveness of global versus local models in predicting solar power generation, utilizing hybrid models that integrate machine learning and physical models. Findings indicate that simpler models, such as linear regression, often outperform complex models under significant weather forecast uncertainties, suggesting that complex models may overfit to noisy data. The research highlights that global models, trained on Numerical Weather Prediction (NWP) data from Germany, generalize well to new installations and outperform local models in scenarios with limited historical data. However, local models exceed global models' performance when sufficient historical data is available, with a critical threshold identified at approximately 130-140 days of historical data. Beyond this threshold, local models continuously improve, providing a guideline for transitioning from global to local models. The study also underscores that data from the UK is unsuitable for predicting German solar power output due to differences in data quality and feature-target relationships. This work contributes to the existing body of knowledge by developing and validating global models based on NWP data and quantifying the necessary historical data for effective local modeling. These insights aim to enhance the scalability and efficiency of solar power forecasting, addressing both data scarcity in new installations and logistical challenges in maintaining numerous bespoke models.

KEY WORDS

Local-global Model; Solar Energy Forecasting

Sustainable Development Goals (SDGs):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
API	Application Programming Interface
ARIMA	AutoRegressive Integrated Moving Average
ARMA	AutoRegressive Moving Average
AWS	Amazon Web Services
CatBoost	Categorical Boosting
CNN	Convolutional Neural Network
DE	Germany
DNN	Deep Neural Network
DWD	Deutscher Wetterdienst (German Weather Service)
EoT	Equation of Time
FCMR	Federal Climate Research
GB	Gradient Boosting
GHI	Global Horizontal Irradiance
GPkN	Global Plant kNear
GPH	Global Plant Holdout
ICON-EU	Icosahedral Nonhydrostatic Model, weather forecast model of the German Weather Service
KNN	k-Nearest Neighbor
LightGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
ML	Machine Learning
NWP	Numerical Weather Prediction
OCF	Open Climate Fix, a London-based non-profit product lab dedicated entirely to reducing greenhouse gas emissions
ORM	Object-Relational Mapper

PatchTST	Patch-based Transformer for Sequential Tasks
PHANN	Physical Hybrid Artificial Neural Network
PV	Photovoltaic
PV System	Photovoltaic System
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SQL	Structured Query Language
TFT	Temporal Fusion Transformer
UK	United Kingdom

1. INTRODUCTION

Solar energy plays a critical role in the transition towards sustainable energy, significantly contributing to the reduction of fossil fuel dependency and the mitigation of climate change. As a clean and renewable source of power, solar energy has the potential to meet a substantial portion of the world's energy demands. However, the integration of solar power into the energy grid poses significant challenges due to the variable nature of solar irradiance, which is influenced by dynamic local weather patterns.

Accurate forecasting of solar power generation is essential for efficient energy management and grid integration. It enables grid operators to balance supply and demand, optimize the use of energy storage systems, and maintain grid stability. Despite its importance, predicting solar output across different sites remains a complex task. Site-specific conditions such as geographical location, weather variability, and system characteristics contribute to the difficulty of accurate forecasting (Iheanetu, 2022).

The current industry practice involves developing bespoke models for each solar installation. While this approach can provide high accuracy for individual sites, it presents two major issues: data scarcity in new installations and scalability challenges. New solar installations lack historical operational data, making it difficult to train robust predictive models. This data scarcity leads to less accurate forecasts, impacting strategic planning and the integration of these new assets into the energy infrastructure. Additionally, custom-tailored forecasting models for each installation create scalability issues. Managing and updating thousands of unique models strains computational resources and operational workflows, reducing the efficiency of large-scale solar energy operations (Lago et al., 2018).

To address these challenges, solutions that generalize well to unseen installations are needed. Global models, which are trained on data from multiple sites, offer a promising approach. These models can overcome data scarcity by leveraging information from existing installations and ensure scalability by reducing the need for bespoke models (Cargan et al., 2023). However, the performance of global models compared to local models, which are trained on data from individual sites, under varying conditions remains an area of active research.

The main goal of this thesis is to develop and validate models that generalize well across different solar installations, particularly those without historical data. The aim is to create global, scalable, and robust predictive models that address issues of data scarcity and logistical bottlenecks present in current practices. Additionally, we seek to provide guidelines on when to prefer local models over global ones based on specific conditions. By comparing different levels of model globality (international, national, regional) to local models in various setups, we aim to determine under which conditions (e.g., the amount of training data) global models offer advantages. This approach will provide insights into when to transition from global to local models based on the availability of training data for new photovoltaic (PV) systems.

In summary, this study contributes to the existing body of knowledge by developing and validating global models based on Numerical Weather Prediction (NWP) data, quantifying the necessary historical data for effective local modeling, and investigating the potential of available data to mitigate weather forecast errors. These insights aim to enhance the scalability and efficiency of solar power forecasting, addressing both data scarcity in new installations and logistical challenges in maintaining numerous bespoke models.

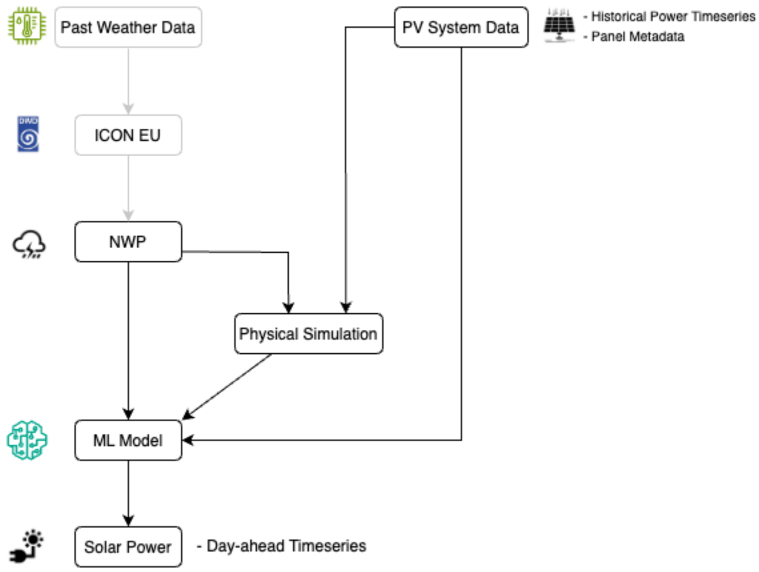

Figure 1-1 Data flow for predicting solar energy production. This figure demonstrates how weather data and photovoltaic (PV) system data, including metadata and time series performance data, are utilized to make predictions. Weather data are processed through the ICON EU model of the German Weather Service, resulting in numerical weather predictions. These predictions, along with PV system data, feed into both a physical simulation and a machine learning model. The output from the physical simulation serves as a first guess and is also used as a feature in the machine learning model. The combined model produces a one-day-ahead forecast with a 15-minute granularity.

This thesis is organized into seven main chapters, each dealing with specific aspects of the research work. The current Chapter 1 introduces the research topic and outlines the importance of accurate solar power forecasting and the challenges associated with current practice. The subsequent Chapter 2 provides the background information needed to understand the context of the study, including the development of solar forecasting methods and various forecasting techniques. Chapter 3 provides a comprehensive literature review that examines existing studies on local and global models for solar power forecasting and identifies gaps that this study aims to fill. Chapter 4 describes the methodology used in this study, including data collection, pre-processing, feature development, and experimental design. Chapter 5 discusses the results of the experiments and provides an in-depth analysis of the findings and their implications for the effectiveness of local and global models. Chapter 6 concludes the thesis by summarizing the key findings from the research, highlighting the practical applications of the results. Finally, Chapter 7 addresses the limitations of the study and suggests areas for further investigation to improve the accuracy and scalability of solar power forecasting models.

2. BACKGROUND

As the global energy transition progresses, the importance of renewable energy sources, particularly solar energy, is increasingly coming to the fore. Solar energy is one of the most abundant and sustainable sources of power, playing a crucial role in reducing greenhouse gas emissions and dependency on fossil fuels. However, this shift towards renewable energy brings new challenges for energy management, particularly due to the variable and locally differing weather conditions that make it difficult to predict solar output accurately.

Accurate forecasting of energy production from photovoltaic (PV) systems is essential for efficient control of the energy flow and for strategic energy supply planning. Precise predictions enable grid operators to balance supply and demand effectively, optimize the use of energy storage systems, and maintain grid stability. Despite its importance, predicting solar output across different sites remains a complex task. Site-specific conditions such as geographical location, weather variability, and system characteristics contribute to the difficulty of accurate forecasting.

In industrial practice, individually tailored models are frequently utilized for specific systems. While these models provide a high level of precision, they are encountering limitations in scalability and efficiency in a rapidly expanding market that continues to introduce a multitude of new systems. Custom models, though accurate for individual sites, require significant resources for development and maintenance, which becomes impractical as the number of installations grows. Moreover, it is challenging to make predictions for newly installed systems without any historical data, leading to less accurate forecasts and impacting strategic planning and integration of these new assets into the energy infrastructure.

In academic circles, studies such as those by Nie et al. (2022) and Cargan et al. (2023) have started to explore the distinctions and potential advantages of local versus global modeling approaches in forecasting solar power. Local models are trained on data from specific installations and can capture site-specific characteristics effectively. However, they require extensive historical data, which is often not available for new installations. On the other hand, global models are trained on data from multiple installations across different locations. These models can generalize well to new sites, addressing the issue of data scarcity, but may miss site-specific details that local models capture.

This part of the thesis introduces the foundations and classifications of solar power forecasting. Understanding the classification of different prediction methods for solar power is imperative to comprehend the framework within which this research is situated. The classification illuminates the various approaches developed over time to address the complexities of solar energy generation forecasting. These approaches range from physical models based on meteorological data and solar radiation principles to statistical methods that leverage historical data patterns to advanced machine learning techniques that can handle complex, non-linear relationships in the data.

The classifications of the mainstream literature are presented below, while the distinction between local and global models is discussed in more detail in the following Literature Review. This review will highlight the strengths and weaknesses of different modeling approaches, the evolution of forecasting techniques, and the current state of research in the field. By examining the literature, this study aims to identify gaps and opportunities for improving solar power forecasting, particularly through the development and validation of models that can effectively generalize across different installations while addressing the limitations of current practices.

2.1. EVOLUTIONARY PERSPECTIVES ON SOLAR PREDICTION METHOD

The evolution of solar power forecasting methods reflects the solar industry's response to the complexity of harnessing sunlight for electricity. Initially, the field relied on physical models, which were directly based on meteorological data and the principles of solar radiation conversion. These methods, detailed and accurate within their design parameters, were quite dependable in stable weather conditions. However, as Iheanetu (2022) pointed out, their effectiveness diminished when faced with unpredictable weather and varying geographical locations.

Table 2-1: Evolution of Solar Power Forecasting

Stage	Improvements	Downsides
Physical Methods	Accurate within design parameters, based on meteorological data and solar conversion principles	Performance declines under variable weather and different geographic locations
Statistical Methods	Greater adaptability, utilizes historical data to identify patterns for future output	Dependence on stationary data and difficulties with non-linear relationships
Machine Learning	Improved accuracy in feature extraction and understanding temporal dynamics	Requires large datasets and computational resources
Hybrid Models	Optimizes features of component methods, synthesizing various predictive methodologies	Complexity increases, challenging to implement and maintain, higher computational load
Transformer-based Methods	Exceptional ability to model complex relationships, parallel computing	Potential overfitting and interpretability issues, high computational cost for training

To address the limitations of physical models, statistical approaches like ARMA and ARIMA were introduced. These methods, as Cherkassky & Ma (2004)described, leveraged historical data to discern patterns and predict future outputs, offering adaptability to a wider range of conditions. Yet, their reliance on stationary data and the challenge of non-linear data relationships often restricted their forecasting accuracy, especially when abrupt weather changes occurred.

The field took a significant leap forward with the advent of machine learning, particularly with deep learning techniques. Yona et al. (2013), and Chen et al. (2013) have underscored the capabilities of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to extract features from complex datasets and comprehend temporal dynamics, greatly enhancing prediction accuracy. These methods have shown promise in addressing the irregularities and uncertainties inherent in solar power generation. This era of forecasting also saw the incorporation of Long Short-Term Memory

(LSTM) networks, and which brought significant improvements in capturing the temporal and sequential patterns in solar irradiance data. Researchers like Goodfellow et al. (2016) have demonstrated how LSTMs are particularly well-suited for modeling time-series data, which is integral in predicting the intermittent nature of solar power generation.

Building upon the progress afforded by machine learning, the solar power forecasting field has witnessed the creation of hybrid models, an innovative synthesis of physical, statistical, and machine learning techniques. These models harness the collective strengths of diverse methodologies to enhance prediction accuracy and reliability. For instance, Fonseca Junior et al. (2015) and Umer et al. (2019) have successfully integrated Artificial Neural Networks (ANNs) with physical forecasting methods. Their hybrid physical ANN (PHANN) models simulate the sky's theoretical model for specific locations using clear sky solar radiation models, optimizing the feature selection process, and enhancing forecast reliability.

Moreover, transformer models, as conceptualized by Vaswani et al. (2017), have revolutionized the field with their ability to handle sequential data without the constraints of sequential processing, allowing for parallelization and significantly faster computation. These advancements have not only enhanced the accuracy of predictions but also opened new possibilities for handling the vast and complex datasets associated with solar power forecasting. In "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," Lim et al. (2020) introduce a novel deep learning model that excels in integrating various data types for enhanced forecasting. This Temporal Fusion Transformer (TFT) model is adept at processing and synthesizing multiple data streams—ranging from static and known future inputs to historical sequences and time-varying covariates. The authors showcase the TFT's robustness in handling complex, real-world datasets where these diverse data types are pivotal in improving prediction accuracy and interpretability. In a practical exploration of this model, Santos et al. (2022) apply the TFT to solar power predictions, demonstrating its effectiveness in this specific domain. Their work, titled "Application of Temporal Fusion Transformer for Day-Ahead PV Power Forecasting," evidences the TFT's superior performance in forecasting solar energy production. By leveraging the model's capabilities to interpret various influential features, they provide valuable insights into the dynamics of solar power generation, affirming the model's operational success in accurate day-ahead forecasting.

2.2. FURTHER CLASSIFICATIONS AND FRAMEWORKS

In the dynamic field of solar energy forecasting, a comprehensive understanding of various prediction methodologies is crucial. The evolution of these methodologies reflects the industry's response to the inherent complexities of predicting solar power generation. Over time, a diverse array of techniques has been developed, each with its own strengths and applications, catering to different aspects of solar forecasting.

This section aims to delve deeper into the classifications and frameworks that underpin solar power forecasting methods. By categorizing these approaches, we can better understand their unique characteristics and the contexts in which they excel. Such classifications not only illuminate the historical progression of solar forecasting techniques but also provide a structured overview of the current state-of-the-art methodologies.

The classifications include distinctions between direct and indirect prediction techniques, various forecasting horizons, and the types of data sources utilized in solar forecasting. Understanding these classifications is essential for grasping how different methods address the challenges posed by the variable nature of solar irradiance and the integration of solar power into the energy grid.

2.2.1. DIRECT VS. INDIRECT PREDICTION TECHNIQUES

Direct forecasting techniques forecast PV power directly from historical data. This method relies on analyzing past PV performance data to identify patterns and relationships that can be used for future performance forecasts (Chen et al., 2013). The direct method is simple and correlates directly with PV performance but requires extensive historical data to be effective. Indirect prediction techniques involve a two-step process. First, they model weather factors such as solar irradiance, which significantly impact PV output. This stage is crucial as solar irradiation is a primary driver of PV power generation. Subsequently, the derived data is converted into predicted PV power. Indirect methods incorporate additional data layers, such as meteorological or weather forecast information, to estimate PV power (Jiang et al., 2023, Gaboitaolelwe et al., 2023). This approach is particularly useful in scenarios where historical PV power data is scarce or unavailable, such as new PV installations. Research has demonstrated that indirect methods can provide more accurate forecasts due to their comprehensive approach, which considers various influencing factors. The method takes into account site-specific variables and losses to predict energy yield values, providing a more nuanced and detailed forecast (Iheanetu, 2022). This approach is especially useful for new PV systems or when weather forecast data has other applications (Gaboitaolelwe et al., 2023).

2.2.2. FORECASTING HORIZONS

In research on predicting photovoltaic (PV) performance, there are varying views and definitions regarding forecast horizons. Although research groups use different classifications, a widely accepted classification has emerged and is frequently used in literature. Iheanetu (2022) classified forecasting horizons into four categories: very short-term, short-term, medium-term, and long-term. It is important to note that this classification may be interpreted differently by different researchers. Examples include Raza et al. (2016), Nespoli et al. (2019), and Sobri et al. (2018).

Table 2-2: Overview of Different Forecast Horizons, Duration, and Applications

Forecast Horizon	Period	Application
Very short-term	<1 min	Controlling Power Distribution
Short-term	1 – many hours	Ensures commitment, scheduling, and dispatch
Medium-term	1 week - 1 month	Smooths the planning of power system and maintenance schedule
Long-term	1 month – 1 year	Helpful for power generation transmission and distribution

Very short-term forecasts cover a period of a few seconds to less than an hour. Ren et al. (2015) and Das et al. (2018) mention the use of these forecasts in electricity marketing and pricing, power smoothing, and real-time power distribution control. Short-term forecasts refer to a period of one to several days. Short-term forecasts are important for economic load balancing and operation of electricity systems, particularly for renewable energies. Medium-term forecasts, which cover a period of one week to one month, are crucial for planning and maintenance of conventional or solar-integrated power systems. Long-term forecasts extend over a period of one month to one year and are crucial for power generation, transmission, distribution, and solar energy planning (Das et al., 2018; Behera et al., 2018). The accuracy of photovoltaic power forecasting (PVPF) models depends significantly on the length of the forecasting horizon. Studies by Lonij et al. (2013) and (Lipperheide et al., 2015) show that forecast accuracy decreases as the horizon lengthens. For example, a PVPF model with a 15 to 90 minute horizon showed varying accuracy levels. The prediction error increased significantly from 3.2% to 15.5% as the forecast horizon increased from 20 to 180 seconds. Monjoly et al. (2019) confirm that also the normalized root mean square error (nRMSE) also increased with the duration of the forecast horizon, regardless of the model type used. These results highlight the significance of choosing an appropriate forecast model based on the forecast horizon, taking into account factors such as cloud movement and other atmospheric conditions that may impact accuracy (Wang et al., 2020; Shi et al., 2012; Mellit et al., 2014; Kühnert, 2016).

