Combining Machine Learning with Physics-based Models for Day-ahead Solar Forecasting

Rong Gu
St Hilda's College

University of Oxford
A thesis presented for the degree of
Master of Science by Research

Trinity 2025

“Today’s posterior distribution is tomorrow’s prior”

— Dennis Lindley

Acknowledgements

I would like to thank my supervisors Michael Osborne and David Howey for their invaluable support and guidance for my thesis. Their dual perspectives perspectives of machine learning and applicational science throughout my degree broadened my understanding of the topics and taught me an incredible amount. I would also like to express my gratitude for my labmates in the Bayesian Exploration and Battery Intelligence labs for always being available to help with my questions and troubleshooting.

A special thank you to Peter Dudfield from OCF for sharing his deep industry insight and suggestions for my research, as well as COSF for their financial support throughout my degree.

Declaration

I, Rong Gu, declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated above through the relevant publications and their corresponding sections. Since some works were carried out under supervisors' guidance, the term "we" will be used instead of "I".

Rong Gu

October 2025

Abstract

This thesis presents a hybrid approach for day-ahead solar power forecasting that integrates a deterministic physical model with a Gaussian process (GP) post-processing component. We begin by reviewing physical forecasting models, focusing on deterministic and ensemble-based models. To improve predictive accuracy, GPs are employed to correct residual errors in the physical model outputs. Two post-processing methods are explored: a time-series GP and an autoregressive GP, each assessed using two hyperparameter inference strategies: maximum likelihood estimation and the no-U-turn sampler.

The hybrid model is evaluated on three datasets: a household photovoltaic (PV) system in Oxford, small to medium-scale PV sites in Hong Kong, and a range of PV systems in various locations in the UK. The approach demonstrates consistent performance across diverse climates and system configurations, effectively capturing fluctuations in solar output and adapting predictions in the absence of complete system metadata, such as shading or inverter characteristics.

The proposed model yields lower forecasting errors compared to a variety of benchmark models. These include statistical methods such as ARIMA and GluonTS, neural network-based approaches including long short-term memory and transformer models, and tree-based ensemble methods such as random forest and XGBoost. These benchmarks are evaluated in both post-processing and direct forecasting settings. In addition, the performance of the model is compared with Quartz, a commercial tool used for direct PV power forecasting. The results highlight the potential of combining physical models with machine learning techniques to improve the accuracy and generalisability of short-term solar power and energy forecasting.

Keywords: solar, forecast, Gaussian process, hybrid, model

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AC ..... Alternating Current

ANN ..... Artificial Neural Network

ARGP ..... Auto-Regressive Gaussian Process

ARIMA ..... Auto-Regressive Integrated Moving Average

BMA ..... Bayesian Model Averaging

CDF ..... Cumulative Distribution Function

CNN ..... Convolutional Neural Network

CRPS ..... Continuous Ranked Probability Score

DC ..... Direct Current

DHI ..... Diffuse Horizontal Irradiance

DNI ..... Direct Normal Irradiance

ECMWF ..... European Centre for Medium-Range Weather Forecasts

EMOS ..... Ensemble Model Output Statistics

FNN ..... Feed-forward Neural Network

GAN ..... Generative Adversarial Network

GHI ..... Global Horizontal Irradiance

GP ..... Gaussian Process

GPR ..... Gaussian Process Regression

GRU ..... Gated Recurrent Unit

HK . . . . . Hong Kong

HKUST . . . . . Hong Kong University of Science and Technology

HMC . . . . . Hamiltonian Monte Carlo

LSTM . . . . . Long Short-Term Memory Network

MAE . . . . . Mean Absolute Error

MAPE . . . . . Mean Absolute Percentage Error

MARS . . . . . Multivariate Adaptive Regression Splines

MCMC . . . . . Markov Chain Monte Carlo

ML . . . . . Machine Learning

MLE . . . . . Maximum Likelihood Estimation

NUTS . . . . . No-U-Turn Sampler

nRMSE . . . . . Normalised Root Mean Square Error

NWP . . . . . Numerical Weather Prediction

POA . . . . . Plane of Array

PV . . . . . Photovoltaic

QRF . . . . . Quantile Regression Forest

RBF . . . . . Radial Basis Function

RMSE . . . . . Root Mean Square Error

RNN . . . . . Recurrent Neural Network

SVM . . . . . Support Vector Machine

XGBoost . . . . . Extreme Gradient Boosting

1 | Introduction

The global deployment of photovoltaic (PV) systems has accelerated dramatically over the past decade, making solar energy one of the fastest-growing renewable energy sources worldwide. The cumulative installed PV capacity increased from just over 100 GW in 2012 to more than 1.6 TW in 2023 [Jäger-Waldau, 2023]. With continued policy support and declining technology costs, projections anticipate global PV installations to exceed 2.8 TW by 2030 [Pourasl et al., 2023]. This growth spans all sectors of the market, from residential rooftop systems to utility-scale solar farms. In particular, utility-scale installations have dominated additions of PV capacity. In 2022, about 60% of global solar PV investment was in utility-scale plants, compared to approximately 30% in residential and 11% in commercial installations [Jäger-Waldau, 2023]. Although large-scale PV systems account for the majority of cumulative capacity, distributed PV systems, such as rooftop systems, have also expanded significantly, positively contributing to national energy portfolios and decentralised energy systems.

The rapid rise of PV generation introduces substantial forecasting challenges. Solar power generation is highly variable, influenced by weather conditions such as irradiance, cloud cover, and temperature, as well as site-specific factors. These fluctuations complicate grid operations, especially as the PV penetration rate increases. In large-scale systems, inaccurate forecasts can lead to suboptimal scheduling, higher reserve margins, and increased operational costs [Leo et al., 2025]. In household settings, forecast errors can result in inefficient battery usage, greater dependence on the grid, and lower long-term financial returns [Mirletz and Guittet, 2023]. Therefore, accurate PV power forecasting is essential for optimising energy usage in household applications and improving grid integration in both commercial and utility-scale systems.

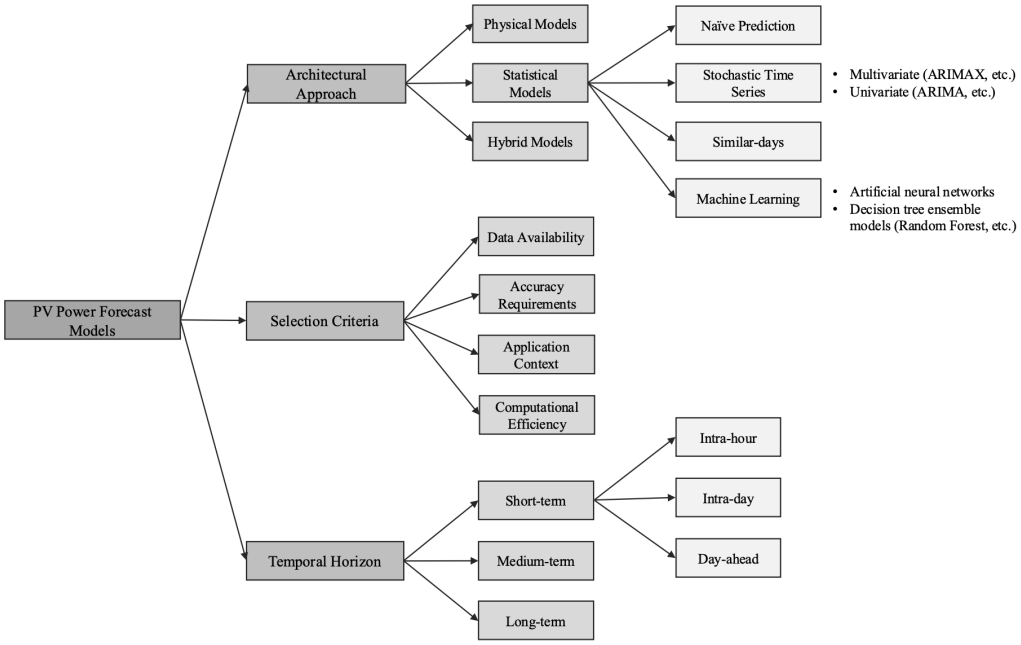

Figure 1.1: Overview of PV power forecasting approaches, categorised by temporal horizon, architecture, and selection criteria. Short-term horizons include d intra-day and day-ahead forecasts with lead times of up to 24 hours, medium-term horizons cover multi-day predictions up to weekly timescales, and long-term horizons extend from multiple weeks to seasonal or even annual timescales.

The overview in Figure 1.1 categorises PV forecasting models along three key dimensions: temporal horizon, architecture, and selection criteria. In terms of forecast horizon, solar forecasting can be categorised into three time types: short-term, mid-term, and long-term forecasts [Leo et al., 2025]. Short-term forecasts, including forecasts for the day and the day ahead, are the most widely used in operational power system management. Intra-hour and intra-day forecasts provide predictions of PV generation from a few minutes up to several hours in advance and are essential for real-time grid balancing and dispatch decisions [Leo et al., 2025]. Their main challenge lies in their high sensitivity to rapid weather fluctuations, such as fast-moving clouds, which can cause sudden and unpredictable changes in solar output. Day-ahead forecasts, which predict PV generation 24 hours in advance, are crucial for scheduling energy markets and unit commitment [Leo et al., 2025]. These forecasts are heavily dependent on numerical weather prediction (NWP) models and are commonly used due to their relevance to next-day operational planning. However, their precision can be significantly affected by meteorological uncertainties, particularly in regions with highly variable weather conditions. Mid-term forecasts, covering several days to weeks, support asset management and seasonal planning, but become less reliable as the accuracy of the weather model decreases over time [Leo et al., 2025, Chen and Zhao, 2023]. Long-term forecasts, covering months to years, are used for long-term strategy decisions such as capacity expansion and policy making, although their accuracy is limited by climate variability and the extended time horizon [Leo et al., 2025, Chen and Zhao, 2023]. Studies have found that under clear skies, day-ahead PV forecasts can achieve a root mean square error (RMSE) as low as 6%, whereas cloudy conditions can drive errors as high as 20–80% [Theocharides et al., 2024, Glassley et al., 2011, Lorenz et al., 2012]. In contrast, even small improvements in forecast accuracy yield substantial benefits. For example, a case study in Cali-

fornia showed that a 35% improvement in intra-day ramp prediction could reduce annual spinning reserve costs by approximately $5 million for a 4 GW solar portfolio [of Energy, 2016]. Similarly, at the household scale, day-ahead PV prediction errors on the order of only 3-14% have been shown to diminish the economic performance of PV-battery systems [Mirletz and Guittet, 2023].

PV power predictions may be forecasted directly from a range of predictors [Ulbricht et al., 2013]. Alternatively, they can be calculated using irradiance forecast data produced by NWP or sky and satellite imagery. PV power calculation methods can be classified into physical, statistical, or hybrid approaches [Antonanzas et al., 2017, Leo et al., 2025]. Physical methods use theoretical simulation models, such as irradiance-to-power conversion models, electrical simulation models, and transposition models, to calculate the output power of a PV system based on its main design parameters. In contrast, statistical methods encompass all data-driven approaches, including both classical statistical modelling and the novel machine learning (ML) algorithms. Finally, hybrid approaches combine the above approaches, consisting of two or more different methods [Antonanzas et al., 2017]. Recent literature highlights the effectiveness of hybrid approaches, which often reduce errors such as MAPE and nRMSE by approximately 7-26% compared to single-method approaches [Leo et al., 2025].

Motivated by the growing importance of PV forecasting and the associated challenges, this thesis proposes a hybrid forecasting framework for day-ahead solar power prediction that combines a deterministic physical model with Gaussian process (GP) post-processing. The GP component aims to learn systematic discrepancies in the physical model output, improving both point forecasts and uncertainty quantification. The hybrid model is evaluated across three distinct datasets: a household PV system in Oxford, small-to-medium scale systems in Hong Kong (HK), and a diverse portfolio of PV sites across the UK. This multi-regional and

multi-scale evaluation enables a thorough examination of forecasting challenges under varying climatic and system conditions.

The following sections review the physical, statistical and hybrid forecasting methods (Section 1.1), provide an overview of available PV time-series datasets along with their limitations (Section 1.2) and outline the main contributions of this thesis (Section 1.3). The available PV time-series datasets often lack critical metadata such as system orientation, tilt, or inverter capacity, which underscores the need for adaptable forecasting methods.

1.1 PV Forecasting Methods

As previously introduced, the three architectures of solar forecasting are physical, statistical, and hybrid. Physical models, often referred to as white-box models, derive from the fundamental principles of PV energy conversion [Amiri et al., 2024]. These models estimate PV output based on system-specific characteristics and environmental conditions, requiring detailed inputs such as solar irradiance, temperature, wind speed, panel orientation, and shading losses, without the necessity of historical data [Al-Dahidi et al., 2024]. Their transparency allows for interpretability and diagnostics, but this also means they are highly sensitive to errors in input data and incomplete system metadata [Chicco et al., 2016]. The physical models can be classified into three types: clear-sky models, transposition models and semi-empirical models [Leo et al., 2025]. Clear-sky models estimate the theoretical maximum irradiance under cloudless conditions using astronomical and atmospheric data, providing a useful baseline for performance assessment. Transposition models convert direct and diffuse irradiance into plane-of-array irradiance for tilted PV surfaces. These models often rely on decomposition models, which divide total solar irradiance into its direct and diffuse components to better capture

PV system behaviour under varying sky conditions. Lastly, semi-empirical models combine physical equations with empirically derived coefficients to estimate PV power output based on environmental factors. In practice, transposition and semi-empirical models are often used in sequence with other physical components as part of an integrated modelling pipeline. A widely adopted approach in physical PV power modelling involves converting global horizontal irradiance (GHI) forecasts from NWP models into PV power forecasts through a 'model chain'. The term 'model chain' refers to a series of physical models that simulate various stages that include solar positioning, the separation of beam and diffuse irradiance, the transposition of horizontal irradiance to a tilted surface, PV cell temperature and performance, shading loss, and other factors [Yang et al., 2019, Wang et al., 2022, Lorenz et al., 2011]. Traditionally, the model chain is a deterministic process, assigning a single value to each time step over the forecast horizon. Recently, probabilistic forecasts have been proposed that use ensemble model chains [Mayer and Gróf, 2021, Mayer and Yang, 2022]. In an ensemble model chain setup, each model chain with combinations of physical models is treated as a probable path for irradiance-to-power conversion, enhancing forecast accuracy and helping to quantify forecast uncertainty. Additionally, NWP models can operate in ensemble mode by generating multiple simulations with varied initial conditions or model physics. Mayer et al. [Mayer et al., 2024] previously discussed the improved performance when combining ensemble NWP models with ensemble model chains. The limitations of physical model chains include their high sensitivity to irradiance and weather forecast errors. Moreover, the effectiveness of these models is contingent on the physical models selected within the model chain, requiring detailed knowledge about the PV system design parameters and site-specific factors such as shading, reflection, and soiling. Furthermore, simpler physical models often exhibit systematic bias and produce forecast ranges that are too narrow, in-

dicating an underestimation of the true variability in PV generation [Mayer and Gróf, 2021].