2.2.3. Data Source

Accurate forecasting of solar energy production relies on diverse data sources that capture varying aspects of weather and atmospheric conditions. Three primary data sources are widely recognized in the field: ground-based sky images, satellite images, and numerical weather predictions (NWPs). Each source offers unique advantages and collectively contributes to a comprehensive understanding of solar irradiance.

Ground-based sky images involve using cameras positioned on the ground to capture high-resolution images of the sky at regular intervals. By analyzing these images, it is possible to detect cloud cover, cloud movement, and other atmospheric conditions that influence sunlight availability. The significant advantage of ground-based sky images is their ability to provide high resolution and accuracy in local predictions, capturing microclimate variations that might be missed by satellite images (Nie, Li, et al., 2022).

Satellite images, on the other hand, offer a broader perspective on weather patterns and cloud cover over extensive areas. These images are crucial for tracking the movement of cloud formations and predicting their impact on solar irradiance across different regions. Although satellite images may have lower resolution compared to ground-based images, they provide a macroscopic view, making them valuable for regional solar forecasts over large geographical areas (Cargan et al., 2023).

Numerical Weather Predictions (NWPs) are advanced computer models that simulate the atmosphere using complex equations that govern the physics of weather. The ICON-EU model, employed by the German Weather Service, is an example of such a model. It uses various meteorological data inputs to forecast future weather conditions, including temperature, wind, humidity, and cloud cover. The ICON-EU model is a mesoscale forecast system that provides detailed predictions for Europe, utilizing a complex grid structure to simulate the atmosphere accurately. NWPs consider a wide range of variables, making them powerful tools for predicting solar irradiance several days in advance. By integrating ground-based observations, satellite data, and predictions from models like ICON-EU, a comprehensive and accurate solar energy production forecast can be achieved (Heppelmann et al., 2017).

In this study, we primarily utilize numerical weather predictions from the ICON-EU model to generate solar energy forecasts. This choice is driven by the model's ability to provide detailed and accurate weather forecasts, which are critical for predicting solar irradiance and, subsequently, solar power generation. The use of NWPs allows for a robust and scalable approach to forecasting, addressing the challenges of data scarcity and variability in new and existing solar installations.

2.3. Incorporation of this Work

This work employs a combination of machine learning and physical models, making it appropriate to describe the models used herein as hybrid models. Unlike traditional time series forecasting, our approach does not utilize sequential input; instead, we match the NWP data and plant-specific data at a given time directly with an output, namely, the power generation.

The ICON-EU model (Icosahedral Nonhydrostatic Model) of the German Weather Service (DWD) is a numerical weather prediction model based on physical laws and mathematical equations to simulate the atmosphere and its processes. In addition to using these forecasts, this work also simulates solar power generation using physical laws and incorporates these simulations as an additional feature for our machine learning models.

We operate within the realm of indirect prediction methods by utilizing the weather forecasts provided by the German Weather Service. Our focus is on one-day-ahead forecasting, categorizing our work within short-term forecasting. Additionally, our data source is the Numerical Weather Predictions (NWPs), as our weather forecast data is derived from these predictions.

In essence, this work integrates both machine learning and physical models to enhance the accuracy and reliability of solar power generation forecasts. By leveraging advanced machine learning techniques alongside the robust physical models of the ICON-EU system, our approach enables more nuanced and precise predictions, effectively addressing the inherent variability and unpredictability in solar power generation.

3. LITERATURE REVIEW

Accurate solar power forecasting is crucial for integrating renewable energy into the grid, balancing supply and demand, and optimizing energy storage. This review examines advancements in forecasting, focusing on local and global models. Local models offer detailed, site-specific forecasts but face scalability issues due to their reliance on extensive historical data. In contrast, global models utilize large, multi-location datasets, providing scalability and robustness but potentially lacking precision for individual sites.

Recent studies highlight the benefits of combining diverse data sources, like satellite imagery and ground-based sky images, to enhance forecast accuracy. This review synthesizes methodologies and findings, identifying strengths, limitations, and research gaps. By comparing local and global models, we aim to outline where each excels and propose areas for further investigation.

The review covers the following: definitions and distinctions of local and global models, current forecasting methodologies, the role of additional data sources, and identification of research gaps addressed by our study.

3.1. Definition: Local and Global Models

Local models are precisely trained on individual (Montero-Manso & Hyndman, 2021) or small groups of organized solar installations (Nie et al., 2022) and are therefore tailored to the specific requirements of these individual installations. While they are excellent for detailed, site-specific analysis, their application can be limited when it comes to making general predictions for the entire data set. An issue that arises with local models is the limited availability of historical data, especially for new plants. The lack of historical operating data makes it difficult to develop reliable forecasting models. This problem is particularly relevant for the prediction of solar power, as there is no past data available for new solar plants that would allow an accurate prediction of future performance. Another disadvantage of local models is their limited scalability and logistical bottlenecks. Creating a separate model for each solar power plant is no longer a viable standard procedure as the number of plants increases. This approach leads to logistical complications, as maintaining and updating numerous individual models requires significant computing resources and complicates operational processes. Consequently, this can significantly affect the efficiency of large solar power plants.

In contrast, global models aim to capture the entire data set or a large part of it with a single model. They identify general patterns and trends that apply to the entire dataset, making them more robust and less prone to overfitting to small data segments. Global models attempt to capture underlying structures or relationships that are consistent across different contexts. However, global models may not capture the details or nuances of local data patterns that are present in local models. While global models are intended to overcome the limitations of local models by capturing overarching patterns in a large data set, accuracy for specific assets must not be compromised due to site-specific differences between assets.

3.2. CURRENT APPROACHES OF GLOBAL MODELS

There is not much literature on the comparison of local or global models for Numerical Weather Predictions (NWP) of solar power. However, there is similar literature from the field of solar irradiance forecasting using satellite data and NWPs (Lago et al., 2018), as well as solar power forecasting from satellite images (Cargan et al., 2023) and sky- image solar power forecasting (Nie et al., 2023). These are reported in the following three sections:

The work “Short-term solar irradiance prediction without local telemetry: a generalized model with satellite data” by Lago et al. (2018) addresses the challenge of predicting solar irradiance without relying on local ground measurements like historical PV performance data. The authors propose a generalized model that uses satellite-based measurements, NWPs, as well as humidity and temperature forecasts for the short-term prediction of solar irradiance. The DNN model is trained using data from a limited number of locations where ground measurements are available and can be easily transferred to other locations without such data. The study shows that this generalized model performs comparable or even better than local models, achieving a relative root mean square error (rRMSE) of 31.31%, compared to 32.01% for the best local model. This approach highlights the potential for scalable and efficient solar forecasting across different geographical locations and emphasises the importance of satellite data for improving forecast accuracy.

In the final review of the paper by Cargan et al. (2023) entitled "Local-Global Methods for Generalized Solar Irradiance Forecasting", the authors explore in depth different techniques to build prediction models by applying standardized machine learning methods such as Random Forests (RFs), Deep Neural Networks (DNNs), Long Short-Term Memory Networks (LSTMs) and Convolutional Neural Networks (CNNs). These models were trained using four approaches: local, global, Global Plant Holdout (GPH) and Global Plant kNear (GPkN), each with its own specific methodology and application. The study emphasizes that global approaches and their extensions are superior compared to local approaches. Despite the higher computational cost of training, global models that combine data from all locations consistently perform better predictions on solar irradiance. In particular, the authors show how global models can enable forecasts for locations with limited historical data. Interestingly, the research demonstrates that CNNs using satellite imagery are superior to other models, especially in scenarios where data from nearby locations must be used as a proxy. This emphasizes the importance of richer weather data for simple point-based data for improving forecast accuracy. However, they point out that further research is needed, especially on the effect of more comprehensive weather data on the performance of the models. Future work will aim to increase the temporal resolution to better capture intra-hourly fluctuations. In addition, the researchers plan to expand the data substitution approach by combining data from multiple locations instead of relying only on the closest location. These findings offer valuable perspectives for the further development of precise forecasting models for the solar energy industry and underline the importance of innovative approaches in renewable energy research.

Furthermore, similar research in the field of solar energy forecasting based on sky image data offers further valuable insights and methods that can be adapted for further development in this area: In the study by Nie et al. (2023), titled "Sky image-based solar forecasting using deep learning with multi-location data: training models locally, globally or via transfer learning?", the development of deep learning-based models for solar power forecasting is investigated. These models use heterogeneous data sets with different weather conditions from different parts of the world. The researchers compare the performance of locally trained models based on single datasets with globally trained models using the combination of multiple datasets. In addition, the potential of knowledge transfer from pre-trained models to a new dataset of interest (transfer learning models) is evaluated. The results of the study suggest that locally trained models perform best in their local context, while significant prediction errors can occur when used outside their training context. Nevertheless, trends and patterns of the PV/radiation time series can be predicted well, suggesting that the feature representations learned by local models are generally transferable, although the learned regression functions remain location-dependent. Training on a global dataset improves the generalizability of the model, as evidenced by better performance in fitting to individual locations, although this may involve increased training effort. Transfer learning proves to be particularly helpful with limited training data, especially when a source model has been pre-trained on a large and diversified dataset. By pre-training solar forecast models on a global dataset and transferring them to the local DEWA dataset, the training effort can be reduced by 80% while improving forecast accuracy by 1% compared to the locally trained base model. The authors call on the research community to contribute to a comprehensive global dataset for solar forecasting that contains massive and diversified data. Providing open-source models based on such extensive datasets could greatly simplify the development of local models and significantly reduce training efforts. Future studies should investigate the benefits of data augmentation (e.g. image blending, rotations, vertical/horizontal flips) and scene representation (e.g. polar coordinates, sun-centered sky images) for knowledge transfer in more diverse contexts. These findings add to the understanding of the effectiveness of local, global and transfer learning models in solar power forecasting and highlight the potential of deep learning approaches in utilizing diversified datasets to improve forecast accuracy.

3.3. RESEARCH GAP

The analysis of existing literature reveals several key insights regarding the use of local and global models for solar power forecasting. Local models, known for their site-specific accuracy, face significant scalability challenges and require extensive historical data. In contrast, global models, despite being computationally intensive, demonstrate robust performance by identifying overarching patterns across multiple sites. Studies indicate that global models, particularly when enriched with comprehensive weather data such as satellite imagery, significantly outperform local models in scenarios with limited historical data.

However, several research gaps remain. As we have seen above, there is a difference between irradiation forecasts and solar energy forecasts. While irradiance determines power generation to a certain extent, there are additional challenges with power generation that complicate the development of a global model. These additional challenges encompass varying efficiencies, shading, differing measurement methods, data quality, system degradation, inverter losses, and other factors. These can vary significantly between PV systems, complicating the development of a global model with comparable performance.

Current studies primarily focus on global models using sky imagery and satellite data. Due to data availability issues with Numerical Weather Predictions (NWPs), there is a notable lack of comprehensive research on applying global models to NWP data for solar energy forecasting. This research aims to fill this gap by leveraging an extensive NWP archive to investigate whether global models can be as effective or even superior to local models when applied to NWP data.

Furthermore, while it is acknowledged that local models require substantial historical data, the specific threshold at which it becomes beneficial to switch from global to local models remains underexplored, despite its importance for practical implementation. This study intends to quantify this threshold, providing clearer guidelines on the data requirements for local models.

As Cargan et al. (2023) state: "Future work will aim to increase the temporal resolution to better capture intra-hourly fluctuations." This study will also apply this to NWPs, interpolating all input data to 15-minute intervals if necessary. Furthermore, they emphasize the importance of richer point-based weather data for improving forecast accuracy. Nie et al. (2023) agree and call on the research community to use a comprehensive global dataset for solar forecasting that includes extensive and diverse data. This study will address this part of the research gap for NWPs by working with two extensive and rich weather data along with historical performance data:

- UK: 958 PV systems x 78,625 (time dimension) x 55 features = 4,142,751,250 (4.14 billion) observations
- DE: 30 PV systems x 33,117 (time dimension) x 55 features = 54,643,959 (54.64 million) observations

Previous research has primarily distinguished between local and global models. Besides geographic differences, numerous data-related factors could influence performance. One such factor is the different measurement methods and data processing techniques used in various datasets. Therefore, this study delves deeper by testing international, national, and regional global models against each other.

These contributions are expected to provide practical insights for the scalable and efficient integration of solar power into energy infrastructure, addressing both data scarcity in new installations and the logistical challenges of maintaining numerous bespoke models.

4. METHODOLOGY

In this section, we detail the methodological approach adopted for our study, aimed at developing and validating forecasting models for solar power generation. Our methodology is structured to systematically address the challenges associated with data scarcity and the scalability of prediction models. By integrating machine learning techniques with physical models, we aim to create robust and generalizable solutions. The following subsections outline our processes for data collection, preprocessing, feature engineering, model development, and evaluation, ensuring a comprehensive and rigorous analysis framework.

4.1. DATA

Effective forecasting relies on the quality and comprehensiveness of the data used. In this subsection, we describe the data sources utilized, the processes involved in collecting and integrating these datasets, and the preprocessing steps undertaken to prepare the data for analysis. Our datasets include historical solar power data from the UK and Germany, as well as weather forecast data from the German Weather Service (DWD). We ensure data consistency and reliability through meticulous preprocessing, including normalization, aggregation, and feature engineering, to enhance the predictive power of our models.

4.1.1. DATA SOURCES

Our three data sources consist of historical solar power data from the UK (secondary data) and Germany (primary data), as well as weather forecast data for both countries from the German Weather Service (DWD).

The primary data source from Germany spans from March 1, 2023, to February 29, 2024. It includes data from 30 photovoltaic (PV) systems, each of which may consist of multiple subsystems. The data is recorded at 15-minute intervals, with solar power output measured in megawatts (MW). The locations of these PV systems in Germany are shown in Figure 4-1, which illustrates the geographical distribution of the systems across the country.

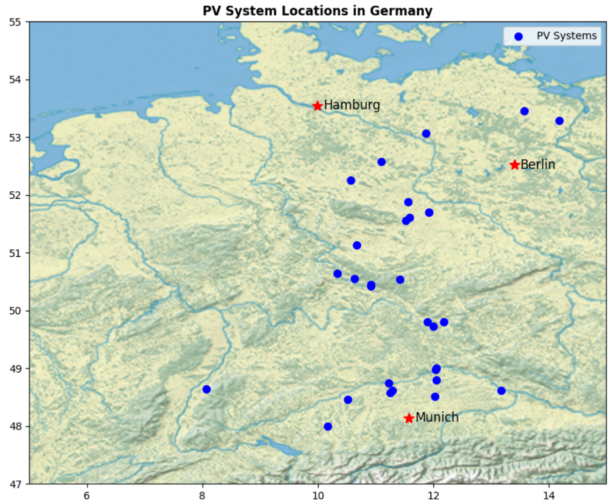

Figure 4-1 PV System Locations in Germany

Additionally, we employ a comprehensive dataset of PV generation performance provided on HuggingFace by Open Climate Fix (openclimatefix, 2022), a London-based non-profit product lab dedicated entirely to reducing greenhouse gas emissions. OCF's focus on an open and collaborative approach, rapid prototyping, and the pursuit of scalable and practical solutions drives its mission to maximize climate impact. The dataset used encompasses performance data from 1,311 PV systems in the United Kingdom, recorded from 2018 to 2021. The geographical distribution of these PV systems in the UK is depicted in Figure 4-2.

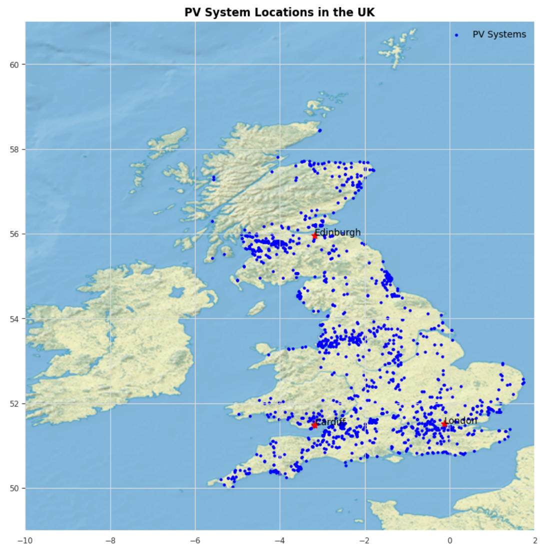

Figure 4-2 PV System Locations in UK

The dataset comprises two main components: the metadata (*metadata.csv*) and the PV performance time series data (*pv.netcdf*).

1. *Metadata (metadata.csv):* This file contains details about the various PV systems. The metadata includes columns for:

    - ss_id: The identification number of the system.
    - latitude_rounded: The latitude of the PV system, rounded to approximately the nearest kilometer.
    - longitude_rounded: The longitude of the PV system, rounded to approximately the nearest kilometer.
    - llsoacd: A variable to be determined.
    - orientation: The orientation of the PV system (azimuth).
    - tilt: The tilt of the PV system.
    - kwp: The capacity of the PV system in kilowatts peak (kWp).
    - operational_at: The date and time the PV system became operational.

It's noteworthy that the metadata contains additional PV systems that do not appear in the PV performance time series data.

2. *PV Performance Time Series Data (pv.netcdf)***:** The time series data is stored in xarray format and contains the solar generation in watts for the PV systems listed in the metadata. Each data variable corresponds to a *ss_id* from the metadata, representing a subset of all *ss_ids* listed in the metadata. The time coordinates are tagged as *datetime*, representing the date and time of the solar generation measurement with a granularity of 5 minutes.

This detailed dataset description forms the foundation for analyzing the PV performance data within our research, aiming to gain insights into the efficiency and performance of PV installations under varying geographical and operational conditions.

Weather forecast data from the regional ICON-EU model of the German Weather Service (DWD) was used to develop and validate the model presented in this thesis. The ICON-EU model is a nested, high-resolution regional model within the global ICON model, which has been in operation since July 21, 2015. There is a tightly coupled two-way interaction between the regional ICON-EU model and the global ICON, which enables a consistent and accurate weather forecast. The native model grid of the ICON-EU has a horizontal grid spacing of 6.5 km, while the output grid has a grid spacing of about 7 km (0.0625°). Vertically, the ICON-EU model is structured in 60 levels up to a height of 22.5 km. The ICON-EU weather forecasts are available for periods up to +120 hours from the four model runs at 00, 06, 12 and 18 UTC and up to +30 hours from the model runs at 03, 09, 15 and 21 UTC. For the forecast period up to +78 hours, the forecast data is available at hourly intervals, while the periods between +81 and +120 hours are covered at 3-hourly intervals. The nested ICON-EU model covers the whole of Europe, with coverage in the west and east extending far beyond the European territory and encompassing the area between 23.5°W-62.5°E and 29.5°N-70.5°N. The forecast data of the ICON-EU regional model, which are routinely distributed in standardized packages per forecast element on the free DWD Open Data Server, were used for this study. The use of this high-resolution and precise weather forecast data enables a well-founded analysis and has the potential to significantly increase the forecast accuracy of the developed model.

4.1.2. DATA COLLECTION

For managing the performance data of photovoltaic (PV) systems, a PostgreSQL database on AWS has been developed. To simplify and automate interaction with the database, we rely on SQLAlchemy, an SQL toolkit and ORM (Object-Relational Mapper) for Python. The structure of this database is detailed in Figure 4-3, ensuring the accurate and exhaustive capture of information regarding the solar PV installations. Central to this architecture is the *LocationMatchingTable*, which offers a dynamic linkage between customer details and location identifiers with an external ID, *id_location_external*. This ID is unique to each *customer* but can overlap across different customers, highlighting the necessity for a unique internal *location_id* for unambiguous association.

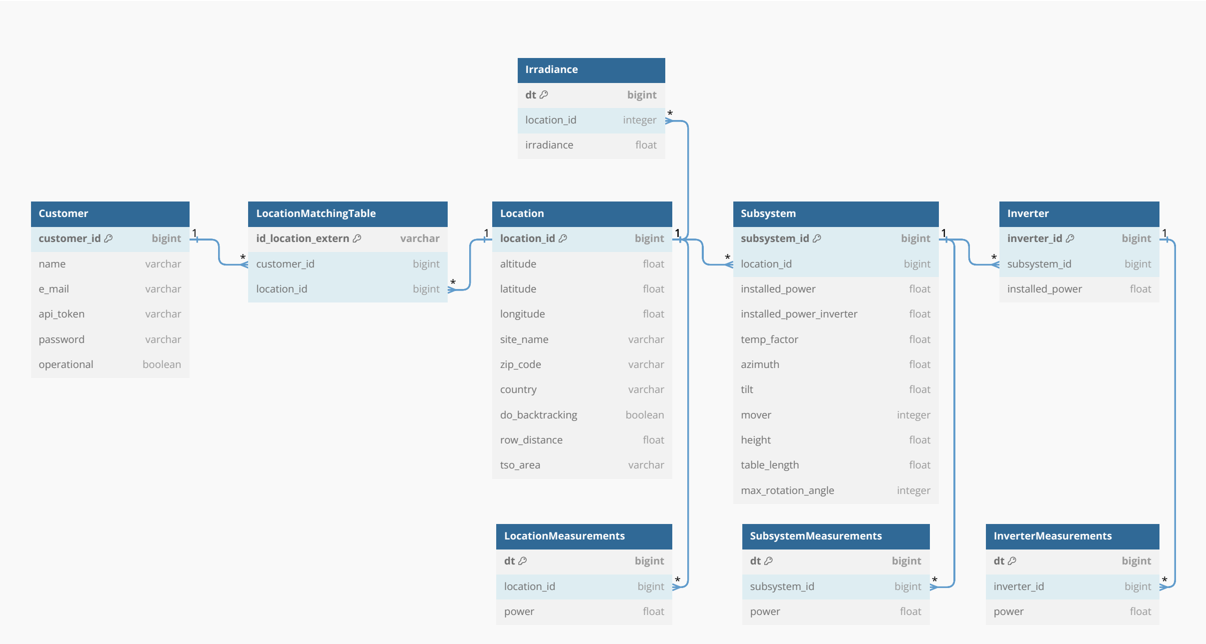

Figure 4-3 Hierarchical database structure for solar PV service management. This diagram illustrates the relational connections between customers, locations, subsystems and inverters

The database is organized in a hierarchical fashion, beginning with customer data at the top and extending through locations, subsystems, and finally to inverters. However, it should be noted that not all customers are able to supply data at the inverter level. This hierarchy, coupled with meticulous data management practices, provides the framework for a robust and adaptable database, setting the stage for in-depth analysis and enhancement procedures. Data ingress to the database is twofold: publicly accessible data is compiled and input manually, while plant-specific data is streamed directly into the database as a time series through an API. The system is designed for customer-driven integration, allowing the addition of new locations or subsystems via the API, thus promoting a seamless and efficient augmentation of the overall asset base. The API supports the normalization of time stamps to a UTC format, as well as the conversion of power data, which can be available in various units of measurement (watts, kilowatts, megawatts) and are standardized to the standard kilowatt (kW) format. To facilitate efficient data retrieval and improve query execution, indexing and partitioning techniques have been employed across the database tables:

Indexing is crucial for the acceleration of data retrieval. In the customer table, the *customer_id* serves as the primary key and is inherently indexed, allowing for swift customer record lookups. Additionally, unique indexes are created on the *e_mail* and *api_token* columns, ensuring that searches based on these fields are efficient and that the values remain unique across the dataset. For the *LocationMatchingTable*, we employ a composite primary key on the *customer_id* and *id_location_extern* columns, which enhances the performance of joint queries with other tables. The *location_id* field is also uniquely constrained and indexed, upholding a one-to-one relationship with the Location table and aiding in the integrity and speed of database operations. The *Location* table has *location_id* as its indexed primary key, facilitating rapid access and manipulation of location data. The *Subsystem* and *Inverter* tables follow a similar pattern, with *subsystem_id* and *inverter_id* respectively acting as primary keys and thus being indexed, ensuring efficient data management within their domains.

Partitioning, on the other hand, is adopted to manage large data volumes by dividing tables into more manageable segments. We've applied partitioning to the *LocationMeasurements*, *SubsystemMeasurements*, *InverterMeasurements*, and *Irradiance* tables, all partitioned by the *dt* column using a range-based approach. This partitioning enables the database to handle time-series data effectively, with improved query performance, especially when dealing with large time series data. Data is segmented into monthly intervals, making operations like insertions, deletions, and aggregations more efficient and manageable.

4.1.3. Preprocessing

In our model for analyzing photovoltaic (PV) system performance within the UK PV Project data, we've taken meticulous steps to ensure only complete datasets were included. We began by cross-referencing the system IDs in the metadata file (*metadata.csv*) with those in the time series data (*pv.netcdf*). This process resulted in a reduction from 1311 to 1310 matching PV systems. One system, identified by ID 27068, was excluded from the time series dataset due to the absence of corresponding metadata, thereby maintaining the data integrity for integration into our database. The overlapping period of data recordings from PV systems and weather forecasts in the UK was from 1 August 2019 to 27 October 2021 and was filtered accordingly (see Figure 4-4). This period was chosen for consistency across all the different sites, which in turn meant a reduction of 1310 to 958 sites, as the remaining 352 sites only included data prior to the start of this period.

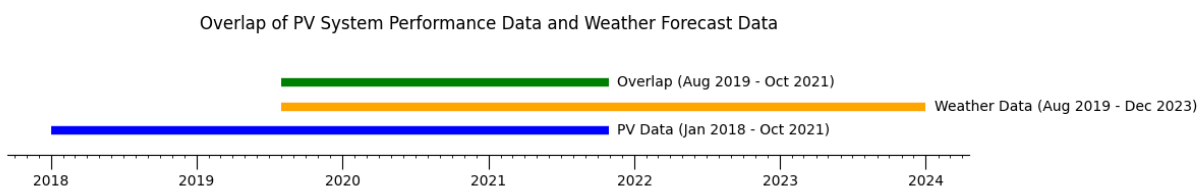

Figure 4-4: Overlap of UK historical PV System Data and Weather Forecast Data

The German solar power data spans from March 2, 2023, to February 9, 2024. Given the availability of NWP data for the entire duration, no data segmentation was required.

For the UK PV power data, we first converted the recorded energy production from watts to kilowatts to make it the same unit as the German data. Historical PV system performance data and weather forecast data had different granularity: 5 minutes for the UK data and 1 hour for the DE and NWP data. To harmonize these datasets, the UK PV performance data was aggregated, and the DE PV performance data and NWP forecast data were interpolated to a uniform granularity of 15 minutes.

In addition to the conversion and aggregation, we normalized the power values by the installed capacity of each PV system. This normalization step is critical as it ensures that comparisons of performance are fair and standardized across a range of system sizes and output capacities, allowing for a precise assessment of relative efficiency and effectiveness. According to openclimatefix (2023), the installed power is not very reliable. Values that were above 1 after normalization were therefore set to 1.

For the weather forecast using the ICON-EU model, we were able to utilize the organization's archive. As the weather forecast is generated daily and extends five days into the future, we worked with a multi-index structure. For each combination of positional index (representing a PV system) and dt_calc (the calculation time point, consistently at 00:00 hours), we had 120 observations at 1-hour intervals, providing a 5-day forecast horizon for a total of 36 features (see Table A-1). For simplification purposes, we have extracted data that covers the period between 24 and 48 hours ahead, as these reflect the one day-ahead forecasts.

4.2. Feature Engineering

Feature engineering is a fundamental step in developing effective forecasting models, as highlighted by Abdelmoula et al. (2022). This process enhances the model's ability to capture underlying patterns and relationships within the data. Given its critical importance, we perform a series of transformations to extract, create, and refine features that improve the model's predictive accuracy. In the following section, we present the specific transformations applied to our dataset, focusing on feature encoding, target encoding, and PV simulation (as a physical model), ensuring each contributes meaningfully to the forecasting process.

From the weather forecast data, we were able to generate many additional features by combining these features with each other or transforming them with physical calculations. Below is a detailed description of the specific transformations and the new features created:

Table 4-1 Description of Feature-encoded Variables

Feature	Description
apparent_temperature	Combines the wind chill index and the heat index to calculate the perceived temperature based on the current temperature, relative humidity and wind speed. The wind chill is determined for temperatures below 10°C and wind speeds above 5 km/h according to the FCMR192003 guideline (FCMR, 2003). The heat index is used when conditions are outside these limits and follows a formula from Rothfusz (NWS, 1990), based on Steadman, (1979).

cos_zenith	This feature calculates the cosine of the solar zenith angle based on geographical coordinates and time data. It uses Spencer's formula from 1971, which considers the Julian date calculation and the position of the Earth in its orbit. This calculation methodology is used to determine values relevant for the correction of bias in photovoltaic system performance data. First, the earth is positioned in its elliptical orbit around the sun and the Julian time of day is determined. The critical components of this calculation include the solar declination and the equation of time (EoT) defined by Spencer and Iqbal and described by Zekai (2008). The actual calculation of cos_zenith is done by applying the cosine function to the calculated angle converted to radians. This precision methodology is essential to ensure the accuracy of solar energy simulations and prediction models.
clear_sky_ghi	This feature calculates the global horizontal irradiance (GHI) assuming a clear sky using the SOLIS model. This model uses the position of the sun, determined by geographical coordinates and time stamps, for the prediction.
time_of_day_sine	This feature is calculated by taking the sine value of the current time in the day cycle (0 to 23 hours).
time_of_day_cosine	This feature is calculated by taking the cosine value of the current time in the daily cycle (0 to 23 hours).
yearly_sinus	This feature is calculated by taking the sine value of the current date in the yearly cycle.
yearly_cosinus	This feature is calculated by taking the cosine value of the current date in the yearly cycle.
time_of_day_numeric	This feature simply represents the numeric time from 0 to 23 hours.
hour_of_day	This feature is derived directly from the time of day (0 to 23 hours).
horizon_in_min	This feature represents the forecast horizon in minutes. It is calculated by subtracting the calculation time (dt_calc) from the prediction time (dt_fore) of an observation.

Target encoding involves creating features based on the historical values of the target variable to capture temporal patterns and trends. When performing target encoding, we ensured that only past data was used to prevent data leaks, thereby maintaining the integrity of the predictive model. This approach guarantees that no future information influences the model, which could otherwise lead to distorted results.

The following features were developed through target encoding:

- *mean_last_1_day*, *mean_last_7_days*, and *mean_last_30_days*: These features are rolling average values calculated from historical data over the past 1 day, 7 days, and 30 days, respectively.
- *std_last_1_day*, *std_last_7_days*, and *std_last_30_days*: These features represent the rolling standard deviation for the past 1 day, 7 days, and 30 days, reflecting the variability of the target variable.
- *max_last_1_day*, *max_last_7_days*, and *max_last_30_days*: These features capture the maximum values of the target variable from the given time periods, indicating the extreme values.

- *cumulative_mean_value*: This feature represents the cumulative average of all previous values of the target variable, considering the entire historical data up to the respective point.
- *value_5day_avg*: Unlike simple averages, this feature calculates the average of the target variables from the same time points over the last five days. By doing so, it better accounts for daily fluctuations and helps the model incorporate these patterns into the forecast, which is crucial for precise and relevant predictions.

These target-encoded features are essential for capturing the short-term and long-term variability and trends in the data, thereby enhancing the forecasting model's accuracy.

In addition to feature and target encoding, we decided to incorporate an additional feature based on PV simulation. This simulation feature estimates the output to be produced at a given time, functioning independently of machine learning. This approach is beneficial because some plants have information on tilt and azimuth angles, while others do not. Our simulation can handle both scenarios, indirectly incorporating these important parameters for all installations. Furthermore, it provides an initial estimate of the expected output, enhancing the model's accuracy The expected output is calculated using a detailed simulation model that comprises several steps. First, a PV system object is created, defining the coordinates (latitude, longitude) and the height of the system. The configuration of the subsystems is also defined, including module tilt, module orientation, system power, inverter power, temperature factor, and type of tracking system. Next, the weather forecast data is prepared by extracting global horizontal irradiation (GHI) and air temperature at a height of 2 meters. This data is converted into a standardized format for the simulation. The solar position is then calculated, determining the solar zenith and azimuth angles for the specified timestamps using the location coordinates. The irradiation is then transposed to the inclined module surfaces, accounting for module inclination and orientation as well as direct, diffuse, and global horizontal irradiation. The system output is simulated by considering the irradiation on inclined surfaces, system temperatures, system configuration, and power losses. Finally, the results are aggregated by combining the simulated performance data for all subsystems into an overall estimate of the expected performance for the entire PV system.

After completing all feature engineering and preprocessing steps, we ended up with two Zarr files. For our investigations, we had the following extensive datasets available:

- UK: 958 PV systems x 78,625 (time dimension) x 55 features = 4,142,751,250 (4.14 billion) observations
- DE: 30 PV systems x 33,117 (time dimension) x 55 features = 54,643,959 (54.64 million) observations

All experiments were then conducted using these Zarr files. Zarr files were used to enable faster reading of the data by the models.

4.3. MODELING & EVALUATION

In this study, we do not do time series forecasting with a sequential input, but we match already existing weather forecasts with a target output value for the solar power at a certain timestamp. Therefore, in this study, we did not use state-of-the-art complex transformer models, but we utilized several regression models, including Linear Regression, LightGBM, Random Forest, CatBoost, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). The following settings for the models remain the same across all experiments to ensure comparability.

Linear Regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It works by fitting a linear equation to the observed data, aiming to minimize the difference between the predicted and actual values (Thombare et al., 2022).

For the LightGBM model, we configured several parameters, including boosting type set to *'gbdt'*, objective set to *'regression'*, metric set to *'rmse'*, and number of leaves set to 31. We also specified a learning rate of 0.05, a feature fraction of 0.9, a bagging fraction of 0.8, and a bagging frequency of 5. LightGBM (Light Gradient Boosting Machine) is a highly efficient gradient boosting framework that uses decision trees as base learners. It operates by building trees sequentially, where each subsequent tree attempts to correct the errors of the previous ones. LightGBM is optimized for speed and performance, handling large datasets and high-dimensional data efficiently (Aksoy & Genc, 2023).

The Random Forest model was initialized with a random state of 42 to ensure reproducibility. Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their predictions. It reduces the risk of overfitting associated with individual decision trees by averaging their results, leading to improved accuracy and robustness (Liu et al., 2019).

Similarly, for the CatBoost model, we set a random seed of 42. CatBoost (Categorical Boosting) is another gradient boosting algorithm specifically designed to handle categorical data effectively. It transforms categorical variables into numerical representations internally, reducing preprocessing effort. CatBoost also mitigates overfitting through ordered boosting and is known for its ease of use and fast training (Prokhorenkova et al., 2019).

The KNN was configured with 10 neighbors. KNN is a non-parametric, instance-based learning algorithm. It predicts the value of a new data point based on the average of its k-nearest neighbors in the feature space. The choice of k and distance metrics are crucial to the performance of the model, with KNN being particularly useful for datasets where the decision boundary is irregular (Song et al., 2017).