Statistical models are also extensively used in PV power forecasting [Antonanzas et al., 2016]. These black-box models are data-driven, relying on historical datasets of solar irradiance and power output without explicit knowledge of the physical processes or the target system's design parameters. Commonly used statistical models include naïve forecasting, similar-days models, stochastic time series, and ML approaches [Ulbricht et al., 2013]. Naïve forecasting and similar-days models are straightforward, as they directly apply historical PV power data. Stochastic time series models establish temporal dependencies in PV power generation data and can be divided into univariate and multivariate categories. Popular univariate models include auto-regressive integrated moving average (ARIMA) and exponential smoothing, and multivariate models include ARIMA with exogenous inputs (ARIMAX) and multivariate adaptive regression splines (MARS) [Friedman, 1991]. By contrast, ML models exploit advanced algorithms to uncover complex patterns and relationships in data [Ogliari et al., 2017b]. Among published work on forecasting PV power with ML, frequently used methods include artificial neural networks (ANNs) [Ogliari et al., 2017b, Theocharides et al., 2020], support vector machines (SVMs), and random forests [Antonanzas et al., 2017]. ANNs are supervised learning models that consist of interconnected nodes capable of learning non-linear relationships among input variables and PV power. SVMs are also supervised learning models that identify an optimal hyperplane to maximise the margin between classes in classification tasks, or to fit a function within a specified margin in regression tasks such as PV power forecasting. Random forests are an ensemble learning method that constructs multiple decision trees and aggregates their outputs to improve predictive accuracy and robustness. Deep learning methods, such as convolutional neural networks (CNNs), recurrent

neural networks (RNNs), and long short-term memory networks (LSTMs) [Wang et al., 2019], extend ANNs and have also been applied to day-ahead PV power forecasting. These models use meteorological inputs to capture complex hierarchical and temporal patterns in PV power data. However, the performance of these data-driven models is highly dependent on the amount and quality of training data available, and their accuracy is limited when less than 1–3 years of historical data are available [Wang et al., 2019].

As mentioned above, hybrid models combine the strengths of multiple models by effectively capturing various data patterns. Combining linear and non-linear models has also been suggested to enhance forecast accuracy [Zhang, 2003]. Recent proposals include hybrid models that merge CNNs and RNNs [Xiang et al., 2024], as well as those that integrate physical and statistical approaches—e.g., incorporating physical calculations into neural networks [Schmelas et al., 2015, Ogliari et al., 2017a], and employing ensemble model output statistics (EMOS) [Horat et al., 2024].

1.2 Key Inputs for PV Forecasting

As shown in Figure 1.1, the selection criteria of PV forecast models include data availability, accuracy requirements, application context, and computational efficiency. In this section, we focus on the availability and quality of input data, which directly influence the forecast accuracy of the chosen model and the suitability of its modelling approach. We also discuss some common data-related challenges in PV forecasting, along with mitigation methods.

In general, PV output is driven by both environmental factors and system-specific characteristics. Table 1.1 summarises the typical datasets required for PV power forecasting models.

Table 1.1: Typical data inputs for PV forecasting models

Data Type	Description	Model Usage
Solar irradiance	GHI, DHI, DNI derived from NWP or satellite.	Physical / Statistical
Meteorological variables	Temperature, wind, cloudiness derived from NWP or satellite.	Physical / Statistical
PV system metadata	Location, tilt, azimuth, capacity, inverter type.	Physical
Historical PV output	Time-series of observed PV generation.	Statistical / Validation

The availability of these datasets often dictates which forecasting model is suitable. For example, a physics model typically requires detailed irradiance forecasts and system parameters. If such granular data or metadata are unavailable, one might opt for a direct data-driven model that predicts power output from historical patterns without explicitly calculating irradiance. In other words, the choice between different model paradigms (physical, statistical, or hybrid) is closely related to data availability. A traditional model that demands high-resolution weather forecasts and complete site specifications may yield accurate results in theory, but it is only practical if those inputs are available and reliable. Therefore, effective forecasting models must be adaptable to the available dataset and robust to missing or uncertain input.

1.2.1 Challenges in PV Time Series Data

In this section, we give an overview of the challenges often present in PV forecasting. Although often overlooked in literature, these issues are critical for real-world applications, and are thus taken into account in the framework design of the proposed method presented in this thesis. PV time-series input data presents several key challenges that hinder accurate forecasting. These include missing values, temporal resolution mismatches, and strong diurnal and seasonal patterns in both input and output data.

Missing Data and Incomplete Metadata: For the variables in Table 1.1, historical PV output often contains missing values due to sensor malfunctions, communication failures, or maintenance activities. These missing entries can occur randomly or during specific periods, such as night-time when measurements are typically absent. Incomplete metadata, such as location, panel orientation, surface azimuth, nominal capacity, and inverter type, can further complicate the use of physical models and normalisation processes. Such data deficiencies may lead to inaccurate clipping estimates or errors in incidence angle calculations and irradiance transposition models. In these cases, it is useful to apply data-driven methods to correct for these offsets.

Irregular Sampling and Outliers: In real-world scenarios, PV data often exhibit irregular sampling intervals due to logging inconsistencies or adjustments to daylight saving time. Moreover, outliers, such as sudden spikes and dips caused by sensor faults or shading events, can significantly bias model training, especially for algorithms sensitive to extreme values. For example, [Osborne et al., 2010] explored ML methods for detecting sensor failures. If not addressed properly, these issues can lead to over-fitting and poor generalisation.

Diurnal and Seasonal Cyclicality: Both PV modelling input (irradiance and weather data) and output (PV power observations) may exhibit strong diurnal and seasonal patterns. Daily production typically follows a bell-shaped curve, peaking at solar noon and dropping to zero at night. Seasonal variations in hourly irradiance affect the amplitude and duration of this curve, influenced by factors such as day length and sun angle. These cyclic patterns introduce non-stationarity into the time series, posing challenges for models that assume constant statistical properties over time [Gayathry et al., 2024].

Temporal Granularity Mismatch: A common issue in PV forecasting is the discrepancy between the temporal resolution of weather inputs and PV output observations. NWP models, such as the Integrated Forecasting System used by the European centre for medium-range weather forecasts (ECMWF), typically produce forecasts at coarse temporal resolutions (e.g., hourly), whereas PV data may be recorded at finer intervals (e.g., every few minutes). This mismatch can lead to aliasing errors and challenges in capturing short-term fluctuations, such as those caused by passing clouds. Upsampling or downsampling techniques are employed to reconcile these differences, but they may smooth out critical variations necessary for accurate forecasting [Inman et al., 2013].

1.2.2 Mitigation Approaches

To address the above challenges, two broad strategies are employed: (1) data preprocessing techniques to clean and transform the input time-series, and (2) forecasting model architectures explicitly designed to be robust against data quality issues. This section surveys established methods in each category.

1.2.2.1 Data Processing

Preprocessing is an essential first step to improve data quality and align the inputs with model requirements. Key techniques include:

Missing Data Imputation: When PV measurements or meteorological inputs are missing or incomplete, analysts either interpolate or use model-based imputation to fill gaps and ensure smooth, realistic fills. Simple approaches for handling missing data include linear interpolation, forward-filling, and backward filling, though these methods can easily introduce bias or reduce accuracy and may not be suitable for modelling highly dynamic variables like irradiance.

In contrast, more sophisticated methods use ML to infer missing values in a way that is consistent with temporal patterns and any available exogenous data. For example, WGAN-LSTM imputation can ensure smooth fills closely matching true distributions. This is achieved by combining Wasserstein generative adversarial networks (GANs) that create realistic data predictions with LSTMs capable of ensuring temporal consistency in data [Fang et al., 2023]. Through this, the method aims to supply reasonable estimates for missing points that maintain continuity and prevent losses in training accuracy. By imputing strategically, one can prevent loss of valuable training segments and avoid bias that would occur from dropping data entirely. Alternatively, fully probabilistic models (e.g. GPs) offer an elegant and principled way to handle missing data by marginalising over the uncertainty in missing inputs during inference [Rasmussen and Williams, 2006].

Resampling and Alignment: To reconcile differences in temporal granularities, data are often resampled to a common timeline. For example, high-frequency PV series may be aggregated to match hourly weather forecasts, or NWP outputs may be interpolated to finer intervals [Di Leo et al., 2021]. Aligning timestamps ensures that input features like irradiance forecasts will correspond to the correct PV output time. By carefully resampling data in this manner, models can learn the correct temporal relationships at the intended forecast frequency without aliasing. Additionally, coarse temporal resolutions in weather data that fail to capture rapid fluctuations can be augmented using specialised downscaling or "temporal super-resolution" methods. For instance, [Inman et al., 2013] suggest that hourly NWP data can be interpolated or supplemented with high-resolution historical observations as training data to achieve highly augmented quasi-minute predictions.

Outlier Detection and Smoothing: Forecasting models are typically sensitive to large errors in their training datasets [Liu et al., 2012]. To minimise their im-

pact, data cleaning routines are often used to identify anomalous data points deviating from the expected physical limits set manually or statistical patterns determined by previous data. They enable outliers (for example, a sudden spike beyond the panel's maximum capacity, or negative irradiance values) to be clipped, removed, or replaced with interpolated values. Techniques such as z-score filtering, or robust statistical methods like median-based filters, are frequently used to automatically detect and flag outliers for data cleaning. For example, [López Gómez et al., 2020] automatically removed all rows in their dataset containing null or duplicate values, then referenced sunrise/sunset times to manually correct misalignments in timestamps caused by this filtering. This realignment is often essential in daily forecasting to ensure that measurements continue to align with expected patterns driven by solar cycles. By ensuring that the training data are free from obvious errors, models can focus on learning genuine PV behaviour without having to account for the impact of noise that would otherwise reduce prediction accuracy.

Handling cyclical variables: To handle diurnal and seasonal cyclicity, transformations or features are often introduced into PV power data that make these cycles easier to use for training of predictive models. Time-of-day and seasonal indicators of time variables are commonly encoded as cyclical variables using sine/cosine transforms, making data continuous so that a model can easily differentiate patterns linked to differences between morning vs. afternoon, or winter vs. summer [Siddiqui et al., 2019]. Another powerful technique is normalising by clear-sky models. This process effectively removes the majority of predictable variations in diurnal or seasonal trends through a multiplicative adjustment. In practice, one can divide the observed PV power or irradiance by the expected clear-sky value (the maximum theoretical irradiance at that moment), yielding a

clear-sky or clearness index. This index represents the fraction of potential power actually generated and ranges from zero to one, where one represents clear sky conditions and lower values represent tconditions affected by cloud cover [Lauret et al., 2022]. By transforming raw power from forecasting data into a clear-sky index that remains approximately stationary and bounded during daylight hours, seasonal drift is reduced and strong contributions of diurnal cycles are largely factored out. This multiplicative de-seasoning is therefore commonly used in solar forecasting preprocessing steps, as it allows models to focus on the effects of weather-induced variability [Boland, 2015].

Similarly, additive de-seasoning can be achieved by simply subtracting a precomputed seasonal trend or long-term average from cyclical data. For example, the mean daily profile of PV power datasets can be found by computing the average value at each timepoint over multiple days, then fitting a Fourier series on this data [Boland, 2020]. Subtracting the predicted daily curve from observed daily data leaves only residual fluctuations driven by weather-based variation or other stochastic components for the model to predict. In summary, features like clear-sky index, sun elevation, or seasonal Fourier terms encoding a periodic structure can help models to account for cyclical trends in data. Doing so improves generalisation across different times and facilitates small adjustments based on additional variables like weather.

Normalisation and Scaling: As previously mentioned when discussing the application of a clear-sky index to data, normalising input and output variables to a dimensionless or standardised range has many benefits. For example, dividing PV power by the installed capacity yields a ratio which is easily comparable across sites and easier for models to handle. Similarly, weather features like temperature or irradiance may be normalised on a scale of zero to one, ensuring that no single

feature dominates over others due to unfair weighting of units. Normalisation can also involve de-trending long-term changes, like gradual PV degradation, so that the model isn't misled by non-stationary means.

Overall, the data preprocessing techniques mentioned can greatly enhance dataset quality; [Leo et al., 2025] go as far as describing normalisation, interpolation and outlier removal as vital preprocessing steps when building forecasting models. However, the power of these techniques to heavily influence data and predictions means that care must be taken to ensure they are applied appropriately.

1.2.2.2 Modelling Architectures

In parallel to preprocessing data, researchers commonly design forecasting models that are inherently better suited to cope with data imperfections and uncertainties not removed by these corrective measures:

Probabilistic Models: Rather than a single deterministic output, probabilistic models output a distribution of PV power predictions. This approach naturally accounts for uncertainty caused by volatility in weather or data noise and reduces the model's sensitivity to outliers. A popular example is Gaussian process regression (GPR), a Bayesian non-parametric method particularly well suited for PV training data due to its effective handling of non-linear relationships, uncertainty estimation and probabilistic forecasts [Islam et al., 2024]. In addition, GPR models can seamlessly handle irregular time steps commonly found in PV datasets, as predictions can be made for any arbitrary time input using the learned covariance function. This also means that missing observations or gaps in data do not pose a problem beyond informing the posterior distribution.

Other commonly used probabilistic approaches include quantile regression, which predicts PV outputs in various distinct quantiles, and Bayesian neural networks.

These forecasting models are typically more robust to data anomalies because they identify and assign outlier values as low-probability events, limiting their ability to skew future predictions.

Ensemble Learning: Ensemble methods improve robustness when handling heterogeneous or imperfect datasets by combining multiple different models or learners to limit the influence of individual errors. By using a diverse collection of models, errors produced by a single model can more easily be identified and removed, or their impact can be reduced after averaging with the outputs of other models. The spread of predictions made by different models can also be used to help quantify uncertainty. In the context of PV power prediction, many ensemble techniques can be used to obtain a more stable forecast.

For example, boosting techniques sequentially train models on the same data and give incorrectly predicted data points more weight in subsequent runs. By doing so, models can put more weight on challenging predictions and correct previously made mistakes in a continuous process of iterative refinement.

Unlike boosting, bagging methods involve training each model on different subsets of the same dataset before combining predictions to reduce variance and overfitting. These methods are commonly applied to decision trees like random forests. Alternatively, stacked generalisation can be used to build meta-models on the predictions of multiple other model predictions. Assembled models combine the beneficial aspects of their constituent models to make more accurate predictions than would be possible with each model individually. For example, [Lateko et al., 2022] develop a stacked ensemble of several algorithms (eXtreme gradient boosting (XGBoost), multilayer perceptron neural networks, etc.) for short-term PV forecasting that achieved significantly higher accuracy than all single-model comparisons.

Models with Built-in Missing Data Handling: Rather than relying on comparisons with other models, some algorithms have been specially designed to independently handle missing data inputs. For example, recurrent neural networks and state-space models can be adapted to continue their learning and predicting processes when applied to datasets that have missing values, without additional user input. This is typically achieved by either inferring the values of missing data points, or skipping them entirely. [Xiang et al., 2024] recently introduced an LSTM specifically designed for PV forecasting that is highly resilient and tolerant of missing data. In their approach, a recurrent neural network was trained to impute missing PV values recursively as part of the sequence prediction. This enables the model to substitute its own previous forecast in place of a missing value at any time step and to continue the sequence. By using a negative log-likelihood loss function to iteratively refine both the imputation and prediction during training, the model is able to both fill gaps and forecast simultaneously.