Additionally, we explored various architectures for our ANN on a random subset of data and determined that Architecture 4, consisting of layers (128, 64, 32), is best suited for our experiments. The other ANN architectures we evaluated included (64,), (128, 64), (64, 64), and (64, 32, 16). ANNs are inspired by the structure and function of the human brain. They consist of interconnected layers of nodes (neurons), including input layers, hidden layers, and an output layer. Each neuron applies a weighted sum of its inputs and passes the result through a non-linear activation function. ANNs are highly flexible and can model complex, non-linear relationships in data, making them suitable for a wide range of regression tasks (Guresen & Kayakutlu, 2011).

These models were chosen for their diverse approaches and strengths, enabling a comprehensive analysis of the data from multiple perspectives. Through rigorous hyperparameter testing for each regressor, we ensured the optimal configuration to achieve the best performance.

As a performance metric, we only used the RMSE (Root Mean Square Error) for validation. The RMSE is a classic metric and is sufficient for our purposes. It is important to remember that the target, i.e. the produced power, was normalized by the installed power. We lose information about the error in absolute values, but the percentage representation allows us to compare the power of different plants and adequately assess the accuracy of the model:

$$\mathrm{RMSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}} (\mathrm{y_i} - \widetilde{y_i})^2 \qquad \text{(Chai \& Draxler, 2014)}$$

4.4. Experimental Design & Implementation

In this section, we outline the experimental design and implementation of our study, aimed at evaluating the effectiveness and generalizability of different forecasting models. We conducted a series of experiments to assess various factors influencing the accuracy and robustness of the models. Each experiment is designed to address specific research questions and provide insights into model performance under different conditions. The experiments include testing temporal and spatial variations, comparing national versus international data, evaluating different degrees of locality and globality, and determining the optimal balance between global and local models. The following subsections detail the methodology, configuration, and results of each experiment, providing a comprehensive analysis of the models' predictive capabilities.

4.4.1. Exp. 1: Test Dimensions – Temporal vs. PV System Variations

The primary objective of this experiment is to determine which factor is more challenging to predict location (PV System) or time. Since we are not performing classical time series forecasting but rather matching an existing forecast to an output at a specific time, we expect that the model should not be temporally biased when predicting data from the same period for different positional indexes. To test this empirically, this experiment serves as the basis for further experiments in this study and is therefore presented first.

For data preparation, the UK dataset was considered to ensure comprehensive analysis and credibility. To facilitate a temporal division, the median date was calculated based on timestamps, splitting the data into two segments: pre- and post-median date. For positional index division, a random (random state 42) 50% of the selected positional indexes were allocated for training, with the remaining 50% designated for testing.

The model configuration involved using a LightGBM model set up for regression purposes. The model was trained using data up to the median date from 50% of the positional indices. For the validation set, a train-test split was performed with a size of 0.2 and a random state of 42.

For test configurations, four distinct test sets were created:

- Test Set 1 (Same Facilities, New Time): Evaluates the model's performance on new time periods for the same facilities.
- Test Set 2 (New Facilities, Full Time): Assesses the model's accuracy on new facilities over the entire period.
- Test Set 3 (New Facilities, New Time): Tests the model's ability to predict for new facilities in new time periods.
- Test Set 4 (Same Time, New Facilities): Evaluates the model's performance on new facilities within the same time frame as the training data.

The model's performance was evaluated using the Root Mean Squared Error (RMSE) across the validation data and the differently configured test sets. The RMSE values were normalized against the validation RMSE to provide a clear comparison of model performance across different test conditions. This normalization process adjusts the RMSE values relative to the validation RMSE, which serves as a baseline, allowing us to discern how much more difficult each test set is to predict compared to the validation set.

Test Set 1 (same facilities, new time)

Systems	Time									
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
x1	train	train	train	train	train	test	test	test	test	test
x2	train	train	train	train	train	test	test	test	test	test
x3	train	train	train	train	train	test	test	test	test	test
x4	train	train	train	train	train	test	test	test	test	test
x5	not used	not used	not used	not used	not used	not used	not used	not used	not used	not used
x6	not used	not used	not used	not used	not used	not used	not used	not used	not used	not used

Test Set 2 (new facilities, full time)

Systems	Time									
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
x1	train	train	train	train	train	not used	not used	not used	not used	not used
x2	train	train	train	train	train	not used	not used	not used	not used	not used
x3	train	train	train	train	train	not used	not used	not used	not used	not used
x4	train	train	train	train	train	not used	not used	not used	not used	not used
x5	test	test	test	test	test	test	test	test	test	test
x6	test	test	test	test	test	test	test	test	test	test

Test Set 3 (new facilities, new time)

Systems	Time									
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
x1	train	train	train	train	train	not used	not used	not used	not used	not used
x2	train	train	train	train	train	not used	not used	not used	not used	not used
x3	train	train	train	train	train	not used	not used	not used	not used	not used
x4	train	train	train	train	train	not used	not used	not used	not used	not used
x5	not used	not used	not used	not used	not used	test	test	test	test	test
x6	not used	not used	not used	not used	not used	test	test	test	test	test

Test Set 4 (same time, new facilities

Systems	Time									
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
x1	train	train	train	train	train	not used	not used	not used	not used	not used
x2	train	train	train	train	train	not used	not used	not used	not used	not used
x3	train	train	train	train	train	not used	not used	not used	not used	not used
x4	train	train	train	train	train	not used	not used	not used	not used	not used
x5	test	test	test	test	test	not used	not used	not used	not used	not used
x6	test	test	test	test	test	not used	not used	not used	not used	not used

Figure 4-5: Illustration of the different Test Set Configurations. This figure illustrates the different configurations used in the experiment to test the model's performance across two dimensions: time and location. The configurations include: (1) Same Facilities, New Time; (2) New Facilities, Full Time; (3) New Facilities, New Time; and (4) Same Time, New Facilities. Each configuration is designed to evaluate the model's ability to generalize across different temporal and spatial conditions by slicing the data accordingly.

To determine if the differences in RMSE were statistically significant, a bootstrap analysis with 10,000 simulations was performed for each test set. In each of the iterations, a bootstrap sample of the actual and predicted values for both the test and validation datasets was drawn. For each bootstrap sample, the RMSE was calculated. The difference between the RMSE of the test dataset and the RMSE of the validation dataset was recorded.

Confidence intervals were calculated to indicate the range within which the true RMSE difference lies with certain confidence levels. The interpretation of the confidence intervals is as follows:

- The confidence intervals provide a range within which the true difference in RMSEs between test and validation datasets lies with a certain probability (95% or 99%).
- A 95% confidence interval means that we are 95% confident that the true difference in RMSEs lies within this range.
- A 99% confidence interval is narrower and indicates that we are 99% confident that the true difference lies within this range.
- If a confidence interval does not contain zero, this suggests a significant difference between the RMSEs of the test and validation datasets.

These steps ensured a thorough evaluation of the model's predictive accuracy and its ability to generalize across different test conditions.

4.4.2. Exp. 2: Global Models - National vs. International

This experiment is divided into a small pre-experiment and a second main experiment. In the preliminary experiment, the main goal was to analyze the SHAP values to understand the differences in feature importance between the UK and DE datasets. This analysis aims to derive insights to improve the prediction performance for the DE dataset and inform the development of a robust global model. The methodology involved several steps. First, we prepared the data by loading the UK and DE datasets and performing a train-test split for both countries with a test size of 0.2 and a random state of 42. Then, we computed the SHAP values for each feature using LightGBM models. We compared the SHAP values to identify the features with the largest positive and negative differences between the UK and DE datasets. Additionally, we performed a correlation analysis to compute and visualize correlation matrices for both datasets, helping us understand the relationships between features and the target variable. The union set of features identified through these SHAP analyses will be used as input features for the subsequent main experiment.

In the main experiment, we investigate whether incorporating additional training data from the UK improves the predictive accuracy for German facilities. The focus is on comparing the performance of models trained on different combinations of datasets and their ability to generalize to unseen data in Germany. The German data spans from March 2, 2023, to February 9, 2024, and includes 30 positional indexes. The UK data covers the period from January 1, 2018, to October 27, 2021, with 958 positional indexes.

The unique test set, applicable to all test configurations, is created from the German data covering the period from August 1, 2023, to October 31, 2023, to ensure comparability.

1. UK Only:
   - Training Data: Entire UK dataset (January 1, 2018 - October 10, 2021)
   - Validation Data: Train-test split on UK data (test size = 0.3, random state = 42)
2. UK Training, Germany Validation:
   - Training Data: Entire UK dataset (January 1, 2018 - October 27, 2021)
   - Validation Data: Combined German dataset (March 2, 2023 - July 31, 2023, and November 1, 2023 - February 9, 2024). Train-test split performed (test size = 0.3, random state = 42)
3. UK + Germany Combined Dataset:
   - Training Data: UK dataset (January 1, 2018 - October 27, 2021) and German dataset (March 2, 2023 - July 31, 2023, and November 1, 2023 - February 9, 2024)
   - Validation Data: Train-test split on the combined training data (test size = 0.3, random state = 42)
4. Germany Only:
   - Training Data: German dataset (March 2, 2023 - July 31, 2023, and November 1, 2023 - February 9, 2024)
   - Validation Data: Train-test split on German data (test size = 0.3, random state = 42)

The model employed for this experiment is a LightGBM regressor, chosen for its efficiency and performance in handling large datasets with complex feature interactions. Each configuration is evaluated using the RMSE metric.

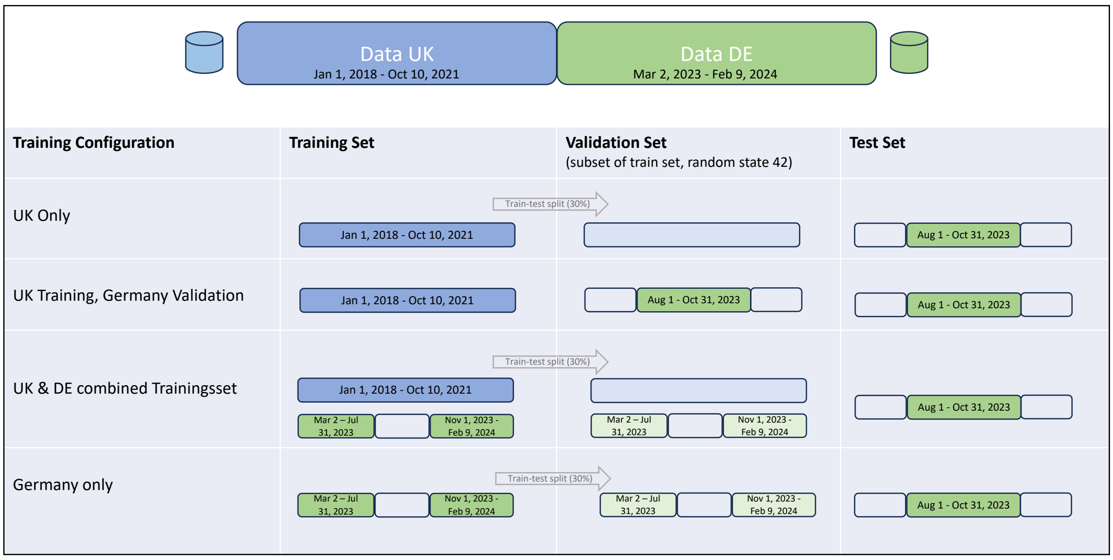

Figure 4-6 Overview of the different training configurations. This figure illustrates the four distinct training configurations used in the experiment: "UK Only," "UK Training, Germany Validation," "UK + Germany Combined Dataset," and "Germany

Only." Each configuration is designed to assess the impact of incorporating UK data on the predictive accuracy for German facilities.

4.4.3. EXP. 3: COMPARING DIFFERENT DEGREES OF LOCALITY AND GLOBALITY

In this experiment, we aimed to determine whether a German global model trained on data from all photovoltaic (PV) plants or a local model trained on individual or clustered plants would yield more accurate predictions. The objective is to compare varying levels of locality and globality in model training. Given the similarity in weather forecast data (and common errors in forecasts for specific regions), a clustered approach might also perform effectively. To investigate this, we designed three distinct scenarios to compare their forecasting performance.

We utilized a dataset containing various meteorological and system-specific features along with the actual PV system outputs (value). The data was divided into a training period (before August 1, 2023) and a testing period (August 1, 2023 to October 31, 2023), known for its challenging prediction conditions.

We conducted the following setups to evaluate the models:

- Setup 1: 29 Training Systems, 1 Test System (30-fold Plant Holdout)
  In this setup, we performed 30-fold plant holdout where each fold had one system as the test set and the remaining 29 systems as the training set. The aim was to assess how well a model trained on most systems could predict an individual system.
- Setup 2: Clustering-Based k-Fold Plant Holdout
  Systems were first clustered into three groups based on their geographical locations using K-Medoids clustering. K-Medoids Clustering is a partitioning method in cluster analysis that divides data points into k clusters, where k corresponds to the number of clusters. K-Medoids uses medoids as cluster centers, which are actual data points within a cluster that have the minimum sum of distances to all other data points in the cluster. Within each cluster, k-fold plant holdout was performed, where k corresponded to the size of each cluster. This involved training on k−1 parts and testing on the remaining part in each fold. This setup aimed to evaluate the performance of localized models within geographically similar groups.

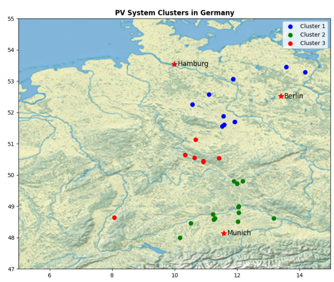

Figure 4-7 PV System Clusters in Germany

- Setup 3: Individual System Models
  Separate models were trained and tested on data from each individual PV system. This setup assessed the performance of models that were specifically tailored to each system.

For each setup and each positional index, the average RMSE across k-plant holdout folds was calculated.

4.4.4. Exp. 4: Optimal Threshold between Global and Local Models

This experiment compares the effectiveness of global and local linear regression training approaches over varying training durations. The dataset is filtered to include data from 30 different locations in Germany, within the time range from August 1, 2023, to August 31, 2023. To generate training data for each positional index, observations from a specified number of days prior to the test period are selected. If insufficient data is available for the specified duration (e.g., in case of 270 days of training data), additional data is taken from the days following the test period to ensure a comprehensive training set. In Experiment 1, we demonstrated that the temporal factor in our model training approach does not introduce a statistically significant bias.

The experiment involves training and evaluating two types of models: global and local. The global model is trained using data from multiple locations and evaluated on a single test location in a holdout setup. A 30-fold plant holdout method is used to ensure robustness, where the dataset is split into 30 different folds, each serving as a test set once while the remaining folds are used for training. In contrast, the local model is trained individually for each location using only the data from that specific location, and the model is evaluated on the same location. This approach aims to understand the performance differences when the model is exposed to diverse versus specific training data.

The primary metric used to evaluate model performance is the Root Mean Squared Error (RMSE), calculated for each positional index daily. This approach ensures that RMSE values are obtained for each day per positional index, providing a detailed comparison between models. Paired t-tests are performed on these RMSE values to statistically compare the performance between global and local models. Confidence intervals are calculated to assess the statistical significance of the observed differences in performance.

The experiment considers multiple training durations ranging from 30 to 340 days. This range helps to understand how the amount of training data affects model performance. For each duration, the RMSE values for global and local models are aggregated and compared. The results are then visualized to highlight the trends and differences in model performance over varying training durations. Additionally, confidence intervals are calculated to assess the statistical significance of the differences observed.

5. RESULTS & DISCUSSION

5.1. EXP. 1: THE IMPACT OF TEST DIMENSIONS – TEMPORAL VS. PV SYSTEM VARIATIONS

Table 5-1 summarizes the RMSE values for the validation set and four distinct test sets to evaluate the model's performance under different conditions. The validation RMSE is 0.0739. To provide a clear comparison, the RMSE values were normalized against the validation RMSE.

Table 5-1 Summary of RMSE Values and Confidence Intervals for Different Test Scenarios This table compares the Root Mean Squared Error (RMSE) values and their confidence intervals for the validation set and four test sets, evaluating the model's performance under various conditions. RMSE values are normalized by deducting the validation RMSE (0.0739) for clear comparison. The table indicates whether predicting new facilities or new time periods presents more difficulty, with confidence intervals showing the statistical significance of these differences. Negative confidence intervals suggest decreased prediction difficulty, while positive values indicate increased difficulty.

Test Scenario	Average RMSE over Test Set	95% Confidence Interval of RMSE Difference (relative to Validation RMSE)	99% Confidence Interval of RMSE Difference (relative to Validation RMSE)
Remaining Time, Same Systems	0.0702	(-0.0018, -0.0012)	(-0.0019, -0.0011)
Same Time, New Systems	0.0723	(0.0004, 0.0010)	(0.0004, 0.0011)
Remaining Time, New Systems	0.0716	(-0.0004, 0.0002)	(-0.0005, 0.0003)
Same Time, New Systems	0.0730	(0.0011, 0.0017)	(0.0009, 0.0018)

*Negative Values in CI: Decrease in prediction difficulty; Positive Values in CI: Increase in prediction difficulty

The normalized RMSE values and their implications for each test set provide valuable insights into the model's performance. For Test Set 1, which involves the same facilities but new time periods, the RMSE shows a significant improvement over the validation set, indicating that the model handles temporal variations well for the same facilities. In contrast, Test Set 2, which involves new facilities over the entire period, shows a marginal increase in prediction difficulty, likely due to the introduction of new facilities. Test Set 3, which involves new facilities in new time periods, shows a slight improvement, suggesting that the model's performance decreases marginally when predicting new times for new facilities, indicating a challenge in generalizing across both dimensions simultaneously. Lastly, for Test Set 4, which involves predicting for new facilities within the same time frame as the training data, the RMSE suggests that this scenario poses the greatest challenge for the model.