Beyond LSTMs, GPs can naturally handle missing data in inputs by marginalising those dimensions, and state-space models like Kalman filters can simply omit the update step when observations are missing. Architectures can therefore explicitly account for missing data to prevent the catastrophic failure that might occur if a standard model encounters a null input. This is achieved either by design, as seen in GPs or Kalman filters [Li et al., 2022], or by training, as seen in the RecLSTM imputation model [Liu et al., 2021]. Both ensure that forecasting can continue gracefully even with incomplete data.

Hybrid models: Finally, hybrid models that combine physical modelling with ML are becoming increasingly popular due to their ability to capitalise on domain knowledge while remaining data-driven [Rudolph et al., 2024]. Hybrid models may use physics-based components, such as a clear-sky PV power calculation

or an equivalent circuit PV model driven by NWP weather forecasts. Statistical models are often then applied to correct any systematic errors made by the physical model. This can be done in an additive way, where error terms are learned and then added or subtracted. For example, one could predict an additive residual $\Delta P_{\text{stat}}(t)$ such that:

$$\hat{P}(t) = P_{\text{phys}}(t) + \Delta P_{\text{stat}}(t) \tag{1.1}$$

As additive approaches rely on absolute error correction, they are only suitable when the bias or error does not scale strongly with the magnitude of P_{phys}.

Alternatively, the output of a physical model $P_{\text{phys}}(t)$ can be adjusted by a multiplicative ML factor $F_{\text{stat}}(t)$ that accounts for clouds and other losses, forming a multiplicative hybrid structure that yields the forecast:

$$\hat{P}(t) = P_{\text{phys}}(t) \cdot F_{\text{stat}}(t) \tag{1.2}$$

In contrast to additive approaches, multiplicative approaches estimate a relative correction based on physical prediction. This makes them better-suited to PV forecasting, in which datasets often contain errors that scale proportionally with the signal (such as errors caused by inverter clipping or shading). Despite being non-stationary, the underlying patterns of errors can be learned by adaptive ML models over time. Moreover, multiplicative hybrid models ensure physical consistency; using a clear-sky index or the output of the physical model as a baseline can remove bulk seasonal effects and allow the machine learner to focus on weather-induced variability [Gargalo et al., 2024].

Model training techniques: In addition to model architecture, the model training process itself can be made robust to outliers and noise. These training approaches can be implemented to complement robust model design and handle

particularly challenging data scenarios. For example, using a Huber loss or quantile loss instead of mean squared error can lessen the influence of large outliers that would otherwise skew training data [Huber, 1964]. Some studies deliberately employ data augmentation or noise injection during training to make the model tolerant to irregularities [Chen et al., 2024]. When adapting models to new sites with short historical records, transfer learning or meta-learning can be used to handle the limited amount of data.

1.3 Contributions

In this thesis, we propose a novel multiplicative hybrid modelling approach that integrates a simple deterministic physics-based PV model with a data-driven correction factor. Specifically, a GP serves as a residual post-processing component to enhance the output of the deterministic model by taking into account site-specific influences, system losses, uncertainties in weather forecast, and temporal dependencies. The multiplicative combination of the physical model and ML correction enhances both forecast accuracy and robustness, compared to conventional physical or purely data-driven models. Furthermore, GPs provide a principled probabilistic framework capable of modelling non-linear relationships and uncertainty through appropriate kernel functions. Thus, this framework offers an effective solution to the challenges outlined in the last section, especially the challenges related to data limitations and model adaptability. The three key contributions of this thesis are as follows:

1. Novel hybrid model for day-ahead solar forecasting combining a deterministic model with GPs. Two GP approaches are explored: an autoregressive GP and a time-series GP.

2. GP performance optimised through kernel selection and hyperparameter

tuning using both the no-U-turn sampler (NUTS) for Bayesian inference and maximum likelihood estimation (MLE) for point estimates.

3. Comprehensive evaluation of the proposed hybrid model against fifteen benchmark models, including eight statistical post-processing methods and seven direct forecasting models. The benchmarks cover physical, statistical, and commercial approaches. The evaluation was performed using three datasets from HK to the UK.

The remainder of this thesis is organised as follows:

- **Chapter 2**: Introduces the three datasets used to evaluate the proposed model and benchmark methods, including a household PV system in Oxford, small to medium-scale PV sites in HK, and a range of PV systems across various locations in the UK.
- **Chapter 3**: Gives an overview of the hybrid approach and introduces the deterministic physical model, comparing prediction performance before and after clipping to a simpler clear-sky model.
- **Chapter 4**: Discusses the suitability of machine learning, statistical and neural network-based approaches for predicting PV output, then evaluates their performance across UK and HK sites.

2 | Data

2.1 Experimental Data

This chapter provides an overview of the datasets used to evaluate the proposed model and benchmark methods. Three sets of PV output data are used in this thesis: (1) a household PV system in Oxford, (2) small to medium-scale PV sites in HK, and (3) a diverse range of household/small PV systems across multiple locations in the UK. Each PV dataset is accompanied by corresponding irradiance and weather data, sampled at the same temporal resolution. An overview of these datasets is presented in Table 2.1. The datasets were obtained from different sources, as specified in Table 2.2.

Irradiance and Weather Data: Typically, NWP models such as those provided by ECMWF [European Centre for Medium-Range Weather Forecasts, 2023] and the Global Forecast System [National Centers for Environmental Prediction (NCEP), 2020] serve as inputs to a model framework for predicting PV power output. However, NWP forecasts are limited in spatial resolution and are not extensively archived, as they are designed mainly for forward-looking use rather than long-term historical analysis. Since the focus of this study is to develop a ML model to enhance the PV power predictions generated by a physical model, we instead use historical irradiance and weather data as inputs to the deterministic model.

Historical irradiance data are obtained from Solcast [Solcast, 2022] and are derived from high-resolution imagery captured by various geostationary meteorological satellites. These data are processed using advanced cloud detection models and retrieval algorithms. The corresponding historical weather data are sourced from the ECMWF ERA5 atmospheric reanalysis dataset [European Centre for

Table 2.1: Overview of datasets used for model evaluation

Name	**Category**[2]	**Peak Capacity (kW)**	T_{physical}[3]	**Resolution**	T_{ML}[4]	**Metadata**
Oxford[1]	Small	3.96	109,585	10 min	761	Yes
HK Site A	Small	2.04	19,994	1 h	833	No
HK Site B	Small	2.93	18,493	1 h	770	No
HK Site C	Small	5.00	9,052	1 h	377	No
HK Site D	Medium	28.99	15,514	1 h	646	No
HK Site E	Medium	28.95	23,972	1 h	999	No
HK Site F	Medium	24.66	33,958	1 h	1,415	No
UK Site 1	Small	2.81	34,368	30 min	716	Yes
UK Site 2	Small	2.85	34,223	30 min	713	Yes
UK Site 3	Small	2.98	34,777	30 min	724	Yes
UK Site 4	Small	3.36	33,503	30 min	698	Yes
UK Site 5	Small	2.97	34,223	30 min	713	Yes

[1] The Oxford site is used as a running example throughout the thesis.
[2] Category is defined by peak capacity: Small (≤ 5 kW), Medium (5–30 kW). Note that peak capacity (P_{peak}) is the clipped maximum output, distinct from nominal capacity (P_{nominal}), the site's rated output.
[3] T_{physical} denotes the number of time steps in the PV power output time series. Although irradiance and weather data are sampled at the same resolution, they are excluded from this count.
[4] denotes the total number of time steps in the ML time series, evaluated at a daily resolution. The ML time series represents all training and testing data points used at that site (each of which make up half of the total ML time steps).

Table 2.2: Data sources for irradiance, weather, and PV power output time series.

Dataset	**Source**	**URL**
Irradiance and weather	Solcast	https://solcast.com
Oxford	Internal	Not publicly available
HK sites	HKUST	https://datadryad.org/dataset/doi:10.5061/dryad.m37pvmd99
UK sites	Open Climate Fix	https://huggingface.co/datasets/openclimatefix/uk_pv

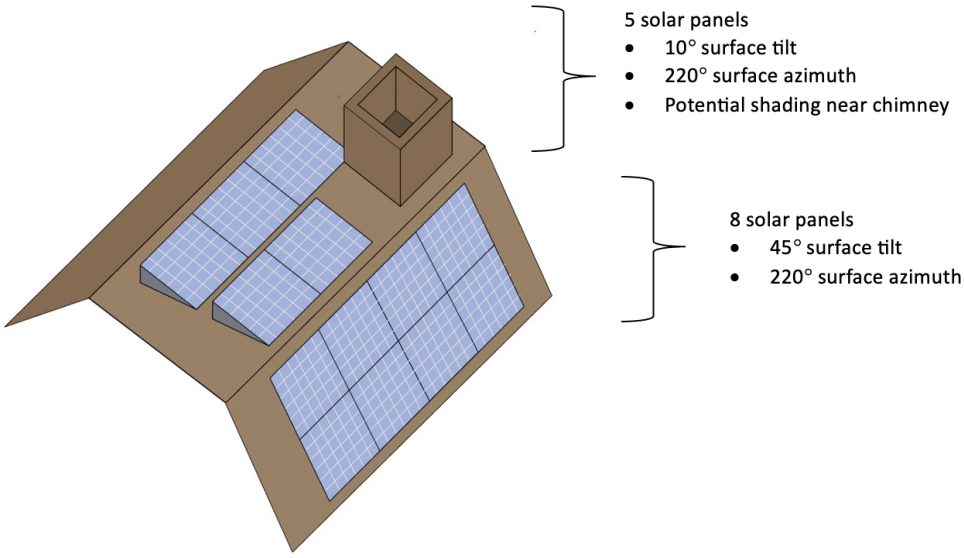

Figure 2.1: Schematic diagram of Oxford household PV site, which is used as a representative example throughout this thesis.

Medium-Range Weather Forecasts, 2017]. For each PV site, the irradiance and weather time-series data are aligned with the PV output data and sampled at the same temporal resolution. For example, if the PV power output is recorded at 10-minute intervals, the corresponding irradiance and weather inputs are also sampled at 10-minute resolution.

Oxford Household PV System Data: The PV power output data from the Oxford site were collected from a residential rooftop PV installation (51.752° N, 1.258° W). A schematic diagram of the system is shown in Figure 2.1. The site consists of 13 solar panels: five panels are installed with a tilt angle of 10° and eight with a tilt of 45°, all oriented at a surface azimuth of 220°. Each panel has a nominal capacity of 385 W, yielding a total nameplate capacity of 5.005 kW. However, due to inverter clipping, the effective peak output is approximately 3.96 kW. The PV power output was recorded in 5-minute intervals from 25th

March 2023 to 24th April 2025. PV output data was converted to 10 minutes by resampling and alignment to account for gaps in data, as well as to match the highest resolution weather data provided by Solcast for use in the proposed deterministic model.

Hong Kong PV System Data: The PV power output data from HK were obtained from a dataset provided by the Hong Kong University of Science and Technology (HKUST). The dataset consists of 60 rooftop PV stations connected to the grid located in the Sai Kung District, a rural coastal area of HK (22.336° N, 114.263° E) [Lin et al., 2024]. Acquired data span a three-year period from 2021 to 2023, with measurements recorded at an hourly resolution and ranging in peak capacity from 2.04 kW to 28.99 kW. For this study, a subset of representative PV sites was selected from this dataset to evaluate the proposed hybrid model and benchmark methods.

UK PV System Data: All PV power output data from the United Kingdom was sourced from the Open Climate Fix dataset [Open Climate Fix, 2022], which compiles generation records from over 30,000 distributed PV systems across Great Britain. Of these systems, most are small-scale residential installations with maximum capacities between 0.47 kWp and 250 kWp. Energy output (measured in Wh) sampled between 2010 and early 2025 was recorded in 30-minute intervals, consistent with the UK electricity market's settlement, and multiplied by 2 to obtain average power in Watts. Each PV system entry contains metadata that includes geographical location, maximum capacity, tilt angle, and azimuth. In addition to the standard 30-minute resolution, roughly 1,300 systems also include higher-frequency data sampled at 5-minute intervals that were multiplied by 12 to calculate average power in Watts. For the purpose of this study, a subset of systems was selected from diverse geographical locations to assess the performance

of the proposed hybrid forecasting framework compared to benchmarks.

2.2 Evaluation Metrics

In this section, we describe the metrics used to evaluate the performance of PV power and energy forecasts. For the test set, we adopt the mean absolute error (MAE), RMSE, and mean absolute percentage error (MAPE), defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|, \tag{2.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}, \tag{2.2}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|, \tag{2.3}$$

where $\hat{y}_i$ is the predicted value, y_i the observed value, and n the number of samples. While MAE and RMSE measure absolute and squared deviations, respectively, MAPE provides a percentage error measurement that is useful when comparing across various PV sites.

Where system metadata are available, the normalised root mean squared error (nRMSE) that scales RMSE by the system's nominal capacity P_{nom} is reported:

$$\text{nRMSE} = \frac{\text{RMSE}}{P_{\text{nom}}} \times 100\%. \tag{2.4}$$

3 | Overview of Modelling Approach

This section introduces the hybrid forecasting framework in full. The approach begins with solar irradiance and weather inputs, together with any necessary assumptions or approximations where data are missing. These inputs feed into a deterministic physical model that converts irradiance into an initial estimate of PV power output.

To capture variability not fully explained by the physical model, such as changes in seasonality, weather, and site-specific conditions, machine learning models are applied. In particular, GPs are trained to learn adjustment factors between observed and predicted outputs. When combined multiplicatively with the physical model, these adjustments yield improved prediction accuracy and are well-suited to correcting PV-related errors such as inverter clipping and shading.

For context, the underlying physical principles are briefly summarised. PV cells generate electricity by converting incident irradiance into direct current that is then transformed into alternating current by the inverter. Efficiency is influenced by material properties (e.g., semiconductor doping), environmental conditions (e.g., temperature, irradiance), and system-level factors (e.g., ageing and inverter design). Physical models attempt to capture these processes at varying levels of complexity: advanced models can offer higher accuracy but require detailed inputs, while simpler models rely on generalised assumptions.

Our framework adopts a deterministic physical model as the baseline. This balances simplicity and data availability with predictive strength, while the integration of GPs provides the flexibility needed to outperform stand-alone physical or machine learning models.

3.1 Multiplicative Hybrid Approach

As introduced in Section 1.1, the general procedure for day-ahead solar forecasting involves converting irradiance and weather forecasts into PV power forecasts via a physical 'model chain'. Traditionally, this approach is deterministic, meaning it produces a single forecast outcome based on the input data. Prior to irradiance-to-power conversion through the model chain, statistical models can be applied to post-process the NWP-based irradiance forecasts to correct biases or improve accuracy. Additionally, or alternatively, post-processing can be applied to PV power forecasts after the irradiance-to-power conversion. In this thesis, we focus on day-ahead power and energy predictions, where the energy forecast refers to a single daily total (in kWh). These predictions are generated once per day by applying ML techniques, such as GPs, to post-process the model chain's predictive power output.

In the proposed model, the deterministic model first forecasts the baseline PV output for each site. For the Oxford installation, this involves predicting the output of the two panel configurations (tilt 10° and 45°, azimuth 220°) then multiplying by the number of panels (5 and 8, respectively) to obtain an initial forecast. For HK sites lacking metadata we assume a standard tilt of 10° and azimuth of 180°. We estimate the number of panels by dividing the system's nominal capacity by a typical panel rating of 330 W. The nominal capacity is inferred from the time-series data by identifying the peak power output and dividing by a performance ratio of 85% [Marion et al., 2005]. The model chain then uses these assumptions along with irradiance, weather data, and panel information including location, tilt, azimuth, and panel count to generate the initial deterministic forecast.