The bootstrap analysis with 10,000 simulations was performed for each test set to determine if the differences in RMSE were statistically significant. The confidence intervals for the difference between validation and respective test RMSE values are summarized in the table. These intervals were normalized against the validation RMSE to allow for a clear comparison. The confidence intervals indicate the range within which the true RMSE difference lies with 95% and 99% certainty. If a confidence interval does not contain zero, it suggests a significant difference between the RMSEs of the test and validation datasets.

The direct comparison of the results of Test Set 1 (cutting off the time dimension for testing) and Test Set 4 (cutting off the facilities as a test set) indicates that predicting new facilities is more difficult than predicting new time periods for the same facilities. The RMSE of Test Set 1 is even lower than that of the validation set, albeit marginally. This finding confirms our suspicion that the time dimension does not have the same impact as it does in classic time series forecasting. Therefore, we can disregard the time factor in future experiments, focusing instead on ensuring that the model generalizes well to new systems.

5.2. Exp. 2: Global Models - National vs. International

In the first part of the experiment, several intriguing observations were made. Of the 58 features listed in Table 0-1, Figures 0-3 and 0-4 display the SHAP summary plots for Germany (DE) and the United Kingdom (UK), respectively. These are summarized in Table 0-2 to provide a clearer overview of which features are considered important in both countries and which are important only in individual countries. Figure 0-5 illustrates the difference in mean SHAP values between the UK and DE.

The results of the main experiment, shown in Figure 5-1, demonstrates that training a model exclusively on German data yields better predictive accuracy for German facilities compared to any other setup that includes UK data.

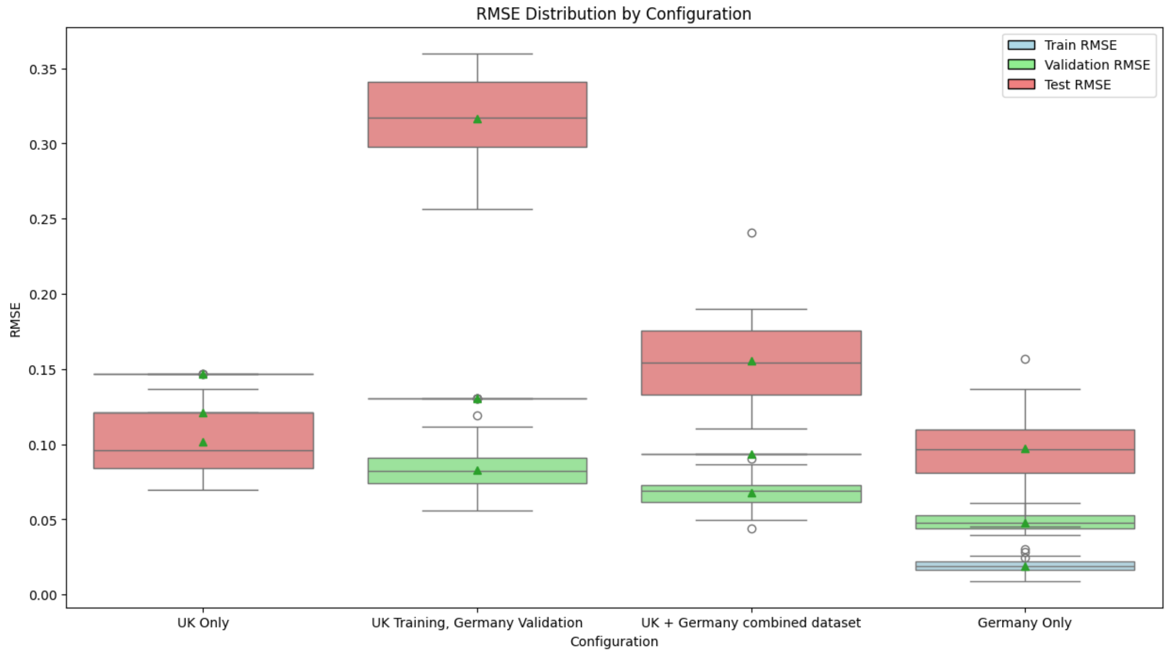

Figure 5-1 RMSE distribution for train, validation, and test sets across different model configurations. The x-axis represents the various configurations (UK Only, UK Training with Germany Validation, UK + Germany Combined Dataset, and Germany Only), while the y-axis shows the RMSE values. The boxplots illustrate the spread of RMSE values, with separate boxes for train, validation, and test RMSEs.

Besides geographic differences, numerous data-related factors could play a role. One of these factors is the different measurement methods and data processing techniques used in different datasets. The UK and Germany might use different measurement methods and sensors, leading to inconsistent datasets. Additionally, different data processing and cleaning methods might be employed, causing variations in the datasets and thus affecting comparability. Another data-related factor is the differences in the types of facilities and technologies used for energy generation. Different technologies and facility types can influence performance data. Furthermore, different maintenance and operational practices can lead to variations in performance data, impacting predictive accuracy. Finally, data quality and availability play a crucial role. The quality and availability of data can vary between the UK and Germany. Differences in the completeness and accuracy of datasets can negatively impact model performance and reduce the ability to generalize to new data.

In Figure 0-6, we observe the difference in correlation with the target (UK values – DE values). What stands out here is that the simulation in the UK has significantly less correlation with the target than in Germany (more than 80% difference), indicating that the simulation in the UK might be inaccurate. The UK data have fixed values for azimuth and tilt, while in Germany, these values change every 15 minutes. Since these two factors are included in the simulation, the error in the UK data might be due to inaccuracies or erroneous data.

Figures 0-7 and 0-8 in the annex depict the target variable (i.e., normalized performance of the PV system) and some important performance-determining features for Germany and the UK for a week, and Figures 0-9 and 0-10 depict the same for a year. From the first two figures, we can see that the data quality in the UK is poorer than in Germany. Not only were nights cut from the UK data, but there are also visible missing values during the day. It is also evident that the performance in relation to performance-determining factors appears significantly higher in the UK than in Germany. This difference is also visible in the annual time series. The figure for Germany shows how the daily maximum value ('value') consistently lies between the *clear_sky_ghi_max* and *cos_zenith_min*. When looking at the maximum values of the target variable 'value' from the UK, it regularly exceeds the *clear_sky_ghi_max* curve. This is another indication that the UK data follow different rules than the German data. A concept drift between the German and UK data would make it difficult to predict German data using UK data.

A significant limitation is that we do not know if the data from Open Climate Fix are correct. They themselves say there can be measurement errors, but they do not specify how many and which facilities. This leads to concept drift when applying the model to other facilities that are likely measured differently and correctly. This limitation could cause problems in accurately predicting German data. Feedback from Open Climate Fix noted that outliers in the data, particularly normalized target values exceeding 1, could be due to noise or incorrect capacity inputs. This includes potential inaccuracies in the measurement of tilt and orientation by installation engineers. These inaccuracies and inconsistencies in data measurement and recording can introduce significant noise, which hinders the model's ability to generalize well from UK data to German facilities.

This feedback aligns with our hypothesis that data-related factors, such as differences in measurement techniques, data processing methods, and overall quality and availability of data, play a significant role in the observed performance gap. Variations in data collection and processing between the UK and Germany can lead to discrepancies that our model cannot effectively reconcile, resulting in poorer predictive accuracy when UK data is included in the training set.

5.3. EXP. 3: COMPARING DIFFERENT DEGREES OF LOCALITY AND GLOBALITY

The results shown in Figures 5-2 and 5-3 indicate that local models (Setup 3) consistently perform better across all regressors. There is a tendency for the clustered data (Setup 2) to have the highest error values compared to the other setups. At first glance, looking at the mean RMSE averaged over the different PV systems, it seems that the LightGBM regressor performs the best in all setups.

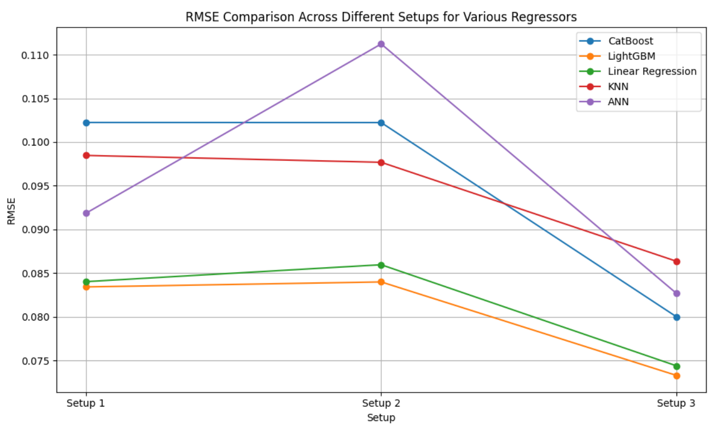

Figure 5-2 Mean RMSE Averaged Across All PV Systems for Different Setups. This figure provides the average RMSE values for each setup, calculated across all photovoltaic systems. It offers a clear comparison of the overall accuracy of different setups in modeling solar power output.

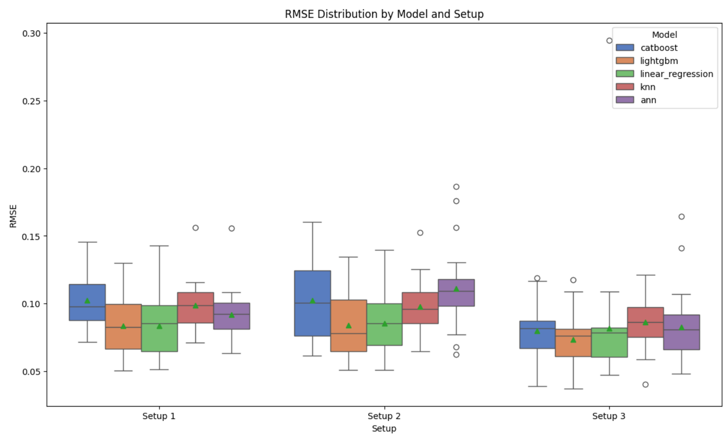

Figure 5-3: Boxplots of RMSE Distribution for Different Setups Across Multiple PV Systems. This radar plot compares the performance of different regressors on various PV systems. For each setup, the RMSE values are scaled using min-max normalization: 1 - (x - x.min()) / (x.max() - x.min()). This normalizes and inverts the values so that lower RMSE values (better performance) are closer to the outside of the radar plot. Each axis represents one PV system.

However, upon closer examination of Figure 5-3, there is an outlier observed with the linear regression, which has significantly higher RMSE values. This is further illustrated in the radar plots in Figure 5-4, particularly in the bottom radar plot for the local models (Setup 3). It becomes evident that for PV system DE-4363, the linear regression has a much higher error, but it continuously performs the best for all other PV systems.

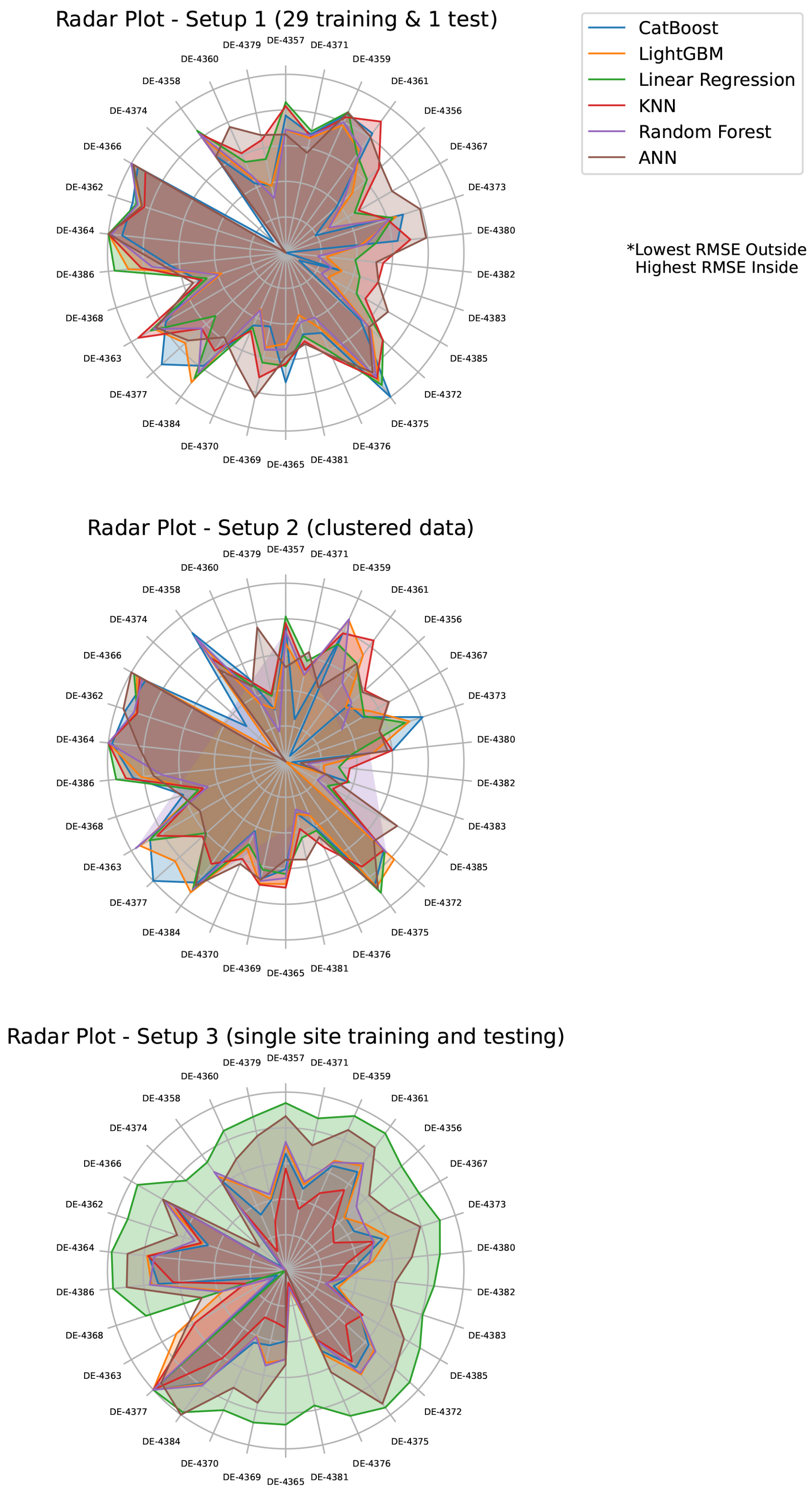

Figure 5-4 Radar Plot Evaluation of Regressors Across Different PV Systems. This radar plot compares the performance of different regressors on various PV systems. Each axis represents a PV system, and the RMSE values indicate the effectiveness of each regressor in modeling the power output for each system. The lowest RMSE are on the outside of the radar plot.

5.4. Exp. 4: Optimal Threshold between Global and Local Models

Table 5-2 summarizes the mean RMSE values for global and local models, along with the confidence levels indicating the statistical significance that the better model outperforms the other.

Table 5-2: Comparison of RMSE values for global and local models trained with varying durations of historical data. The first column, "Duration (days)," lists the number of training days used for model training. The second column identifies the superior model (local or global). The third column provides the confidence level (in percentage) indicating the statistical significance of the superior model. The last two columns show the RMSE values for both the global and local models.

Duration (days)	Better Model	Confidence for better Model	Global RMSE	Local RMSE
30	global	98,56%	11,43%	18,49%
90		57,65%	9,63%	10,28%
120		36,12%	9,61%	9,97%
130		24,99%	9,64%	9,88%
140	local	21,01%	9,61%	9,47%
150		42,04%	9,67%	9,42%
160		92,77%	9,59%	8,98%
170		96,63%	9,58%	8,91%
180		97,23%	9,57%	8,88%
230		98,28%	9,56%	8,85%
340		98,55%	9,57%	8,84%

The results indicate a clear trend in model performance based on the amount of training data available. For shorter training durations (up to 130 days), global models consistently outperform local models, as evidenced by higher confidence levels in the statistical tests. For example, with only 30 days of training data, the global model achieves a significantly lower RMSE (11,42%) compared to the local model (18,48%), with a confidence level of 98.56% indicating that the global model is better.

However, as the training duration increases beyond 140 days, local models begin to outperform global models. This is evident from the RMSE values and the increasing confidence levels. For instance, with 160 days of training data, the local model achieves an RMSE of 8.98%, which is lower than the global model's RMSE of 9.59%, with a high confidence level of 92.77%. This trend continues, with local models consistently achieving lower RMSE values compared to global models for training durations of 170 days and beyond. By 340 days, the local model achieves an RMSE of 8.84% compared to the global model's 9.57%, with a confidence level of 98.55%.

The results are visually represented in Figure 5-5, which plots RMSE values for both global and local models across different training durations. The chart clearly shows the crossover point around 140 days, where the performance of local models begins to surpass that of global models. This visual aid reinforces the statistical findings and highlights the importance of sufficient training data for improving model accuracy in local contexts.