Next, GPs are applied to predict the ratio between the actual observed PV power

and the predicted model chain output, forming the basis of the multiplicative hybrid model:

$$P_{\text{final}}(t) = P_{\text{model chain}}(t) \times \delta_{\text{GP}}(t) \tag{3.1}$$

where $\delta_{\text{GP}}(t)$ is the time-series correction factor predicted by the GP. Other models, such as additive and mixed additive-multiplicative models, were also tested but are not discussed further in this chapter due to their comparatively poor performance.

The motivation for using GPs to model the observed-to-predicted adjustment ratio is due to the significant variability and seasonal patterns observed in the time-series data, as illustrated in Figure 3.1. For example, in the case of the Oxford site, the adjustment factor is not consistently centred around 1, but instead exhibits pronounced volatility, particularly during periods of cloud cover or rapid weather transitions. This variability arises not only from weather impacts but also from factors such as inverter clipping, shading, soiling, and other site-specific influences that are not fully accounted for by the deterministic model. These unmodelled effects introduce non-linear and time-varying interactions between environmental inputs and PV output. Furthermore, the deterministic model is highly sensitive to the selection of sub-models used for irradiance separation, transposition, and other intermediate steps. Simpler configurations, such as single-run deterministic models, often exhibit higher bias and reduced forecast dispersion compared to ensemble-based approaches [Mayer and Gróf, 2021]. In practice, the observed PV power can deviate considerably from deterministic forecasts. Conditioning the GP on the entire training set allows it to generalise over noisy fluctuations while minimising the influence of short-term outliers. Therefore, GPs offer a flexible approach to dynamically correct the physical model output by learning from historical deviations and accounting for time dependencies.

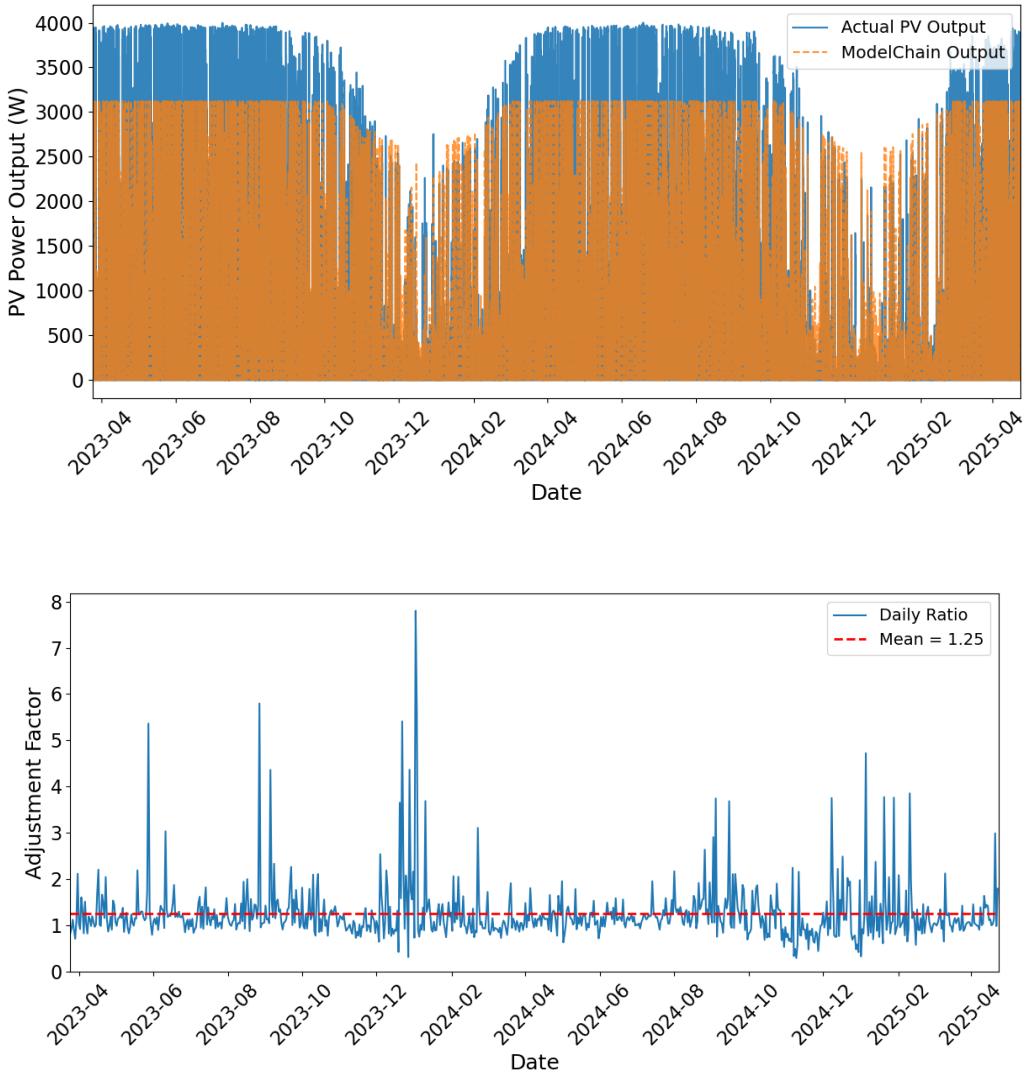

Figure 3.1: (Top) Comparison of physical model predictions and observed PV power output. (Bottom) Corresponding adjustment factors computed as the ratio between observed and predicted PV values. Red dashed line represents the mean adjustment factor value across all data (mean=1.25). These figures are based on the Oxford dataset.

Figure 3.2 presents a detailed flowchart of the proposed hybrid forecasting approach. The adjustment ratio is initially computed by dividing the observed PV power output by the corresponding deterministic model prediction at each time step, excluding all zero values and predicted values below 0.01 W. Although the native resolution of the PV datasets varies: 10-minute intervals for the Oxford site, 1-hour intervals for HK sites, and 5- or 30-minute intervals for UK sites, all adjustment ratios δ_{GP} are aggregated to a single value per day by averaging the valid ratios from 00:00 to 23:59 local time. These daily averaged ratios form the time series used to train the GP models. The dataset is divided into training and testing periods, where the GP is trained on the historical adjustment ratios to predict the next day's ratio at 00:00 local time. As described in Chapter 4, the predicted ratio is then used to scale the deterministic model output, effectively refining the day-ahead forecast by correcting for systemic and time-dependent biases.

3.2 Physical Models

This section outlines the operating principles and key characteristics of PV panels, as well as the physical models used to forecast their output. To begin, the PV conversion process is summarised, showing how solar cells generate electrical current from incident irradiance and convert direct current (DC) into alternating current (AC). Next, the main environmental and system-level factors affecting PV efficiency are described, including temperature, shading, and inverter behaviour. The section then introduces the modelling approaches used to simulate PV generation. Deterministic physical models are explained, followed by probabilistic post-processing methods that refine predictions by accounting for uncertainties. Finally, the performance of deterministic AC models is assessed across UK and HK PV sites and compared against deterministic DC models and clear-sky base-

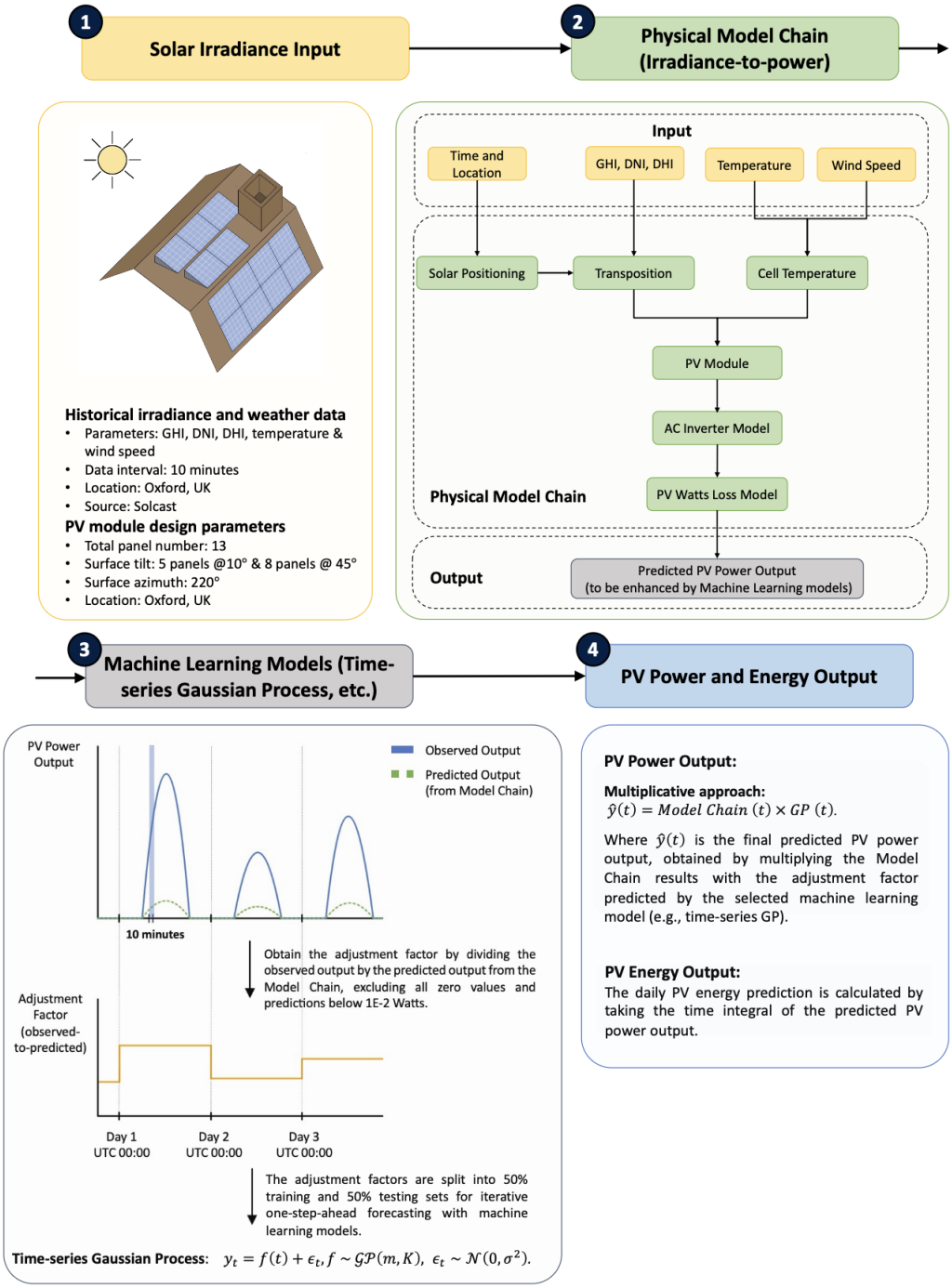

Figure 3.2: Process flow of hybrid forecasting framework: (1) Solar irradiance and weather data are collected; (2) A physical model generates baseline PV power predictions; (3) A GP model learns a daily adjustment factor from observed-to-predicted ratios to refine the final forecasts.

lines. This analysis highlights the influence of inverter clipping, irradiance inputs, and system losses on forecast accuracy.

3.2.1 PV Panel Characteristics

The cells of solar panels rely explicitly on the excitation of electrons by energy directly transferred from sunlight to create DC electricity. In insulators, the bandgap is too large for solar photons to excite electrons across it, while in conductors the absence of a bandgap prevents efficient generation and separation of charge carriers. Semiconductors such as monocrystalline or polycrystalline silicon have a bandgap in the optimal range of 1-1.6 eV, allowing photons in the solar spectrum to excite electrons from the valence to the conduction band, thereby generating usable charge carriers for photovoltaic conversion [Pastuszak and Węgierek, 2022]. Excitation of these electrons from light enables the photovoltaic effect to occur: the energy packet transferred from a single photon releases an electron that it contacts, creating an electron-hole pair and allowing a current to flow. By doping semiconductor material with impurities that either provide excess mobile electrons ("n-type" elements) or are electron-deficient "p-type" elements), p-n junctions can be made that generate electric fields and promote the directional flow of light-generated current towards an external circuit, overcoming any concentration gradient in the process [Al-Ezzi and Ansari, 2022] (Figure 3.3). The consistent uni-directional flow of excited charge carriers generates DC, which is collected by a transparent conducting front electrode and a reflective back electrode, then passed along a closed circuit. The DC generated by this process can be converted to AC using solar inverters for later use.

Whereas the majority of AC electricity transferred to homes for use in appliances is typically generated by alternators relying on rotating magnetic fields to reverse current flow, DC electricity from solar panels is converted to AC using invert-

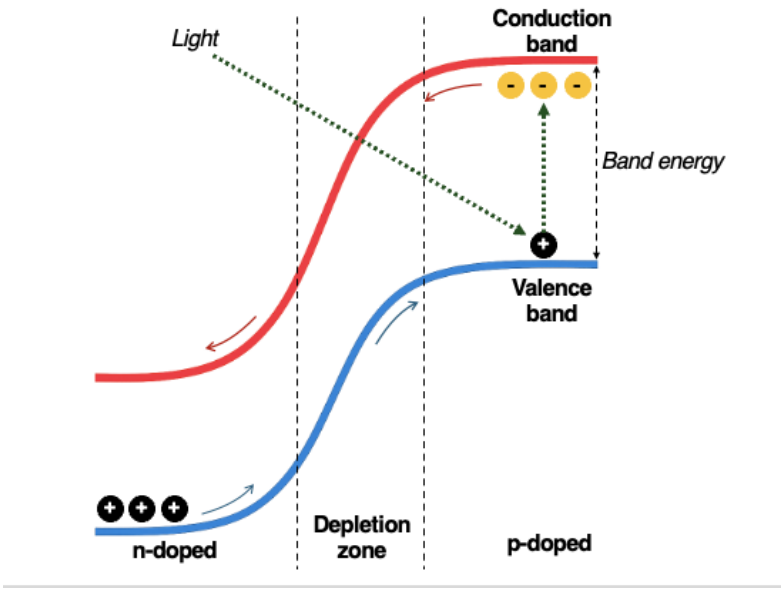

Figure 3.3: Schematic diagram of solar cell p-n junction. P-doped regions contain an excess of positively charged 'holes', while n-type regions contain excess electrons. Incident light excites electrons sufficiently to exit the valence band and enter the conduction band, leaving behind holes and generating a potential difference through charge separation. Arrows along the conduction and valence bands show the directional flow of electrons and holes in the electric field through drift, opposed by counteracting diffusional forces in the opposite direction. Metal contacts complete the circuit and allow current to flow, with electrons later returning to holes after completing work.

ers. These rapidly switch DC voltage on and off with transistors that flip polarity to produce a current waveform. Waveforms can vary in profile, so are typically filtered through capacitors or inductors to produce cleaner, ready-to-use AC sine waveforms with sufficiently low total harmonic distortion for use [Abbas et al., 2021]. Inverters are rated to a maximum AC output capacity; if the DC input exceeds this capacity, for example due to favourable conditions causing spikes in electricity production, this energy is wasted and efficiency is reduced [Micheli et al., 2024].

In addition to limitations imposed by equipment, the efficiency of solar cells in converting solar energy to electrical current is subject to many factors that may vary with weather, location, or age. For example, temperature plays a critical role in the voltage output of semiconductors. At higher temperatures, the increased random motion of electrons leads to higher internal resistance that reduces voltage for the same given current [Löper et al., 2012]. Electrons in hotter environments may also be more likely to recombine with electron-deficient atoms before effectively contributing to the current [Shaker et al., 2024]. The quantifiable effects of thermal agitation on efficiency are relatively linear, with every 1 °C increase above room temperature reportedly reducing the efficiency of crystalline cells by 0.25% [Al Siyabi et al., 2019]. In contrast, the convective effect of wind can increase PV efficiency by reducing the surface temperature of solar cells and removing obstructive debris that would otherwise cause soiling [Csavina et al., 2014]. In addition to environmental variables affecting temperature, the direct amount and intensity of sunlight reaching solar panels is highly variable and may be reduced in the short-term by obstructions causing shading throughout the day. Equally, irradiance may decrease more permanently through the effects of soiling [Menoufi, 2017]. Shading and soiling also cause panels to be irradiated non-uniformly, further impacting productivity.

The listed factors represent only a subset of some of the most commonly discussed influences impacting PV efficiency. For example, [Kazem and Chaichan, 2015] reported that humidity negatively impacts efficiency more than all other parameters measured, including temperature and soiling, through both temperature- and irradiance-related effects. As the physical factors mentioned have the potential to massively constrain PV energy conversion efficiency, forecasting models that account for their influence can greatly improve power output predictions.

3.2.2 Physical Model Chain

As introduced in Section 1.1, physical model chains output PV power by sequentially applying different physics-based models, with each stage adding an incremental level of accuracy to the predicted output. This modelling approach typically includes the following steps:

1. **Solar positioning:** Calculate solar zenith and azimuth angles from time and location using the Reda & Andreas algorithm [Reda and Andreas, 2004].

2. **Clear-sky modelling:** Compute idealised clear-sky GHI, DNI, and DHI for example using the Ineichen model [Ineichen and Perez, 2002], used in this thesis as a benchmark for maximum possible irradiance under cloudless conditions.

3. **Irradiance transposition:**Convert GHI, DNI, and DHI (either from measured/forecasted data or from the clear-sky model) to plane-of-array (POA) irradiance using a diffuse sky model such as Hay-Davies [Hay and Davies, 1980].

4. **Module modelling:** Estimate DC output from POA irradiance and cell temperature using a PV module performance model (e.g., Sandia, CEC, PVWatts), incorporating thermal and soiling losses.

5. **Inverter conversion:** Convert DC to AC output by applying an inverter model (e.g., Sandia, ADR, PVWatts) and accounting for balance-of-system efficiency.

Several libraries implement physical PV model chains, among which `pvlib`'s `ModelChain` class provides a modular and standardised interface to compute PV power from a weather time series [Holmgren et al., 2018]. A schematic diagram of the deterministic model used in this study was presented in Figure 3.2.

The workflow for this model requires system-specific inputs, including GHI, wind speed, air temperature, module and inverter parameters, tilt, and azimuth. Time and location are also required to compute sun-position variables such as zenith and incidence angles using the algorithm proposed by [Reda and Andreas, 2004].

In a typical `pvlib` implementation, the clear-sky irradiance is first estimated using the Ineichen model [Ineichen and Perez, 2002], which calculates idealised GHI, DNI, and DHI from site pressure, Linke turbidity, and solar geometry.

$$\mathrm{DNI}_{\mathrm{cs}} = I_0 \cdot e^{-\frac{\tau_b \cdot m_a}{f_b}} \tag{3.2}$$

$$\mathrm{DHI}_{\mathrm{cs}} = \mathrm{GHI}_{\mathrm{cs}} - \mathrm{DNI}_{\mathrm{cs}} \cdot \cos(\theta_z) \tag{3.3}$$

$$\mathrm{GHI}_{\mathrm{cs}} = \mathrm{DHI}_{\mathrm{cs}} + \mathrm{DNI}_{\mathrm{cs}} \cdot \cos(\theta_z) \tag{3.4}$$

where I_0, τ_b, m_a, f_b, and θ_z are parameters describing extraterrestrial irradiance, atmospheric transmission, and solar position. In this study, the clear-sky model serves as a simple and idealised benchmark scenario, passed through the same `pvlib` model chain as the AC and DC forecasts to enable a like-for-like performance comparison under maximum possible irradiance conditions.

In cases where DNI and DHI are not provided, separation models are used to

derive these parameters from GHI. Since both the Solcast data and the clear-sky irradiance from the Ineichen model used in this thesis already include these components, no separation step was required. The Hay–Davies model is then used for transposition to POA irradiance [Hay and Davies, 1980]. Cell temperature is estimated using the Sandia module model [Marion et al., 2005], which accounts for incident irradiance, ambient temperature, and wind speed. DC output is then computed by applying one of several module performance models, such as the single-diode, CEC, or PVWatts formulations, to the dependence:

$$P_{\text{DC}} = f(E_{\text{POA}}, T_{\text{cell}}, \theta) \tag{3.5}$$

where f is the chosen performance model, E_{POA} is the POA irradiance, T_{cell} is the cell temperature, and θ is the incidence angle. AC output is obtained by applying an inverter model (e.g., Sandia, ADR, or PVWatts) to the DC power:

$$P_{\text{AC}} = \eta_{\text{inv}} \cdot P_{\text{DC}} \tag{3.6}$$

where η_{inv} is the inverter efficiency. It is worth noting that inverter efficiency is generally a non-linear function of input power, with performance degrading at both very low and very high loads. In this study, inverter conversion was modelled using the CEC inverter model available in `pvlib`, which captures these non-linear characteristics and provides a more realistic representation of partial-load efficiency and clipping behaviour.

To evaluate the sensitivity of AC results to inverter modelling, AC outputs were compared with the corresponding DC outputs prior to conversion. To isolate the impact of irradiance and system losses, the clear-sky model, representing the theoretical maximum output, was processed through the same deterministic workflow, providing a baseline benchmark for both AC and DC model comparisons.

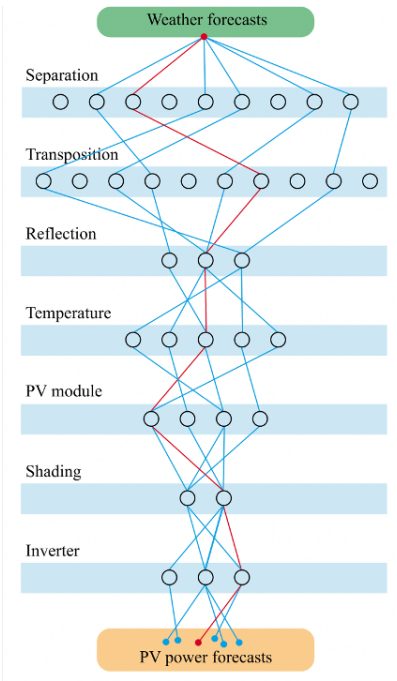

Figure 3.4: Schematic of ensemble PV power forecasting model [Mayer and Yang, 2023].

3.2.2.1 Deterministic and Ensemble Model Chains

Deterministic models typically generate a single best-guess output based on fixed input parameters and model settings, which may not adequately represent the inherent uncertainties in weather predictions and PV system behaviours. In contrast, ensemble model chains use multiple sets of conditions or parameter variations to produce a range of potential outcomes, thus offering a probabilistic forecast that accounts for various sources of uncertainty. For instance, an ensemble might consist of forecasts from K different weather model members or from varying system configuration assumptions. The resulting set of forecasts $\{f_k(t)\}_{k=1}^{K}$ allows for summary statistics and probabilistic post-processing. The setup of an ensemble of model chains is shown in Figure 3.4.

Ensemble methods are motivated by the intrinsic uncertainties in meteorology and PV modelling. Uncertainty in initial conditions, weather variability, and sys-

tem characteristics makes any single forecast potentially biased [Chahboun and Maaroufi, 2021]. Using ensembles improves accuracy and enables the forecast to quantify uncertainty through prediction confidence intervals. In order to do so, raw ensemble outputs are post-processed using statistical techniques. Common methods include ensemble model output statistics, quantile regression forests (QRFs), and Bayesian model averaging (BMA).

Ensemble Model Output Statistics: The EMOS approach proposed by [Gneiting et al., 2005] is one of the most widely used post-processing methods for ensemble setups. EMOS assumes that the PV power output P_{PV} follows a parametric probability distribution, typically a left-truncated normal distribution. In a left-truncated normal distribution, the forecast distributions are truncated at zero, and the probability mass of the negative values is redistributed to the positive values. This is to ensure that the modelled power output remains non-negative. Let $\bar{f}$ denote the ensemble mean and S^2 the ensemble variance, then the output power is:

$$P_{\text{PV}} \sim \mathcal{N}_{\geq 0}\left(a + b\bar{f},\ c + dS^2\right) \tag{3.7}$$

where a, b, c, d are coefficients estimated by minimising the continuous ranked probability score (CRPS) over a historical training set. This calibration step adjusts the ensemble forecast for both bias and underdispersion. The closed-form expression for the CRPS of a truncated normal distribution is provided in [Jordan et al., 2019].

Quantile Regression Forests: Quantile regression forests (QRFs) are a non-parametric extension of random forests that predict the full conditional distribution of a response variable. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ denote the training data, where $\mathbf{x}_i$ are input features and y_i the corresponding PV power outputs. In the context of an

ensemble model chain setup, each $\mathbf{x}_i$ may include features such as individual PV forecasts $f_k(t)$, their mean $\bar{f}$, variance S^2 and etc. For a new input $\mathbf{x}$, the predicted cumulative distribution function (CDF) is given by:

$$\hat{F}(y|\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) \cdot \mathbf{1}\{y_i \leq y\} \tag{3.8}$$

In this case, $w_i(\mathbf{x})$ denotes the weight assigned to sample i. The indicator function $\mathbf{1}\{y_i \leq y\}$ equals 1 when the condition holds, and 0 otherwise. This function represents the estimated cumulative probability that the target PV power output will be less than or equal to y, given the input conditions $\mathbf{x}$.

To obtain the prediction intervals and the predictive mean, the estimated CDF can be inverted. In PV forecasting, QRF is widely used to produce calibrated uncertainty estimates, particularly under variable weather conditions [Meinshausen, 2006].

Bayesian Model Averaging: BMA is a probabilistic ensemble post-processing method that treats each ensemble member forecast f_k as a component in a weighted mixture model. Specifically, it assumes each forecast f_k is associated with a conditional density function $p(y \mid f_k)$, which is typically a Gaussian or truncated Gaussian centred around f_k. The predictive distribution is then computed as a weighted average of these densities:

$$\hat{p}(y) = \sum_{k=1}^{K} w_k \cdot p(y \mid f_k) \tag{3.9}$$

where $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$. These weights represents the posterior distributions of models and are estimated from historical data via MLE or expectation-maximisation methods.

For PV forecasting, studies such as [Gneiting et al., 2005] and [Raftery et al., 2005] demonstrate that BMA can outperform simpler ensemble averaging methods, particularly when ensemble forecasts exhibit systematic differences in quality or dispersion. Unlike EMOS, which fits a single parametric distribution, BMA forms a mixture of conditional distributions, allowing greater flexibility in capturing forecast uncertainty.

3.2.3 Results

To evaluate the performance of deterministic physical models (AC, DC, and clear-sky) in predicting power and energy outputs, datasets from multiple small PV sites in the UK and small- to medium-scale sites in Hong Kong's Sai Kung District are analysed. The selected sites enable assessment of model robustness across different climatic conditions, geographical contexts, and system configurations.

3.2.3.1 UK sites

Six small PV sites were selected in the UK, spanning latitudes 50° to 55° (see Table 2.2). While all lie within the UK's temperate maritime climate, the sites differ slightly in elevation, cloud cover, and coastal proximity to capture environmental variation [Benchimol et al., 2014].

Table 3.1 summarises power and energy prediction errors from AC, DC, and clear-sky models. Across all sites, the AC deterministic model consistently yielded the lowest MAE, RMSE, MAPE, and nRMSE values, with no clear influence from geography. Instead, lower nominal capacity sites generally achieved predictions closer to observed outputs, except at Oxford, where high-accuracy metadata likely improved model performance.

At the Oxford site with nominal capacity 5.005 kW, (Figure 3.5), the clear-sky

Table 3.1: Power and daily cumulative energy prediction performance across UK sites for the deterministic model with (AC) and without inverter conversion (DC), as well as the clear-sky model.

Model	Power			Day-ahead energy		
	MAE (W)	RMSE (W)	nRMSE (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
Oxford						
MC (AC)	286.30	652.93	11.32	3.82	5.15	52.07
MC (DC)	317.33	751.85	16.29	4.38	5.70	70.90
Clear-sky	1463.45	2033.71	50.83	17.191	18.240	544.93
Site 1						
MC (AC)	127.13	278.68	8.29	2.13	3.06	56.18
MC (DC)	136.03	286.36	8.52	2.44	3.41	59.90
Clear-sky	748.27	1315.56	39.15	18.02	19.54	502.93
Site 2						
MC (AC)	144.02	304.85	9.07	2.13	2.93	48.13
MC (DC)	151.57	317.25	9.44	2.44	2.36	52.73
Clear-sky	619.49	1110.74	33.06	16.24	17.94	401.33
Site 3						
MC (AC)	131.12	301.90	8.99	2.51	3.31	55.24
MC (DC)	146.89	317.41	9.45	3.00	3.95	65.80
Clear-sky	641.56	1127.71	33.56	16.76	18.18	802.72
Site 4						
MC (AC)	147.66	327.24	9.74	3.36	4.85	184.28
MC (DC)	220.50	457.85	13.63	4.22	5.72	224.63
Clear-sky	569.68	1007.81	29.99	13.68	14.91	1307.10
Site 5						
MC (AC)	174.31	390.15	11.61	3.07	4.79	195.24
MC (DC)	323.23	677.51	20.16	3.31	5.11	142.29
Clear-sky	657.62	1187.92	35.35	15.88	17.67	617.05

model that assumes optimal irradiance produced the highest predictions year-round, peaking in summer and dipping in winter. Both AC and DC models followed similar seasonality in predictions, but with consistently lower output as the AC model's clipping constraint reduced the effect of consistent overestimation by the DC model. Inverter clipping effects are also incorporated in the clear-sky model, preventing its predictions from ever reaching the nominal capacity. These trends are consistent across all UK sites (see Appendix A.1.1). While Oxford saw summer underprediction for both chains, DC models at other sites typically overpredicted. This may be due to differences in inverter size, panel setup, or the lower time resolution of other UK sites (see Table 2.1).

3.2.3.2 Hong Kong sites

To test reproducibility across climates, PV data from the Sai Kung District in HK (characterised by higher daytime irradiance and temperatures than the UK) was analysed. Table 3.2 displays predicted error values from power and energy calculations when using AC, DC and clear-sky models across all HK sites A-F.

Small sites A-C, with peak capacities ranging from 2.04 kW–5.00 kW, show comparable error for predicted power and energy to UK small sites with peak capacities between 2.81 kW–3.36 kW. Medium sites D–F exhibited notably higher errors, likely due their larger capacity amplifying effects caused by lower temporal resolution.