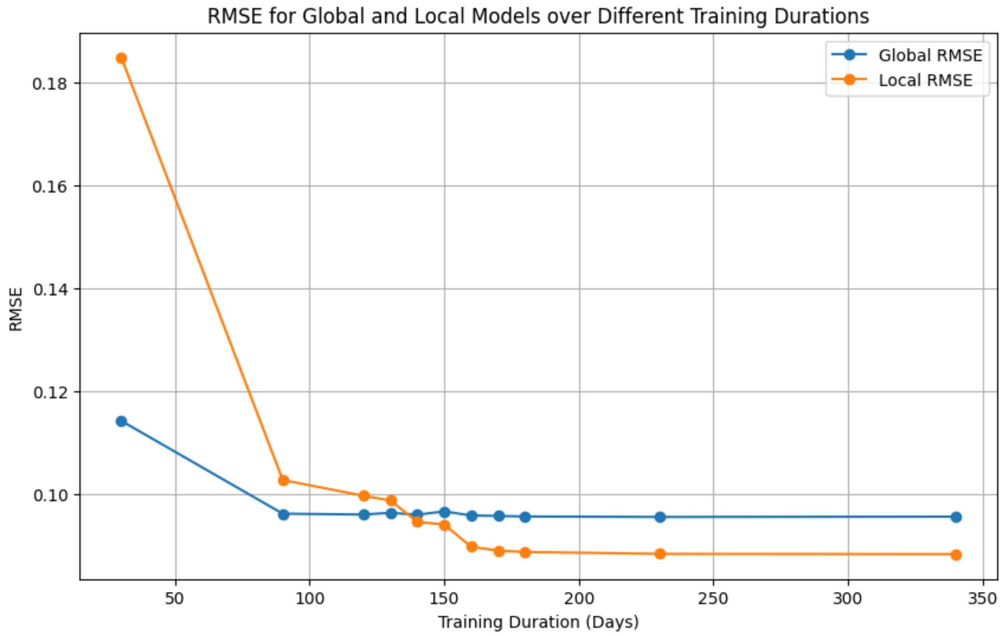

Figure 5-5: RMSE values for global and local models over different training durations, illustrating how the model performance changes with varying amounts of historical data. The graph shows that with fewer training days, global models perform better, but around the 140-day mark, local models begin to outperform global models.

In summary, the experiment shows not only that global models are more effective with limited training data, but also when and with what reliability local models become superior as the amount of data increases.

6. CONCLUSIONS

This thesis aimed to address several critical challenges in the field of solar power forecasting, specifically focusing on data scarcity in new installations and the scalability of existing forecasting models. We examined the efficacy of global versus local models in predicting solar power generation, utilizing hybrid models that integrate both machine learning and physical models.

Our findings indicated that simpler models, such as linear regression, often outperformed more complex models in scenarios characterized by significant uncertainties in weather forecasts. This suggests that complex models may overfit the noise inherent in weather forecast data.

The study demonstrated that national global models trained on Numerical Weather Prediction (NWP) data within Germany can generalize effectively to unseen installations, thereby mitigating the issue of data scarcity in new solar installations. In the context of various degrees of globality, our experiments compared international, national, and regional global models against local models. The results showed that global models outperformed local models when historical data was limited, specifically for training periods up to approximately 130 days. However, this transition is not abrupt but rather gradual. Specifically, the performance advantage of global models is statistically significant with 95% confidence for training periods of 30 days or less. For training periods between 30 and 170 days, the confidence in the superiority of global over local models decreases gradually and remains below 95%. Only after 170 days can we state with 95% confidence that the local models begin to outperform the global models, with their performance continuing to improve with additional historical data.

Additionally, our study highlighted that the UK data was not suitable for predicting German solar power output, underscoring significant differences in the relationship between feature and target data between the UK and Germany. Issues with data quality in the UK datasets further complicated the model's ability to generalize across these regions.

In conclusion, this study provides a nuanced understanding of the conditions under which global and local models excel. It offers practical guidelines for determining the optimal approach based on the availability of historical data, thereby enhancing the scalability and efficiency of solar power forecasting models.

This project was significant for the company as it enabled us to create precise solar forecasts for new installations with limited historical data for the first time. Thanks to these models, our start-up can now compete with established competitors on the German market. In addition, we now know exactly when to use local and when to use global models, which makes our forecasts even more reliable.

7. LIMITATIONS & FUTURE WORK

One significant limitation of this study is the inconsistency in the impact of key influencing features on solar performance between the German and UK datasets. The UK data suffered from poor quality, characterized by instances where the power normalized by the installed power exceeded 1 at certain times, necessitating manual corrections, as well as frequent missing values. The exact reasons for these discrepancies are unclear. Future work should focus on ensuring that data collection methods are standardized internationally. This could involve using data from internationally operating energy suppliers and implementing uniform protocols for historical power data collection. Such standardization would enable more accurate performance comparisons across different locations and minimize distortions in the models.

Due to the varying availability of data between the UK and Germany, features like azimuth and tilt were incorporated indirectly through physical simulations in this study. Using azimuth, tilt, and other factors, solar performance was simulated with a physical model and used as an input feature for the machine learning models. Future research should aim to directly integrate these features into the machine learning models, provided that the underlying data for these features are consistently accurate across all datasets.

Improving the accuracy of weather forecasts and overall model performance could be achieved by incorporating additional data. Several strategies can be explored: Integrating input data from the ICON-EU weather models could provide explanatory data to correct errors in weather predictions, aiding in error mitigation. This study used only one-day-ahead data from a five-day forecast. Future work could leverage the remaining four days of forecasts to reduce weather prediction errors in solar power forecasting. Additionally, using multiple weather models could minimize the impact of errors from individual weather services on the results. More historical data from PV systems would help to make the models more stable and shorten the transition period of 140 days between the respective 95% confidence levels of the global and local models.

In our methodology, we trained our models on 15-minute interval data from photovoltaic (PV) systems. The weather forecast data were originally provided in 1-hour intervals and then interpolated to 15-minute intervals. This high-frequency data integration provides detailed insights into PV performance but may also introduce significant noise due to the interpolation process and short-term fluctuations in solar power output. These fluctuations are challenging to predict accurately, which impacts the performance of our forecasting models. To address this purely using Numerical Weather Predictions (NWPs) and historical performance data, future research could explore a denoising approach. A promising method would be to aggregate the solar power data to hourly intervals, matching the granularity of the original weather data. Training the models on this smoothed, hourly data could improve model accuracy by focusing on more stable, less noisy trends. After training, the model outputs can be interpolated back to 15-minute intervals for detailed prediction purposes. This method aims to harmonize the data sources by finding a common temporal resolution, potentially enhancing the overall forecast accuracy. Furthermore, future work should combine various data sources, which could further enhance the accuracy and robustness of solar power forecasting models. For instance, integrating satellite data, sky-image data, and other relevant information can make the models more resilient and accurate in predicting solar power output. As noted by Cargan et al. (2023), future research should aim to combine satellite and ground-based weather data and increase the temporal resolution to better capture intra-hourly variations. By leveraging these diverse data sources, models can be significantly improved in their prediction capabilities.

Future research should explore the use of advanced deep learning techniques capable of handling sequential input data effectively. Techniques such as PatchTST (Patch-based Transformer for Sequential Tasks) or Temporal Fusion Transformer could be particularly beneficial in enhancing model performance.

BIBLIOGRAPHICAL REFERENCES

Abdelmoula, I. A., Elhamaoui, S., Elalani, O., Ghennioui, A., & Aroussi, M. E. (2022). A photovoltaic power prediction approach enhanced by feature engineering and stacked machine learning model. Energy Reports, 8, 1288–1300. https://doi.org/10.1016/j.egyr.2022.07.082

Aksoy, N., & Genc, I. (2023). Predictive models development using gradient boosting based methods for solar power plants. Journal of Computational Science, 67, 101958. https://doi.org/10.1016/j.jocs.2023.101958

Behera, M. K., Majumder, I., & Nayak, N. (2018). Solar photovoltaic power forecasting using optimized modified extreme learning machine technique. Engineering Science and Technology, an International Journal, 21(3), 428–438. https://doi.org/10.1016/j.jestch.2018.04.013

Cargan, T., Landa-Silva, D., & Triguero, I. (2023). Local-Global Methods for Generalised Solar Irradiance Forecasting (arXiv:2303.06010). arXiv. https://doi.org/10.48550/arXiv.2303.06010

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? https://doi.org/10.5194/gmdd-7-1525-2014

Chen, S.-M., Chang, Y.-C., Chen, Z.-J., & Chen, C.-L. (2013). Multiple fuzzy rules interpolation with weighted antecedent variables in sparse fuzzy rule-based systems. International Journal of Pattern Recognition and Artificial Intelligence, 27(05), 1359002. https://doi.org/10.1142/S0218001413590027

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1), 113–126. https://doi.org/10.1016/S0893-6080(03)00169-2

Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., Horan, B., & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews, 81, 912–928. https://doi.org/10.1016/j.rser.2017.08.017

FCMR. (2003). Report on Wind Chill Temperature and Extreme Heat Indices: Evaluation and Improvement Projects. Yumpu.Com. https://www.yumpu.com/en/document/view/10797752/report-on-wind-chill-temperature-and-extreme-heat-indices-

Fonseca Junior, J. G. da S., Oozeki, T., Ohtake, H., Takashima, T., & Ogimoto, K. (2015). Regional forecasts of photovoltaic power generation according to different data availability scenarios: A study of four methods. Progress in Photovoltaics: Research and Applications, 23(10), 1203–1218. https://doi.org/10.1002/pip.2528

Gaboitaolelwe, J., Zungeru, A. M., Yahya, A., Lebekwe, C. K., Vinod, D. N., & Salau, A. O. (2023). Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison. IEEE Access, 11, 40820–40845. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3270041

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Guresen, E., & Kayakutlu, G. (2011). Definition of artificial neural networks with comparison to other networks. Procedia Computer Science, 3, 426–433. https://doi.org/10.1016/j.procs.2010.12.071

Heppelmann, T., Steiner, A., & Vogt, S. (2017). Application of numerical weather prediction in wind power forecasting: Assessment of the diurnal cycle. Meteorologische Zeitschrift, 26. https://doi.org/10.1127/metz/2017/0820

Iheanetu, K. J. (2022). Solar Photovoltaic Power Forecasting: A Review. Sustainability, 14(24), Article 24. https://doi.org/10.3390/su142417005

Jiang, X., Jiang, M., & Zhou, Q. (2023). Day-Ahead PV Power Forecasting Based on MSTL-TFT. https://doi.org/10.48550/arXiv.2301.05911

Kühnert, J. (2016). Development of a photovoltaic power prediction system for forecast horizons of several hours [Phd, Universität Oldenburg]. https://oops.uni-oldenburg.de/2841/

Lago, J., De Brabandere, K., De Ridder, F., & De Schutter, B. (2018). Short-term forecasting of solar irradiance without local telemetry: A generalized model using satellite data. Solar Energy, 173, 566–577. https://doi.org/10.1016/j.solener.2018.07.050

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2020). Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting (arXiv:1912.09363). arXiv. https://doi.org/10.48550/arXiv.1912.09363

Lipperheide, M., Bosch, J. L., & Kleissl, J. (2015). Embedded nowcasting method using cloud speed persistence for a photovoltaic power plant. Solar Energy, 112, 232–238. https://doi.org/10.1016/j.solener.2014.11.013

Liu, J., Cao, M. Y., Bai, D., & Zhang, R. (2019). Solar radiation prediction based on random forest of feature-extraction. IOP Conference Series: Materials Science and Engineering, 658(1), 012006. https://doi.org/10.1088/1757-899X/658/1/012006

Lonij, V. P. A., Brooks, A. E., Cronin, A. D., Leuthold, M., & Koch, K. (2013). Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. Solar Energy, 97, 58–66. https://doi.org/10.1016/j.solener.2013.08.002

López Santos, M., García-Santiago, X., Echevarría Camarero, F., Blázquez Gil, G., & Carrasco Ortega, P. (2022). Application of Temporal Fusion Transformer for Day-Ahead PV Power Forecasting. Energies, 15(14), Article 14. https://doi.org/10.3390/en15145232

Mellit, A., Massi Pavan, A., & Lughi, V. (2014). Short-term forecasting of power production in a large-scale photovoltaic plant. Solar Energy, 105, 401–413. https://doi.org/10.1016/j.solener.2014.03.018

Monjoly, S., André, M., Calif, R., & Soubdhan, T. (2019). Forecast Horizon and Solar Variability Influences on the Performances of Multiscale Hybrid Forecast Model. Energies, 12(12), Article 12. https://doi.org/10.3390/en12122264

Montero-Manso, P., & Hyndman, R. J. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. International Journal of Forecasting, 37(4), 1632–1653. https://doi.org/10.1016/j.ijforecast.2021.03.004

Nespoli, A., Ogliari, E., Leva, S., Massi Pavan, A., Mellit, A., Lughi, V., & Dolara, A. (2019). Day-Ahead Photovoltaic Forecasting: A Comparison of the Most Effective Techniques. Energies, 12(9), Article 9. https://doi.org/10.3390/en12091621

Nie, Y., Li, X., Paletta, Q., Aragon, M., Scott, A., & Brandt, A. (2022). Open-Source Ground-based Sky Image Datasets for Very Short-term Solar Forecasting, Cloud Analysis and Modeling: A Comprehensive Survey (arXiv:2211.14709). arXiv. https://doi.org/10.48550/arXiv.2211.14709

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A Time Series is Worth 64 Words: Long-term Forecasting with Transformers (arXiv:2211.14730). arXiv. https://doi.org/10.48550/arXiv.2211.14730

Nie, Y., Paletta, Q., Scott, A., Martín Pomares, L., Arbod, G., Sgouridis, S., Lasenby, J., & Brandt, A. (2022). Sky-image-based solar forecasting using deep learning with multi-location data: Training models locally, globally or via transfer learning? https://doi.org/10.48550/arXiv.2211.02108

Nie, Y., Paletta, Q., Scott, A., Pomares, L. M., Arbod, G., Sgouridis, S., Lasenby, J., & Brandt, A. (2022). Sky-image-based solar forecasting using deep learning with multi-location data: Training models locally, globally or via transfer learning? (arXiv:2211.02108). arXiv. https://doi.org/10.48550/arXiv.2211.02108

NWS. (1990). Heat Index Equation. https://www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml openclimatefix. (2022). Openclimatefix Dataset at Hugging Face. https://huggingface.co/datasets/openclimatefix/uk_pv

openclimatefix. (2023, November 7). Openclimatefix/uk_pv · PV system capacity and generation data. https://huggingface.co/datasets/openclimatefix/uk_pv/discussions/5

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). CatBoost: Unbiased boosting with categorical features (arXiv:1706.09516). arXiv. https://doi.org/10.48550/arXiv.1706.09516

Raza, M. Q., Nadarajah, M., & Ekanayake, C. (2016). On recent advances in PV output power forecast. Solar Energy, 136, 125–144. https://doi.org/10.1016/j.solener.2016.06.073

Ren, Y., Suganthan, P. N., & Srikanth, N. (2015). Ensemble methods for wind and solar power forecasting—A state-of-the-art review. Renewable and Sustainable Energy Reviews, 50, 82–91. https://doi.org/10.1016/j.rser.2015.04.081

Shi, J., Lee, W.-J., Liu, Y., Yang, Y., & Wang, P. (2012). Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. IEEE Transactions on Industry Applications, 48(3), 1064–1069. IEEE Transactions on Industry Applications. https://doi.org/10.1109/TIA.2012.2190816

Sobri, S., Koohi-Kamali, S., & Rahim, N. Abd. (2018). Solar photovoltaic generation forecasting methods: A review. Energy Conversion and Management, 156, 459–497. https://doi.org/10.1016/j.enconman.2017.11.019

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26–34. https://doi.org/10.1016/j.neucom.2017.04.018

Steadman, R. G. (1979). The Assessment of Sultriness. Part I: A Temperature-Humidity Index Based on Human Physiology and Clothing Science. Journal of Applied Meteorology and Climatology, 18(7), 861–873. https://doi.org/10.1175/1520-0450(1979)018<0861:TAOSPI>2.0.CO;2

Thombare, S., Pande, V. N., Kulkarni, R. S., & Kakade, S. M. (2022). Prediction of Solar Power Using Linear Regression. https://doi.org/10.1007/978-981-16-9033-4_53

Umer, F., Aslam, M. S., Rabbani, M. S., Hanif, M. J., Naeem, N., & Abbas, M. T. (2019). Design and Optimization of Solar Carport Canopies for Maximum Power Generation and Efficiency at Bahawalpur. International Journal of Photoenergy, 2019, e6372503. https://doi.org/10.1155/2019/6372503

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T., & Shi, M. (2020). A day-ahead PV power forecasting method based on LSTM-RNN model and time correlation modification under partial daily pattern prediction framework. Energy Conversion and Management, 212, 112766. https://doi.org/10.1016/j.enconman.2020.112766

Yona, A., Senjyu, T., Funabashi, T., & Kim, C.-H. (2013). Determination Method of Insolation Prediction With Fuzzy and Applying Neural Network for Long-Term Ahead PV Power Output Correction. IEEE Transactions on Sustainable Energy, 4(2), 527–533. IEEE Transactions on Sustainable Energy. https://doi.org/10.1109/TSTE.2013.2246591

Zekai, S. (2008). Solar Energy Fundamentals and Modeling Techniques. https://link.springer.com/book/10.1007/978-1-84800-134-3

APPENDIX A

Table A-0-1: Comprehensive overview of all the available features for the experiments conducted in the master thesis