At site A with nominal capacity 2.40 kW (Figure 3.6), both AC and DC model predictions closely align with observed values, following seasonal peaks in summer and troughs in winter. A similar trend is observed for all small and large PV sites in HK (see A.1.2). As expected, the clear-sky model generated the highest theoretical outputs, but its reliance on empirical data rather than direct irradiance

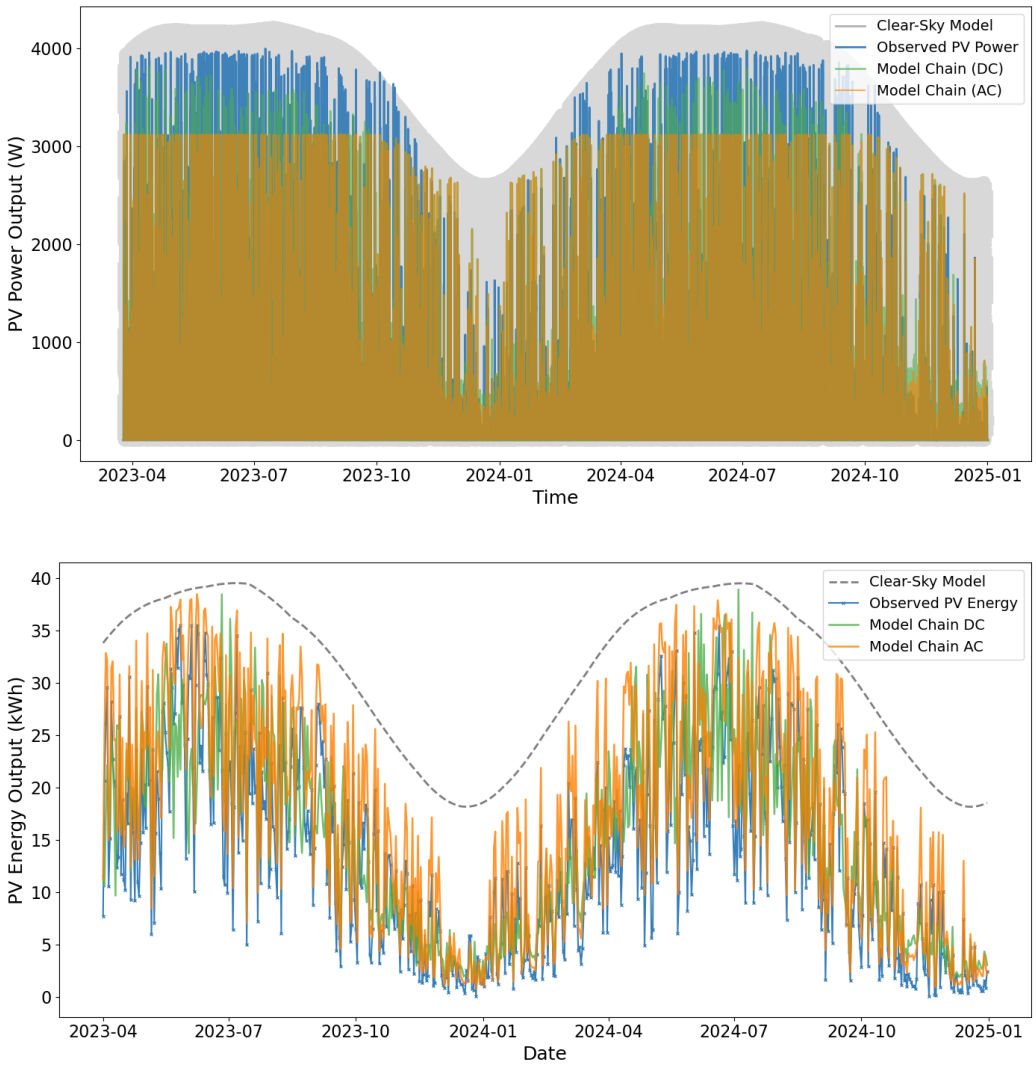

Figure 3.5: UK Oxford Site: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV daily energy data alongside deterministic model predictions with and without inverter clipping. The nominal capacity is 5.005 kW while the peak capacity is approximately 3.96 kW. True observed values shown in blue, clear-sky model in grey, AC model in orange, and DC model in green.

Table 3.2: Power and daily cumulative energy prediction performance across HK sites for the deterministic model with (AC) and without inverter conversion (DC), as well as the clear-sky model.

Model	**Power**			**Day-ahead energy**		
	MAE (W)	**RMSE (W)**	**nRMSE (%)**	**MAE (kWh)**	**RMSE (kWh)**	**MAPE (%)**
Site A						
MC (AC)	76.88	152.56	6.60	1.03	1.39	27.17
MC (DC)	80.10	158.01	6.84	1.18	1.56	32.27
Clear-sky	276.85	524.02	22.68	6.21	7.43	206.15
Site B						
MC (AC)	138.42	273.58	3.78	4.85	5.73	53.51
MC (DC)	140.92	275.79	3.81	4.29	5.12	49.62
Clear-sky	539.35	985.16	13.60	20.37	21.94	499.90
Site C						
MC (AC)	286.48	551.80	7.27	5.01	6.00	32.61
MC (DC)	281.16	570.30	7.51	4.74	6.44	36.84
Clear-sky	637.91	1250.83	16.48	21.59	26.13	205.32
Site D						
MC (AC)	1504.40	3268.28	8.25	25.80	39.00	355.72
MC (DC)	1531.44	3531.16	8.92	27.33	44.91	404.28
Clear-sky	5597.42	10408.64	26.28	128.83	152.33	935.79
Site E						
MC (AC)	1043.96	2128.01	5.37	15.81	21.79	65.95
MC (DC)	1231.67	2462.55	6.22	22.78	29.54	85.07
Clear-sky	4179.35	7546.43	19.06	97.26	110.44	350.72
Site F						
MC (AC)	1101.88	2258.51	5.70	16.96	24.39	227.26
MC (DC)	1293.69	2656.80	6.71	25.17	32.82	271.57
Clear-sky	4759.20	8355.73	21.10	107.60	119.55	750.34

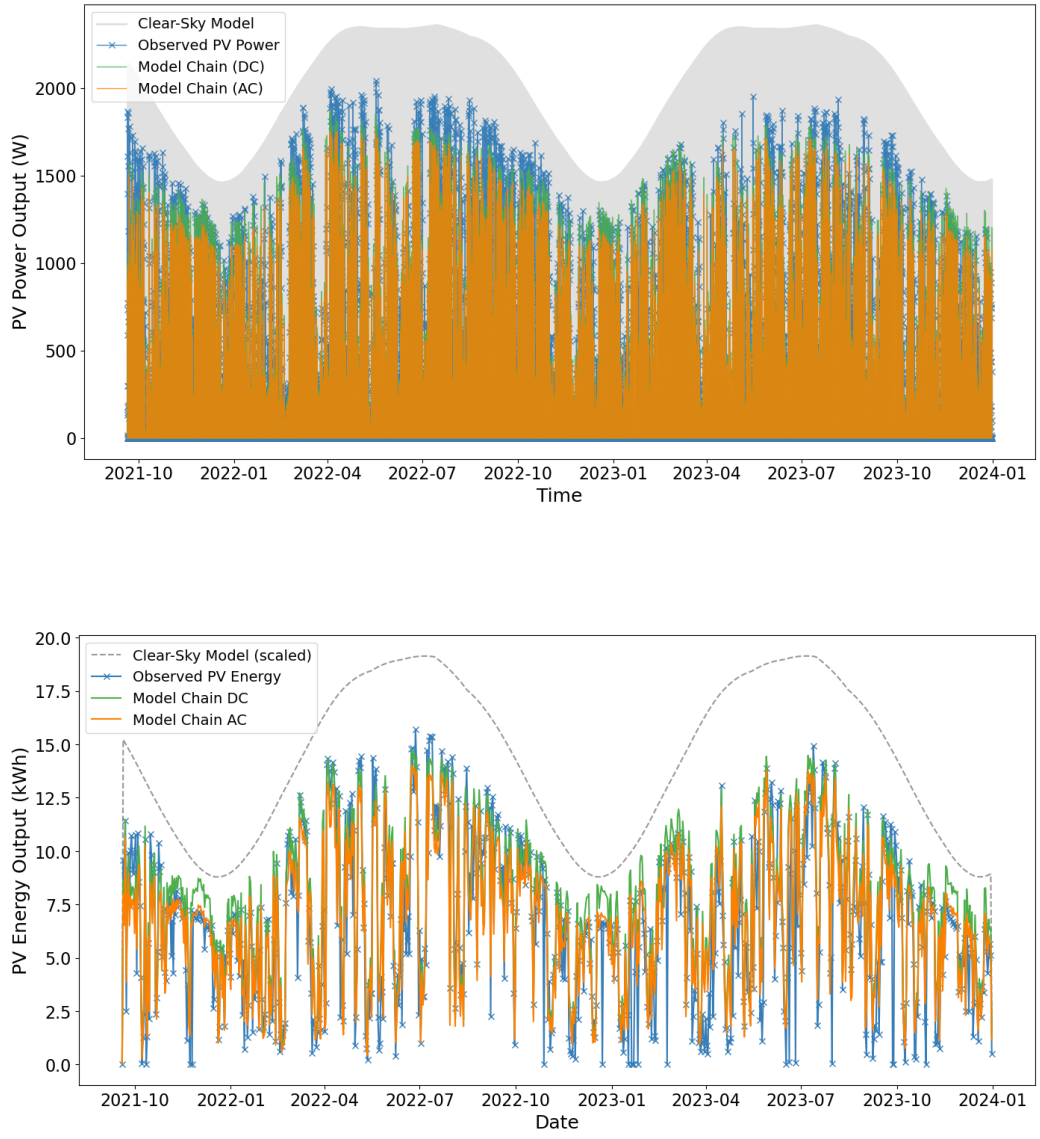

Figure 3.6: HK Site A: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed daily cumulative PV energy data alongside deterministic model predictions with and without inverter clipping. The nominal capacity is estimated to be 2.4 kW while the peak capacity is approximately 2.04 kW. Clear-sky PV output predictions are shown in grey, the AC model is shown in orange, the DC model is shown in green, and true values are shown in blue.

inputs from Solcast is likely responsible for occasional underprediction on winter days with unexpectedly high irradiance. The DC model always outputted lower power and energy predictions than the clear-sky model in summer, reflecting the incorporation of additional environmental effects into the model such as irradiance, cell temperature, and wind speed. Incorporating inverter clipping, The AC model achieved the closest match to observations at all small sites by incorporating inverter clipping effects to always generate lower predictions than the DC model. Both models typically overpredicted outputs in most small sites similarly to those in the UK, but consistently underestimated the highest observed power outputs in medium sites, indicating a need for further refined inverter modelling with greater flexibility.

3.2.3.3 Overall Performance

The figures below summarise the overall performance of the deterministic model with and without inverter modelling and clear-sky model, comparing their nRMSE and MAPE values across all combined sites. MAPE was chosen as the error metric for energy because daily cumulative energy values are comparable across sites, ensuring that percentage errors are meaningful and interpretable. By contrast, instantaneous PV power values frequently approached zero (e.g., in early morning and late evening times, or under heavy cloud cover). This causes the percentage error to become extremely large and fluctuate wildly between sites, making it less informative metric. Instead, nRMSE was used for power, as it remains well defined even when actual values are near zero and provides a more stable comparison across sites of different sizes.

Figure 3.7 shows how AC and DC output models achieve a lower nMRSE for predicted power values than the clear-sky model, highlighting the improved predictive accuracy of the deterministic model. Although the model predictions be-

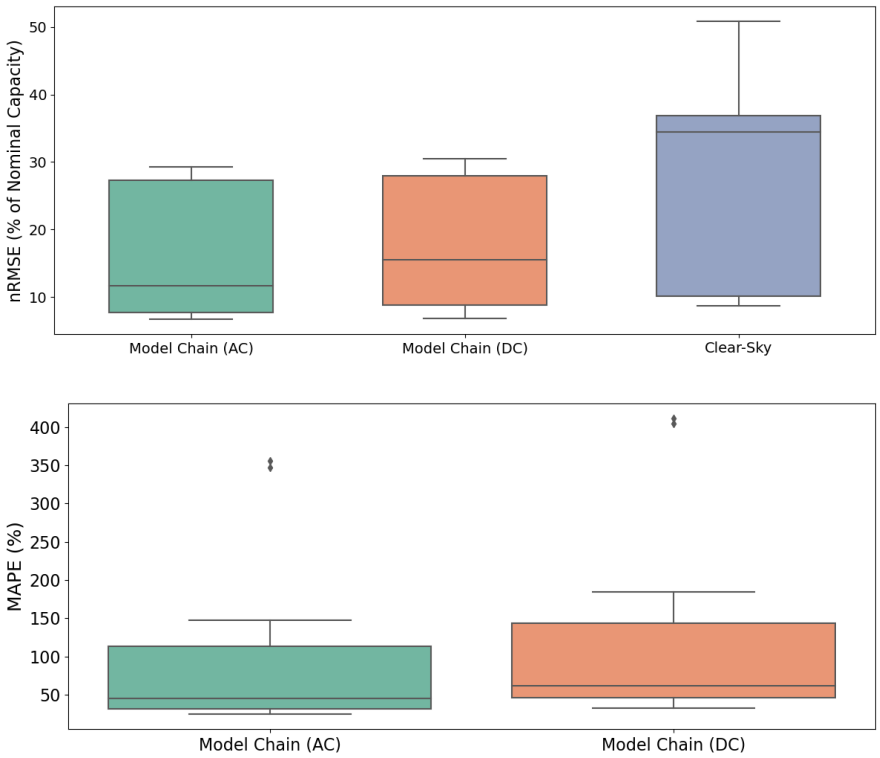

Figure 3.7: (Top) Box-plot showing nRMSE and MAPE of power output predictions for the three PV models tested across 12 PV sites. (Bottom) Box-plot showing MAPE for day-ahead energy output predictions of all three PV models across 12 PV sites. AC model shown in green, DC model shown in orange, and clear-sky shown in blue.

fore and after the inverter modelling are comparable in median and distribution, AC inverter modelling further reduced predicted nMRSE error at all sites (see Table 3.2). A similar trend was observed for the MAPE of day-ahead energy predictions; these results indicate that the AC model is the most robust of all 3 predictive models, followed closely by the DC model without inverter modelling, and lastly the clear-sky model.

3.3 Conclusion

This section introduced our proposed hybrid model approach for day-ahead solar power and energy forecasting, outlining the importance and contribution of the GP post-processing method to enhance the initial PV estimates from the deterministic physical model.

Multiplicative hybrid models are advantageous over additive ones for their ability to handle errors in highly non-linear relationships. Due to the capacity of non-linear kernels, GPs are applied to the deterministic model to calculate adjustment factors that account for additional sources of uncertainty. Multiple different approaches can be used to handle PV data, including ensemble models with statistical techniques that may confer additional flexibility and improve forecast accuracy. However, a deterministic approach was selected for its simplicity and the computational expense of GP post-processing.

The deterministic physical model (both with and without inverter clipping) consistently outperformed simpler clear-sky model predictions in both the UK and HK datasets. Prediction accuracy varied with PV site location and size, highlighting the need for parameter tuning or improved flexibility to adapt to diverse conditions and warranting further testing on PV sites in other climates.

The prediction accuracy of deterministic models used may be improved by ensemble methods that combine multiple predictions and assign them varying probabilities to capture temporal dependencies, or by the incorporation of GPs into predictions that leverage probabilistic relationships between inputs without the computational cost of large ensembles.

4 | Machine Learning Models

Chapter 4 presents the machine learning components of the hybrid forecasting framework introduced in Chapter 3. While the deterministic physical model provides an initial baseline forecast, its systematic discrepancies motivate the use of statistical and machine learning post-processing methods. This chapter therefore explores the suitability of GPs for post-processing, as well as a range of benchmark models spanning statistical, neural, and ensemble approaches. Doing so enables assessment of how different architectures capture temporal dependencies in PV data, as well as quantification of their predictive uncertainty.

We begin with GPs (Section 4.1.1), which form the core of the proposed hybrid framework. Two formulations are studied: a time-series GP designed to capture seasonal and short-term dependencies, and an auto-regressive GP that leverages lagged residuals. For each, both maximum likelihood hyperparameter estimation and fully Bayesian inference via the no-U-turn sampler are implemented, enabling a comparison between point-estimated and marginalised hyperparameters. The role of kernel selection and hyperparameter uncertainty is discussed in detail.

Next, we introduce several other benchmark machine learning models (Section 4.2). These are grouped into three categories: statistical time-series models (ARIMA and GluonTS), neural-network approaches (LSTM and transformer), and ensemble tree-based methods (random forest and XGBoost). Each benchmark is evaluated in both post-processing and direct forecasting settings, allowing for comparisons against the GP-based methods. Their inclusion also highlights the trade-offs between different modelling approaches, computational efficiency, and robustness to site-specific variability.

The evaluation methodology builds on the cross-validation framework introduced

earlier, adapted here for machine learning models. Section 4.3 presents results for 12 sites across the UK and HK. Performance is assessed in terms of point accuracy (MAE, RMSE, MAPE, nRMSE) and probabilistic calibration, providing a unified basis for comparing diverse model classes.

Overall, this chapter demonstrates the effectiveness of GPs as Bayesian post-processors of physical models by contextualising their performance within the broader landscape of machine learning approaches for solar forecasting.

4.1 Gaussian Processes

GPs are non-parametric Bayesian models for regression and classification that define a distribution over functions. A GP is specified by a mean function $m(x)$ and a covariance (kernel) function $k(x, x')$, such that for any finite set of input points the associated function values have a joint multivariate Gaussian distribution. [Rasmussen and Williams, 2006] formalised GPs in the ML context, emphasising their role as infinite-dimensional generalisations of Gaussian priors in Bayesian linear models. In practice, GPs offer flexible function fitting with principled uncertainty estimates, making them powerful tools in time-series forecasting and spatial modelling. In this thesis, GPs are used to post-process the PV power output generated by the deterministic model.

4.1.1 Gaussian Process Regression

In GP Regression, we assume that observations y are noisy evaluations of an underlying latent function f at inputs x:

$$y = f(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2), \tag{4.1}$$

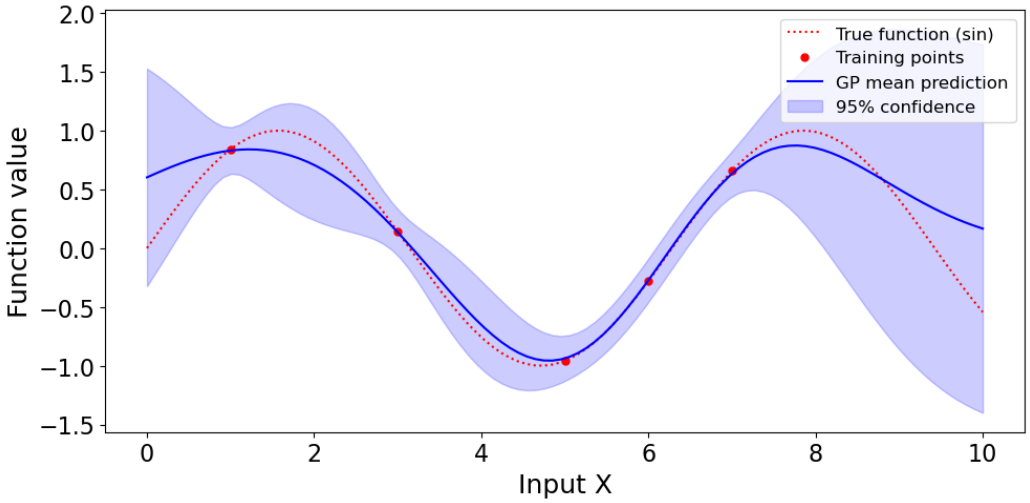

Figure 4.1: GP as a multivariate distribution over functions

The latent function f is modelled using a GP prior:

$$f \sim \mathcal{GP}(m(x), K(x, x')), \tag{4.2}$$

where $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ is the mean function and $K(\mathbf{x}, \mathbf{x}') = \text{cov}(f(\mathbf{x}), f(\mathbf{x}'))$ is the covariance function, also known as the kernel. Typically, the mean function can be taken as an explicit prior mean of the function or assumed to be zero for simplicity (since the second-order statistics encode the observed data). The choice of kernel encodes assumptions about smoothness, periodicity, linearity, and other structural properties of the latent function. Commonly used kernels include the squared-exponential Radial-Basis Fuction (RBF) kernel, Matérn kernels, and periodic kernels. A more detailed discussion on kernels and their hyperparameters is presented in Section 4.1.2.

Given a training dataset $\{(x_i, y_i)\}_{i=1}^n$, the GP prior implies a joint Gaussian distribution for the latent function values at the training and test points. Conditioning on the observed data yields a posterior predictive distribution for new inputs. In

particular, for test inputs $\mathbf{X}_* \in \mathbb{R}^{m \times d}$ (where m is the number of test points and d is the input dimension), we have:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_X \\ \mathbf{m}_* \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right), \tag{4.3}$$

where $K(X,X)$ is the $n \times n$ covariance matrix over the training inputs, $K(X,X_*)$ is the $n \times m$ cross-covariance between the training and test inputs, and $K(X_*,X_*)$ is the $m \times m$ covariance matrix over the test inputs. To model noisy observations at the test points, Gaussian white noise can be added to $K(X_*,X_*)$ by including the term $\sigma_n^2 I$.

Assuming the training observations are independent and identically distributed (i.i.d.) with Gaussian noise, the predictive distribution for the latent function values at the test inputs, $f_* = f(X_*)$, is Gaussian with the following mean and covariance:

$$\mathbb{E}[f_*] = \mathbf{m}_* + K(X_*,X)\left[K(X,X) + \sigma_n^2 I\right]^{-1}(\mathbf{y} - \mathbf{m}_X), \tag{4.4}$$

$$\mathrm{Cov}[f_*] = K(X_*,X_*) - K(X_*,X)\left[K(X,X) + \sigma_n^2 I\right]^{-1} K(X,X_*). \tag{4.5}$$

The $\mathbf{m}_X$ and $\mathbf{m}_*$ here denote the prior mean vectors $m(X)$ and $m(X_*)$, respectively. Equations (4.4) and (4.5) yield both a predictive mean function and a predictive variance that quantifies the uncertainty at each test point. This process involves inverting the matrix $K(X,X) + \sigma_n^2 I$, which is typically done via Cholesky decomposition for numerical stability and computational efficiency.

If we denote $K_y = K(X,X) + \sigma_n^2 I$, its Cholesky decomposition is

$$K_y = LL^\top, \tag{4.6}$$

where L is a lower triangular matrix. This decomposition enables us to avoid explicitly computing the inverse of K_y, thereby improving numerical stability and efficiency.

In practice, to solve $K_y^{-1}(\mathbf{y} - \mathbf{m}_X)$, we first solve the system

$$L\mathbf{v} = \mathbf{y} - \mathbf{m}_X, \tag{4.7}$$

$$L^\top \boldsymbol{\alpha} = \mathbf{v}, \tag{4.8}$$

where $\mathbf{v}$ is the intermediate vector ($L^\top \boldsymbol{\alpha}$) and $\boldsymbol{\alpha} = K_y^{-1}(\mathbf{y} - \mathbf{m}_X)$. Using $\boldsymbol{\alpha}$, the predictive mean at test inputs X_* is given by

$$\mathbb{E}[f_*] = \mathbf{m}_* + K(X_*, X)\,\boldsymbol{\alpha}. \tag{4.9}$$

Similarly, the predictive covariance can be expressed in terms of L without directly inverting K_y, which not only reduces computational complexity but also mitigates numerical errors.

In time-series regression, which is the focus of this thesis, two common approaches are used: time-series GPs and auto-regressive GPs (ARGPs). Time-series GPs place a joint GP prior over all time points, modelling time as a continuous input. In contrast, ARGPs use past observations as inputs to predict future values. More details are discussed in the following sections.

4.1.1.1 Time-series GP

GPs can be directly applied to model time-series data by treating the time index t as the input and the observation y_t as the output. In this formulation, the latent

function f is placed under a GP prior, as shown in equations (4.4) and (4.5):

$$y_t = f(t) + \epsilon_t, \quad f \sim \mathcal{GP}(m, K), \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{4.10}$$

By choosing appropriate kernels, time-series GPs can flexibly capture trends, periodicities, and other temporal structures. For example, the RBF and Matérn kernels induce variably smooth functions, while the periodic kernel models seasonality. Stationary kernels such as the RBF, Matérn and periodic kernels depend only on time differences $|t - t'|$, implying time-invariant behaviour. In contrast, non-stationary kernels like the linear or neural-network kernel allow for time-dependent variation in signal characteristics [Remes et al., 2017].

In the application of PV adjustment factor forecasting, higher adjustment values are typically observed in summer months due to underperformance of the deterministic model with inverter clipping, while lower adjustment values occur in winter since the model aligns more closely with the actual generation. To model this temporal behaviour, composite kernels consisting of both stationary and non-stationary components can be considered. These allow simultaneous modelling of long-term trends and seasonal effects, for example by combining stationary kernels such as RBF or Matérn with non-stationary components like the linear kernel [Roberts et al., 2013]. Although the seasonal PV power variation appears non-stationary due to its changing mean and variance across time, it can be regarded as stationary periodic behaviour since the summer–winter effect repeats annually in theory. In practice, however, a pure periodic kernel did not provide a good fit, largely because of power peak fluctuations and evolving seasonal amplitudes. This motivated the use of a combination of stationary and non-stationary kernels to capture both the fluctuating seasonal effects and long-term trends.

A key advantage of GPs is their ability to handle irregular sampling and provide

full predictive distributions. However, exact GP inference scales cubically with the number of observations, $\mathcal{O}(n^3)$, due to the need to invert the covariance matrix. In practice, this is performed via Cholesky factorisation, which is numerically more stable and scalar-multiple more efficient than directly computing the inverse, but still retains $\mathcal{O}(n^3)$ complexity. If the time series is uniformly sampled and the kernel is stationary, the covariance matrix becomes Toeplitz. In this case, efficient algorithms such as the Yule-Walker equations can be used to compute predictions in $\mathcal{O}(T^2)$ time by exploiting this structure [Golub and Van Loan, 1996]. These equations estimate the parameters of an equivalent autoregressive process using autocovariance values and can be solved efficiently using the Levinson-Durbin algorithm. Further improvements are possible with *state-space* GP formulations, which recast the GP as a linear dynamical system. In this setting, recursive inference using Kalman filtering enables $\mathcal{O}(T)$ time complexity [Hartikainen and Särkkä, 2010]. Alternatively, sparse GPs approximate the full covariance structure using a smaller set of inducing points, typically achieving $\mathcal{O}(nm^2)$ complexity with $m \ll n$, while preserving much of the expressiveness of the full GP model.

4.1.1.2 Auto-regressive Gaussian Processes

An ARGP treats the time series as a regression problem on its own lagged outputs. Define the lag-L vector $\mathbf{x}_t = [y_{t-1}, y_{t-2}, \ldots, y_{t-L}]^\top$. The GP becomes:

$$y_t = f(\mathbf{x}_t) + \epsilon_t, \quad f \sim \mathcal{GP}(m, K), \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{4.11}$$

This is analogous to a classical autoregressive model of order L, denoted AR(L), where the current observation y_t is modelled as a linear function of its L most recent past values. More details on autoregressive statistical models, including

ARMA, ARIMA, and DeepAR (via GluonTS), are discussed in Section 4.2.1.

$$y_t = \sum_{i=1}^{L} \phi_i y_{t-i} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{4.12}$$

where ϕ_i are fixed autoregressive coefficients and ϵ_t is a Gaussian white noise term. In contrast with AR models, the ARGP generalises this setup by replacing the linear mapping $\sum \phi_i y_{t-i}$ with a non-parametric function $f(\cdot)$ drawn from a GP prior.

The input dimension is L, and the kernel $K(\mathbf{x}, \mathbf{x}')$ can capture non-linear dependencies between past values and the future. The ARGP is typically trained on historical data $\{y_1, \ldots, y_T\}$ to learn kernel hyperparameters. As suggested by [Quinonero-Candela et al., 2003], iterative one-step-ahead forecasting generally outperforms direct multi-step forecasting, especially in nonlinear and noisy systems. In direct multi-step forecasting, the model attempts to predict y_{t+H} directly from the current state $\mathbf{x}_t$. This approach requires learning a complex mapping and integrating over intermediate uncertainty, which is computationally challenging for GPs due to intractable future-step integrals. In contrast, one-step-ahead iterative forecasting proceeds recursively: at each step $t = T_{train} + 1, \ldots, T_{train} + H$, the input $\mathbf{x}_t$ is formed from the most recent L observations, the GP predictive mean $\hat{y}_t$ is computed, and this prediction is appended to the sequence for the next input. However, since predicted values are used recursively as inputs, uncertainty may accumulate over time.

In the context of PV adjustment factor predictions, we avoid this issue by using the true observed y_t at each step during forecasting. This is justified by the daily availability of adjustment factors, as discussed in Chapter 2. The ARGP one-step-ahead iterative forecasting procedure is outlined in Algorithm 1.

Algorithm 1 Iterative ARGP Forecasting

Train GP on pairs $([y_{t-1}, \ldots, y_{t-L}], y_t)$ for $t = L+1, \ldots, T_{\text{train}}$
for $t = T_{\text{train}} + 1$ to $T_{\text{train}} + H$ **do**
    $\mathbf{x}_t = [y_{t-1}, \ldots, y_{t-L}]$ ▷ Lagged inputs
    $\mu_t, \sigma_t^2 = \text{GP.predict}(\mathbf{x}_t)$ ▷ Posterior mean/variance
    $\hat{y}_t = \mu_t$ ▷ One-step forecast
    Append y_t to series y ▷ Observed value
end for
return Forecasts $\{\hat{y}_t\}$

The flexibility of kernels and Bayesian uncertainty quantification gives ARGP an advantage in modelling nonlinear dynamics over classical AR models in real-world time series. In this thesis, we also apply an analogous iterative one-step-ahead forecasting approach to time-series GP models. Since time-series GPs produce a full joint posterior over multiple time points, direct multi-step forecasting becomes computationally expensive for long horizons. Moreover, the PV adjustment factor data (see Figure 3.1) exhibits substantial variability and complex temporal dependencies. The iterative one-step-ahead forecasting approach allows the model to adapt to these temporal patterns by updating predictions step-by-step. The performance of both ARGP and time-series GP models using this iterative approach is discussed in the results section of this chapter.

Algorithm 2 Iterative Time-Series GP Forecasting

Training GP on data $\{(t_i, y_i)\}_{i=1}^{T_{\text{train}}}$
for $t = T_{\text{train}} + 1$ to $T_{\text{train}} + H$ **do**
    $\mu_t, \sigma_t^2 = \text{GP.predict}(t)$
    $\hat{y}_t = \mu_t$
    Append (t, y_t) to test set
end for
return Forecasts $\{\hat{y}_t\}_{t=T_{\text{train}}+1}^{T_{\text{train}}+H}$

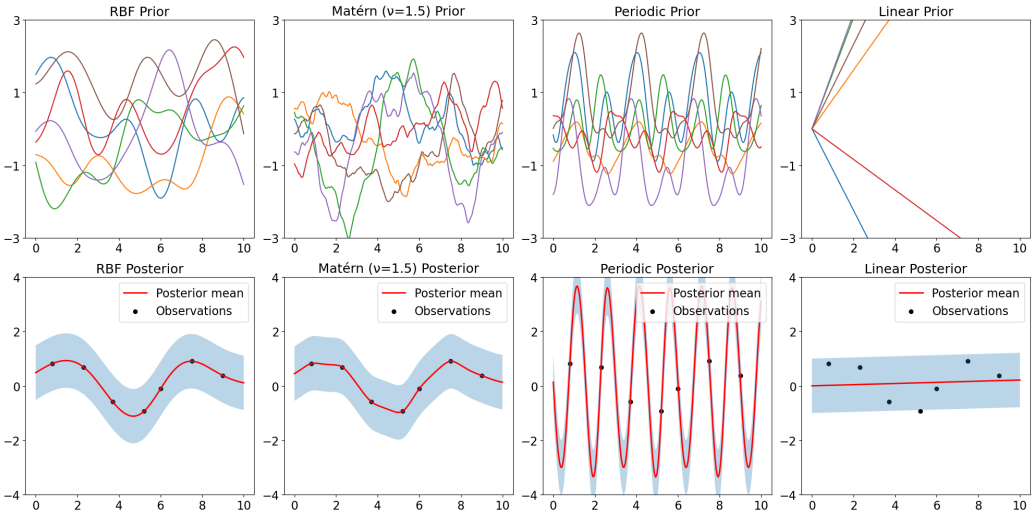

Figure 4.2: Sample function draws from GP priors (top row) and corresponding posteriors after observing data points (bottom row) using four different covariance kernels: RBF, Matérn ($\nu = 1.5$), periodic, and linear. This illustrates how different kernels encode different prior assumptions about smoothness and structure, and how these affect the resulting posterior fits to data.

4.1.2 Kernel and Hyperparameter Selection

As previously introduced, the kernel $K(x, x')$ fully characterises the properties of a GP, determining how function values at different inputs co-vary and thereby controlling smoothness, periodicity, or other structural assumptions. By specifying a kernel function, we implicitly define a distribution over functions, describing how we expect the function to vary before observing any data. Figure 4.2 illustrates sample function draws from GP priors and the corresponding posteriors after observing data. This highlights how the choice of kernel profoundly shapes the GP's behaviour, allowing us to encode smoothness, roughness, or repeating patterns as needed. In addition to the kernel, a GP may also include hyperparameters from the prior mean function (e.g., slope and intercept if a linear mean is used) and from the likelihood function (e.g., observation noise variance), which must likewise be estimated or specified.

As observed, the RBF kernel produces very smooth, infinitely differentiable functions [Rasmussen and Williams, 2006], whereas Matérn kernels can generate rougher functions depending on their smoothness parameter ν. Kernels like the rational quadratic can model functions with multiple length-scales by acting as a scale mixture of RBFs. Periodic kernels induce strictly repeating patterns, making them ideal for modelling seasonal or cyclic effects. Lastly, linear and affine linear kernels are non-stationary and can capture global linear trends in the data.

Each kernel is parameterised by a set of hyperparameters, such as the length-scale ℓ, signal variance σ_f^2, periodicity parameter T and etc. (see Section 4.1.2.1), that control the flexibility and shape of the resulting functions. These hyperparameters play a crucial role in determining the GP's fit to data and are typically estimated from observations by maximising the marginal likelihood or via Bayesian inference through posterior sampling. Further details regarding hyperparameter selection are provided in Section 4.1.2.2.

4.1.2.1 Kernel Selection

In this section, we discuss typical choices of kernels for time-series GPs. We briefly describe commonly used kernels, focusing on those relevant to modelling PV adjustment factors. While this list is not exhaustive, it covers most covariance functions suitable for time-series analysis. We note that valid kernels can be combined via sums or products to form new valid covariance functions, thereby allowing multiple explanatory hypotheses to be encoded flexibly. For further technical details, readers are referred to [Rasmussen and Williams, 2006, Roberts et al., 2013, Osborne, 2010].

Affine linear kernel

$$k_{\text{affine}}(x, x') = \sigma_b^2 + \sigma_v^2(x - c)(x' - c). \tag{4.13}$$

The affine linear kernel models global linear trends in the data and is non-stationary. Here, σ_v^2 controls the variance of the linear slope, σ_b^2 governs the bias term which is also the vertical offset, and c can act as a centring constant. When $\sigma_b^2 = 0$ and $c = 0$, the affine kernel reduces to the standard linear kernel.

Squared-exponential

$$k_{\text{RBF}}(x, x') = \sigma_f^2 \exp\left[-\frac{(x - x')^2}{2\ell^2}\right]. \tag{4.14}$$

The RBF kernel yields very smooth, infinitely differentiable functions with a characteristic length-scale ℓ. The parameter σ_f^2 controls the output-scale (amplitude) of the function. This kernel is widely used for modelling functions that are expected to vary smoothly across the input space.

Rational quadratic

$$k_{\text{RQ}}(x, x') = \sigma_f^2 \left(1 + \frac{(x - x')^2}{2\alpha\ell^2}\right)^{-\alpha}. \tag{4.15}$$

Here, α is known as the shape parameter or index, which determines how heavily different length-scales are mixed. As discussed by [Rasmussen and Williams, 2006], the rational quadratic kernel can be interpreted as an infinite scale mixture of RBF kernels with different length-scales, enabling it to model functions with varying degrees of smoothness. As $\alpha \to \infty$, the rational quadratic kernel converges to the RBF kernel.

Matérn A flexible family of kernels $k_\nu(x, x')$ parameterised by a smoothness parameter ν, which controls the differentiability of sample paths. For example, $\nu = \frac{3}{2}$ and $\nu = \frac{5}{2}$ produce functions that are once and twice differentiable, respectively. As $\nu \to \infty$, the Matérn kernel converges to the RBF kernel. At $\nu = \frac{1}{2}$, it reduces to the exponential (Laplace) kernel, leading to very rough and less smooth functions:

$$k_{\text{Matérn}\frac{3}{2}}(x, x') = \sigma_f^2 \left(1 + \frac{\sqrt{3}|x - x'|}{\ell}\right) \exp\left(-\frac{\sqrt{3}|x - x'|}{\ell}\right). \tag{4.16}$$

Periodic and quasi-periodic

$$k_{\text{per}}(x, x') = \sigma_f^2 \exp\left[-\frac{2\sin^2(\pi(x - x')/T)}{w^2}\right], \tag{4.17}$$

which enforces strict periodicity with period T. This kernel captures exactly repeating patterns and is typically used for modelling cyclic phenomena. When multiplied with an RBF kernel, it forms a quasi-periodic kernel that allows periodic patterns to vary smoothly over time, e.g. to model slowly changing seasonal amplitudes.

Composite kernels Due to the positive semi-definite nature of kernels, they can be combined through addition or multiplication to capture more complex functional behaviours. Additive kernels, for example $k(x, x') = k_1(x, x') + k_2(x, x')$, model superpositions of effects such as global trends combined with seasonal variations. Multiplicative kernels, for example $k(x, x') = k_{\text{RBF}}(x, x') \times k_{\text{per}}(x, x')$, capture interactions where a periodic pattern has amplitude or smoothness modulated by another kernel. Scalar-weighted combinations are also possible, such as $k(x, x') = \alpha k_1(x, x') + \beta k_2(x, x')$, where coefficients α and β control the relative influence of each component.

GPs can also be extended to handle structural breaks and richer data structures. Change-point kernels allow for modelling sudden regime shifts by using different covariance functions before and after a transition point x_c. This is typically achieved by multiplying stationary kernels with step or sigmoidal transition functions, enabling the GP to apply one kernel before x_c and another after, thus capturing abrupt changes in trends or variances [Roberts et al., 2013].

Furthermore, GP kernels naturally generalise to multiple inputs and outputs. For multi-dimensional inputs, a common approach is to define a product kernel across input dimensions, for example $k(\mathbf{x}_i, \mathbf{x}_j) = \prod_e k^{(e)}(x_i^{(e)}, x_j^{(e)})$, allowing each input feature to contribute flexibly to the overall covariance. In multi-output settings (such as correlated time series from multiple sites, or spatial-temporal measurements from a single sensor), one can incorporate output indices directly into the kernel. For example, $k([l_m, t_i], [l_n, t_j]) = k_t(t_i, t_j)\, k_l(l_m, l_n)$, where l denotes the output label. These formulations enable shared learning across related outputs and capture complex structured dependencies. Although these extensions offer substantial modelling flexibility, they are not employed in this thesis since the focus remains on univariate PV adjustment factor time-series modelling. For further technical details on these advanced constructions, readers are referred to Roberts et al. [2013], Alvarez et al. [2012].

In this thesis, even though the PV sites at the same location may exhibit weather-related correlations, we adopt a single-input, single-output GP framework for PV adjustment factor prediction. This choice is motivated by model simplicity, computational efficiency, and clearer interpretability of adjustment dynamics at each individual site.

For the ARGP model, a combination of an affine linear kernel and a Matérn kernel was used to capture both global linear trends and moderately smooth nonlinear

deviations:

$$k_{\text{ARGP}}(x, x') = k_{\text{affine}}(x, x') + k_{\text{Matérn}\frac{3}{2}}(x, x'). \tag{4.18}$$

For the time-series GP model, we employed a composite kernel that combines a scaled RBF kernel and a Matérn kernel to simultaneously capture smooth seasonal trends and local variations:

$$k_{\text{TSGP}}(t, t') = \gamma\, k_{\text{RBF}}(t, t') + k_{\text{Matérn}\frac{3}{2}}(t, t'), \tag{4.19}$$

where γ is a scaling coefficient controlling the magnitude of the RBF component.

4.1.2.2 Hyperparameter Marginalisation

GP hyperparameters are used to control model flexibility and are typically estimated by maximising the marginal likelihood (also known as Type-II maximum likelihood) or by performing Bayesian inference using priors.

In terms of MLE, we aim to minimise the negative log marginal likelihood, also referred to as the loss function:

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbf{y}^\top (K_\theta + \sigma_n^2 I)^{-1}\mathbf{y} + \frac{1}{2}\log|K_\theta + \sigma_n^2 I| + \frac{n}{2}\log(2\pi). \tag{4.20}$$

This loss can be minimised using gradient-based optimisation methods such as L-BFGS [Liu and Nocedal, 1989]. In practice, hyperparameters are often parameterised in log space (e.g., $\log \ell$, $\log \sigma_f$) to enforce positivity constraints during optimisation. Gradients of the loss with respect to each hyperparameter can be derived analytically and are efficiently computed via closed-form expressions or automatic differentiation frameworks. Although the Type-II MLE approach is computationally efficient, it provides only a point estimate of the hyperparameters and may underestimate posterior uncertainty, as it does not account for hyperpa-

rameter variability.

Alternatively, maximum a posteriori (MAP) estimation introduces a prior $p(\theta)$ over the hyperparameters but seeks only the mode of the posterior $p(\theta \mid \mathbf{y})$. In practice, MAP is similar to MLE but with an added prior term that regularises the optimisation, making it more robust to overfitting than pure MLE, while still much cheaper than full Bayesian inference.

Furthermore, an entirely Bayesian approach is possible by placing a prior $p(\theta)$ on the hyperparameters and sampling from the posterior $p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta)$ using libraries like Pyro and GPyTorch.

Markov Chain Monte Carlo Markov Chain Monte Carlo (MCMC) methods are commonly employed to approximate a posterior by generating samples from it. In the Bayesian treatment of Gaussian processes, the predictive distribution requires marginalization over the hyperparameters:

$$p(\mathbf{y}_* \mid \mathbf{y}) = \int p(\mathbf{y}_* \mid \mathbf{y}, \theta)\, p(\theta \mid \mathbf{y})\, d\theta, \tag{4.21}$$

where $p(\theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \theta)p(\theta)$ is the posterior over hyperparameters. This integral is analytically intractable, motivating the use of sampling-based approaches. MCMC constructs a Markov chain whose stationary distribution is the desired posterior; after sufficient iterations, the samples approximate $p(\theta \mid \mathbf{y})$. At each iteration, MCMC algorithms propose a new state which in this case is a candidate set of hyperparameters and decide whether to accept it, thus gradually exploring the posterior distribution. A particularly effective class of MCMC methods for continuous and high-dimensional parameter spaces is the Hamiltonian Monte Carlo (HMC) method. HMC leverages gradient information of the log-posterior to propose new states, which is more efficient than simple random-walk Metropolis-

Hastings methods. However, HMC requires careful tuning of parameters such as the trajectory length and step size.

No-U-Turn Sampler The NUTS method [Hoffman and Gelman, 2014] extends HMC by adaptively determining when to stop each trajectory to prevent it from turning back on itself. This adaptivity makes NUTS robust and less sensitive to manual tuning, providing efficient sampling performance for hyperparameter inference.

In summary, hyperparameters can be estimated using a Type-II MLE approach. Alternatively, fully Bayesian marginalisation can be employed to provide more comprehensive uncertainty quantification, yet at a higher computational cost [Rasmussen and Williams, 2006, Turner, 2011, Roberts et al., 2013]. In this thesis, we explore and compare GP model performance using both MLE and NUTS-based fully Bayesian approaches.

4.1.3 Results

In this section, we fit the ARGP and time-series GP models to the adjustment factor, defined as the ratio of observed to predicted power in Section 3.1, across twelve sites in the UK and Hong Kong. Hyperparameters are inferred using both MLE and NUTS to contrast point-estimate inference with full Bayesian marginalisation. The kernel functions for ARGP and the time-series GP correspond to equations (4.18) and (4.19), respectively.

Model evaluation is based on point accuracy (MAE, RMSE) and probabilistic reliability using calibration scores. The calibration score is defined as the proportion of true observations that fall within a model's predictive interval. A well-calibrated model should yield calibration scores close to the nominal confidence

Table 4.1: Point accuracy and empirical coverage of ARGP and time-series GP with different hyperparameter inference methods at Oxford site.

Model	**MAE**	**RMSE**	**Cal. 68%**	**Cal. 95%**	**Cal. 99.7%**
ARGP (MLE)	0.369	0.598	69.7	91.9	95.1
ARGP (NUTS)	0.368	0.613	72.0	92.4	95.3
Time series GP (MLE)	0.342	0.560	76.5	92.4	94.6
Time series GP (NUTS)	0.338	0.549	78.2	93.7	95.8

level. For a dataset of size n and nominal coverage p (e.g. 95%), the calibration score is

$$\mathrm{Cal}(p) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\big\{y_i \in [q_{i