Source		Feature Name	Description	Unit
target		**value**	**Normalized solar power**	**Percent (%) of Installed Capacity**
Feature Engineering	Feature Encoding	apparent_temperature	Apparent temperature	Degree Celsius (℃)
		time_of_day_sinus	Sine value of the time of day (for cycle calculation)	Dimensionless
		time_of_day_cosinus	Cosine value of the time of day (for cycle calculation)	Dimensionless
		yearly_sinus	Sine value of the yearly cycle (for cycle calculation)	Dimensionless
		yearly_cosinus	Cosine value of the yearly cycle (for cycle calculation)	Dimensionless
		time_of_day_numeric	Numeric value of the time of day (0-23)	Hours
		hour_of_day	Hour of the day (0-23)	Hours
		horizon_in_min	Forecast Horizon in minutes	Minutes
		cos_zenith	Cosine of the solar zenith angle (sun elevation)	Dimensionless
		clear_sky_ghi	Global radiation under cloudless sky	Watt per square meter (W/m^2)
	target encoding	mean_last_1_day	Mean value of the last 24 hours	Percent (%) of Installed Capacity
		mean_last_7_days	Mean value of the last 7 days	Percent (%) of Installed Capacity
		mean_last_30_days	Mean value of the last 30 days	Percent (%) of Installed Capacity
		std_last_1_day	Standard deviation of the last 24 hours	Percent (%) of Installed Capacity
		std_last_7_days	Standard deviation of the last 7 days	Percent (%) of Installed Capacity
		std_last_30_days	Standard deviation of the last 30 days	Percent (%) of Installed Capacity
		max_last_1_day	Maximum value of the last 24 hours	Percent (%) of Installed Capacity
		max_last_7_days	Maximum value of the last 7 days	Percent (%) of Installed Capacity
		max_last_30_days	Maximum value of the last 30 days	Percent (%) of Installed Capacity
		value_24h_ago	Value 24 hours ago	Percent (%) of Installed Capacity
		value_5day_avg	Average of the values at the same time of day from the last 5 days.	Percent (%) of Installed Capacity
	Simulation	simulation	Expected normalized solar power, calculated by physical model	Percent (%) of Installed Capacity
icon-EU, German Weather Service		air_temperature_2m	Air temperature at 2 meters height	Degree Celsius (℃)
		air_temperature_500	Air temperature at 500 hPa pressure level	Degree Celsius (℃)
		air_temperature_850	Air temperature at 850 hPa pressure level	Degree Celsius (℃)
		convective_rain	Convective precipitation	Millimeter (mm)
		dewpoint_2m	Dew point temperature at 2 meters height	Degree Celsius (℃)
		geopotential_height_500	Geopotential height at 500 hPa pressure level	Meter (m)
		geopotential_height_850	Geopotential height at 850 hPa pressure level	Meter (m)
		global_horizontal_irradiance	Global radiation	Watt per square meter (W/m^2)
		high_level_clouds	High-level cloud cover	Percent (%)
		low_level_clouds	Low-level cloud cover	Percent (%)
		max_air_temperature_2m	Maximum value of air temperature at 2 meters height	Degree Celsius (℃)
		max_wind_gust_10m	Wind gust peak at 10 meters height	Meter per second (m/s)
		medium_level_clouds	Medium-level cloud cover	Percent (%)
		min_air_temperature_2m	Minimum value of air temperature at 2 meters height	Degree Celsius (℃)
		mixed_layer_cape	Convective available potential energy	Joule per kilogram (J/kg)
		pressure_mean_sea_level	Mean sea level pressure	hPa
		relative_humidity_2m	Relative humidity at 2 meters height	Percent (%)
		snow_depth	Snow depth	Centimeter (cm)
		specific_humidity_1500	Specific humidity at 1500 hPa pressure level	Gram per kilogram (g/kg)
		specific_humidity_3000	Specific humidity at 3000 hPa pressure level	Gram per kilogram (g/kg)
		surface_temperature	Surface temperature	Degree Celsius (℃)
		total_cloud_cover	Total cloud cover	Percent (%)
		total_precipitation	Total precipitation	Millimeter (mm)
		weather_synop_code	Weather SYNOP code	N/A
		wind_direction_10	Wind direction at 10 meters height	Degrees (°)
		wind_direction_129	Wind direction at 129 meters height	Degrees (°)
		wind_direction_20	Wind direction at 20 meters height	Degrees (°)
		wind_direction_211	Wind direction at 211 meters height	Degrees (°)
		wind_direction_309	Wind direction at 309 meters height	Degrees (°)
		wind_direction_64	Wind direction at 64 meters height	Degrees (°)
		wind_speed_10	Wind speed at 10 meters height	Meters per second (m/s)
		wind_speed_129	Wind speed at 129 meters height	Meters per second (m/s)
		wind_speed_20	Wind speed at 20 meters height	Meters per second (m/s)
		wind_speed_211	Wind speed at 211 meters height	Meters per second (m/s)
		wind_speed_309	Wind speed at 309 meters height	Meters per second (m/s)
		wind_speed_64	Wind speed at 64 meters height	Meters per second (m/s)

For weather_synop_code, there is no specific unit as it represents a numerical representation of weather according to the SYNOP code

Table A-0-2: Utilization of Features in Models for the UK and Germany. This table lists the feature names and indicates their usage in the respective models for the UK and Germany. A check mark in a column signifies that the feature was employed in the model for that country. This comparison highlights which features are deemed relevant and utilized for predicting solar power output in each geographical context.

Feature	UK	DE
clear_sky_ghi	✓	✓
global_horizontal_irradiance	✓	✓
time_of_day_sinus	✓	✓
simulation	✓	✓
cos_zenith	✓	✓
pressure_mean_sea_level	✓	✓
std_last_30_days	✓	✓
yearly_sinus	✓	✓
yearly_cosinus	✓	✓
time_of_day_cosinus	✓	✓
geopotential_height_850	✓	✓
relative_humidity_2m	✓	X
surface_temperature	✓	X
max_last_30_days	✓	X
mean_last_30_days	✓	X
_cos_zenith	✓	X
max_last_7_days	✓	X
low_level_clouds	✓	X
mean_last_7_days	✓	X
total_precipitation	✓	X
max_last_1_day	X	✓
std_last_1_day	X	✓
mean_last_1_day	X	✓
total_cloud_cover	X	✓
specific_humidity_1500	X	✓
air_temperature_500	X	✓
std_last_7_days	X	✓
geopotential_height_500	X	✓
specific_humidity_3000	X	✓

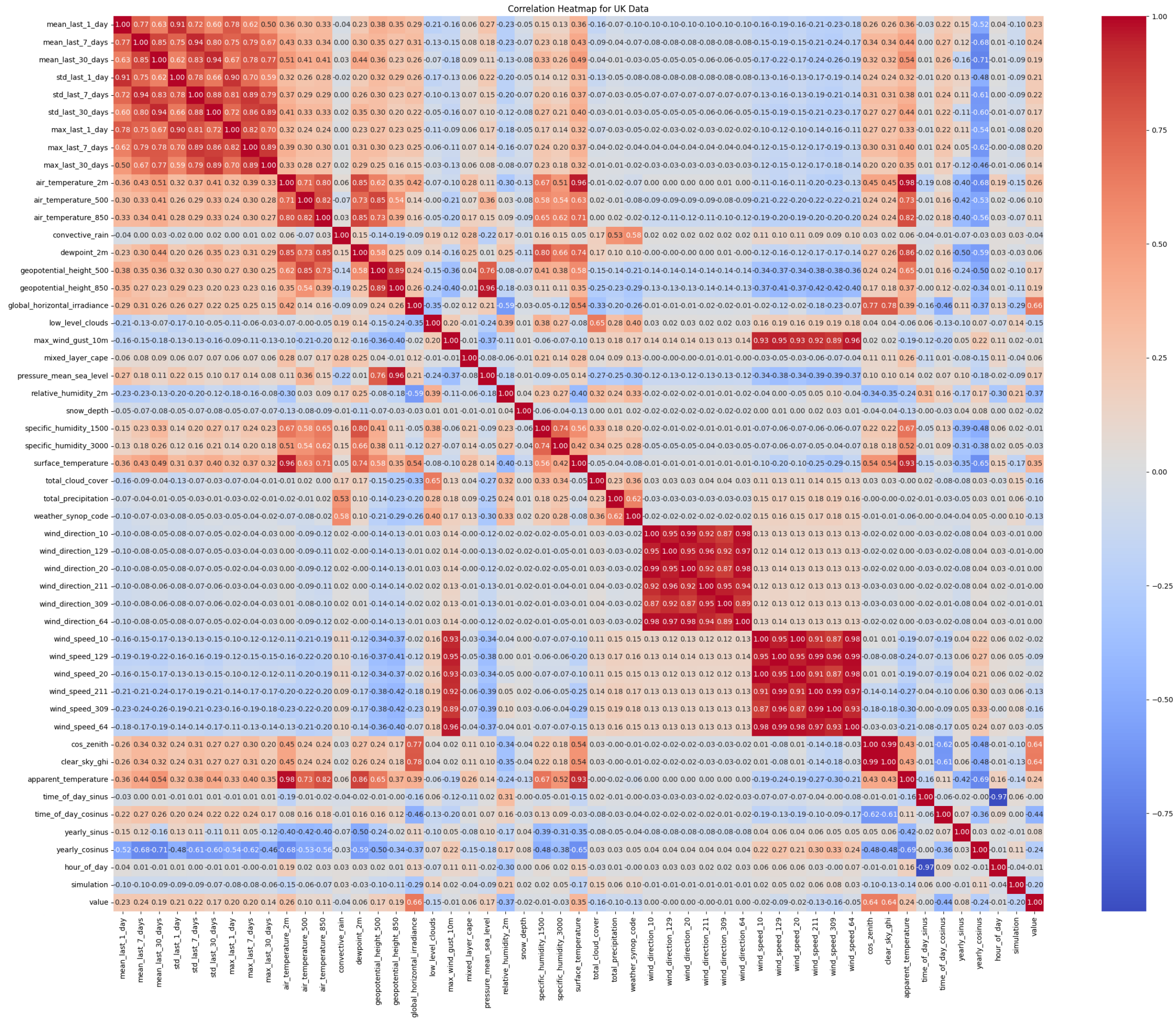

Figure A-0-1: Pearson Correlation Heatmap for UK

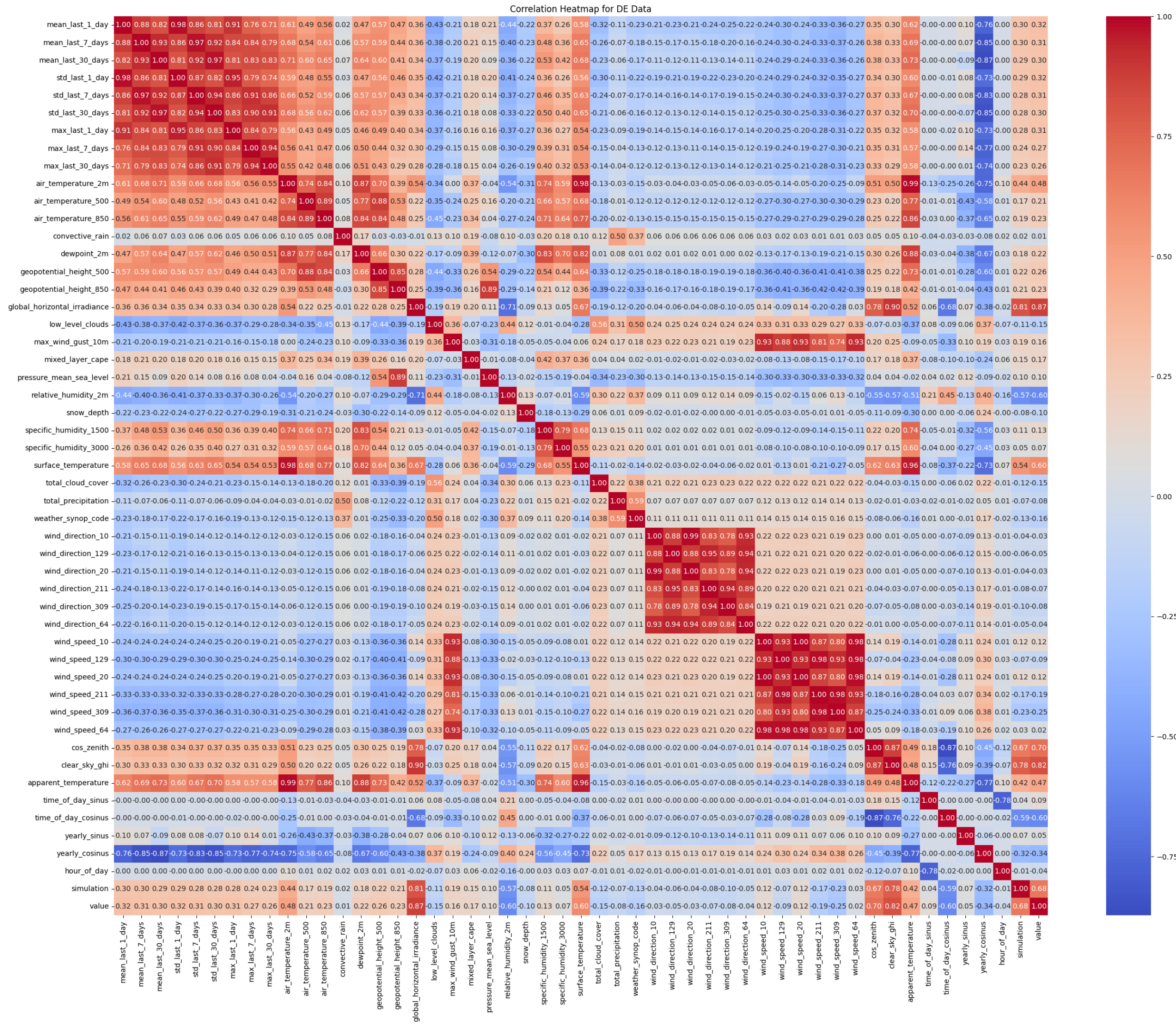

Figure A-0-2: Pearson Correlation Heatmap for DE

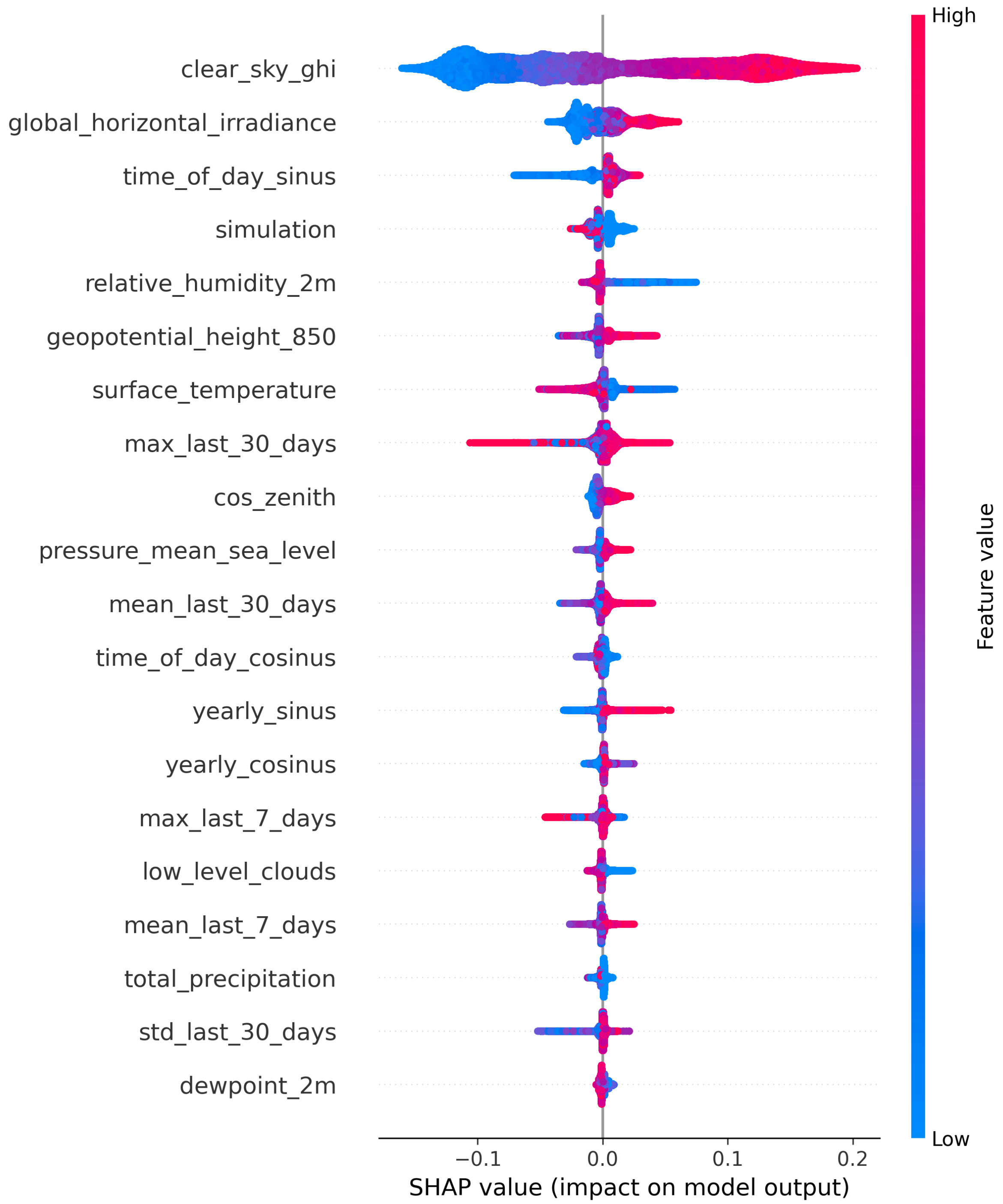

Figure A-0-3: SHAP Summary Plot for UK

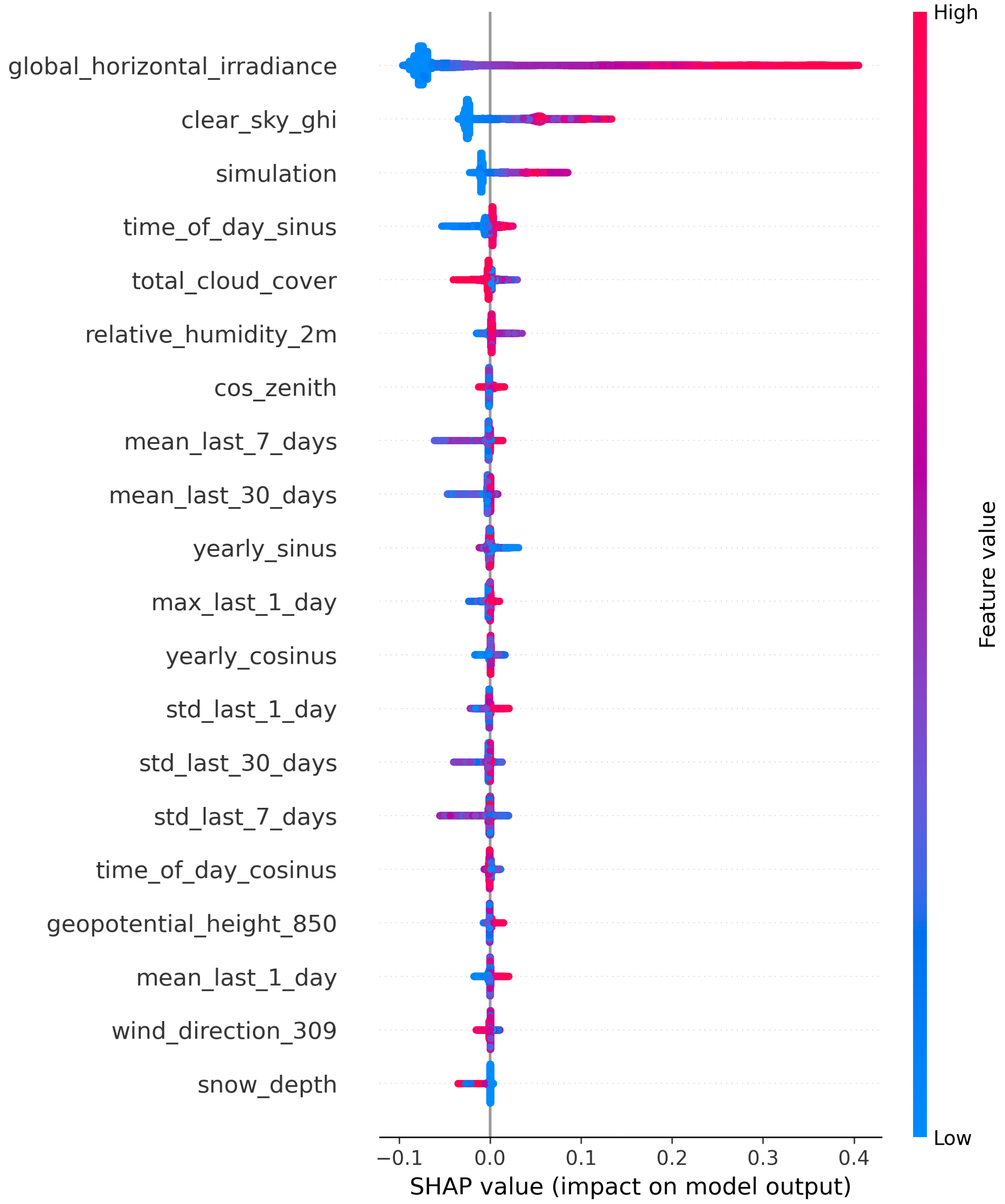

Figure A-0-4: SHAP Summary Plot for Germany

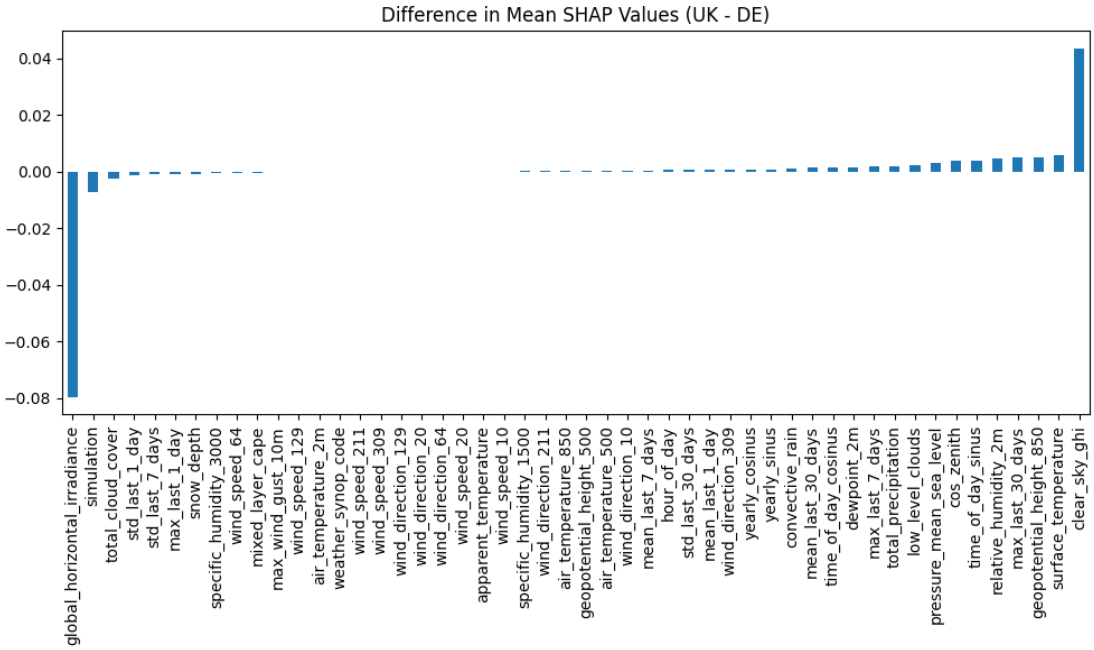

Figure A-0-5: Difference in mean SHAP values between the UK and Germany. The values were calculated by subtracting the mean SHAP values of Germany from the mean SHAP values of the UK, meaning negative values are more importante in Germany, whereas positive values are more importante in the UK.

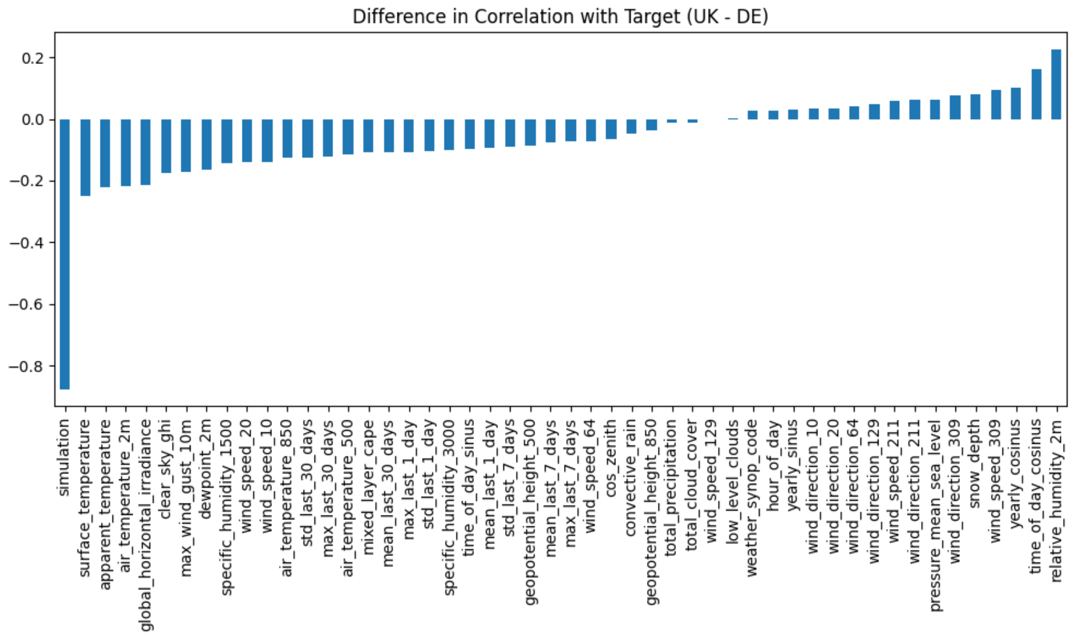

Figure A-0-6 Difference in Correlation with Target, the correlations of the features in DE are subtracted from the correlations of the same features in the UK

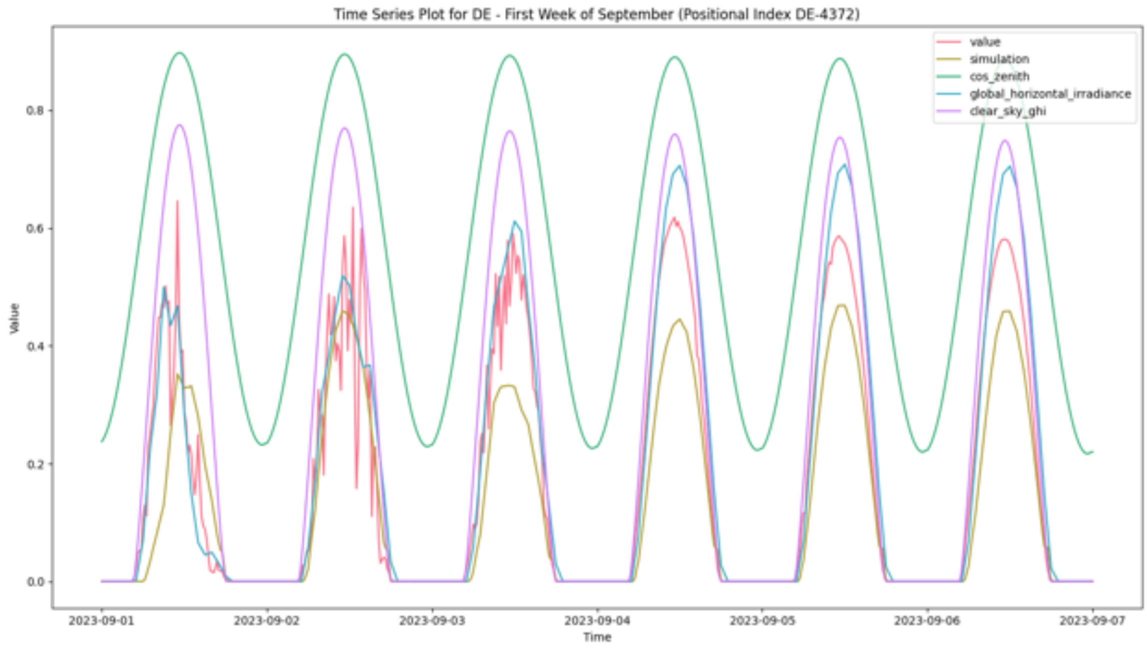

Figure A-0-7 Temporal development of the target variable and important features over a period of one week in Germany

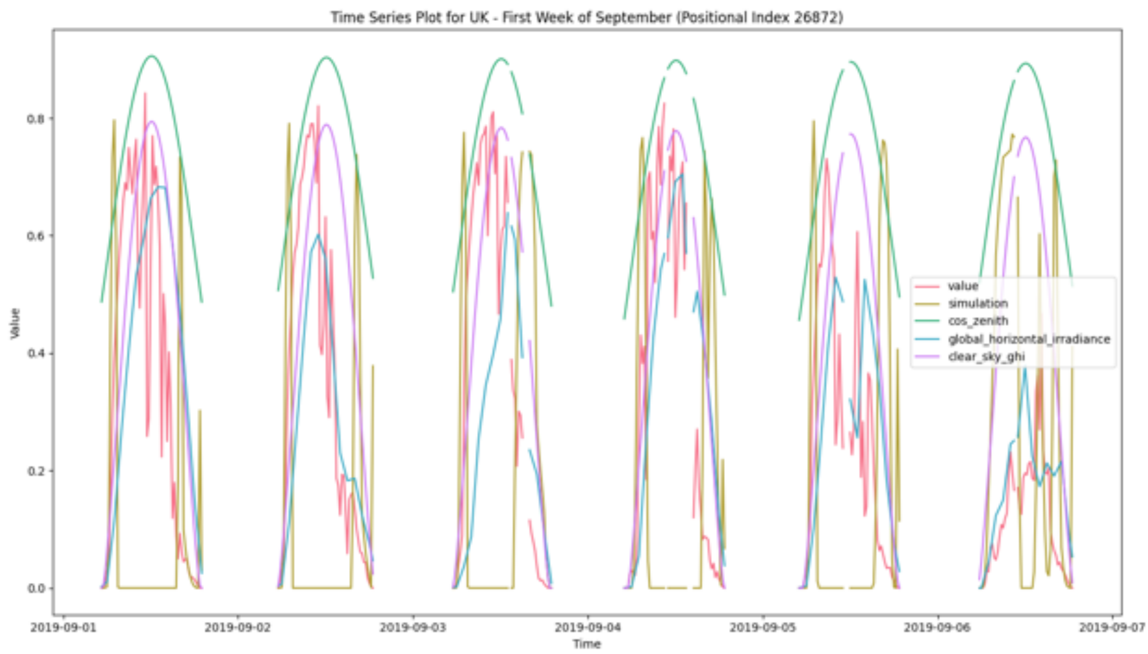

Figure A-0-8 Temporal development of the target variable and important features over a period of one week in the UK

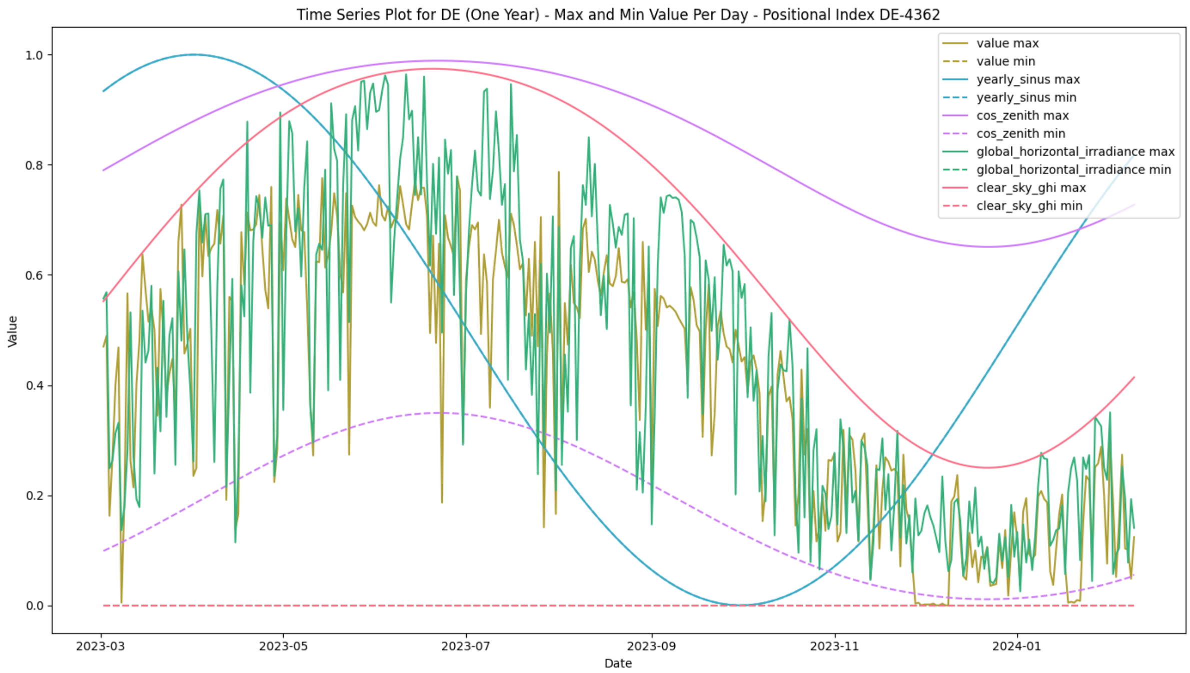

Figure A-0-9 Temporal development of the target variable and important features over a period of one year in Germany

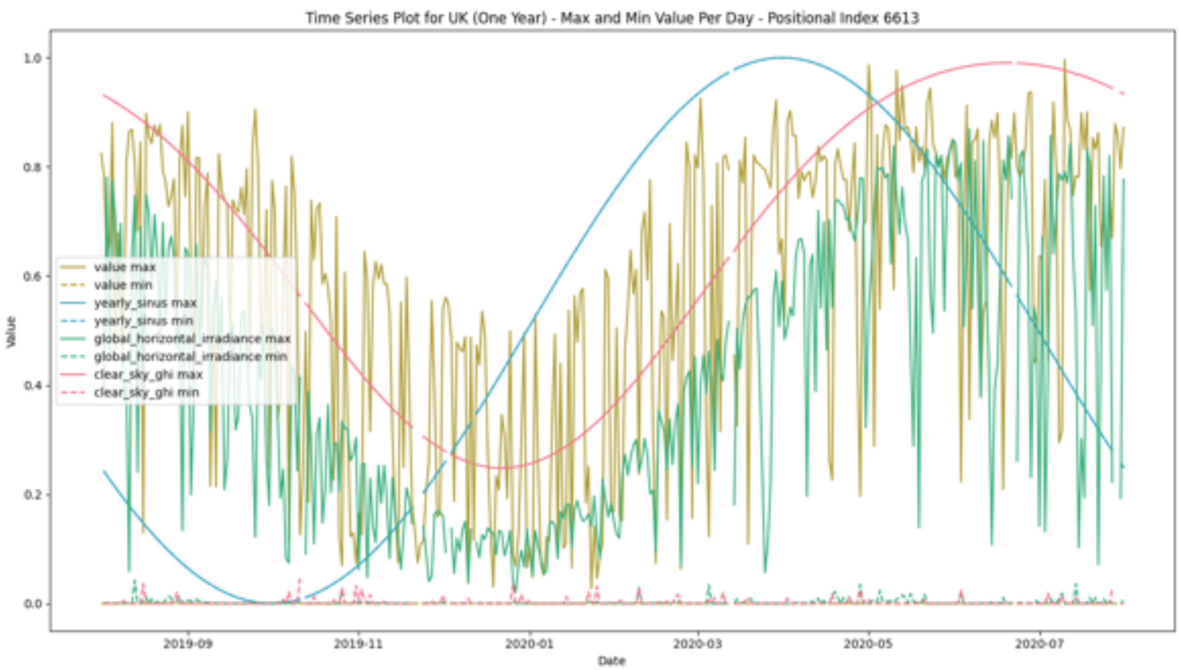

Figure A-0-10 Temporal development of the target variable and important features over a period of one year in the UK

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa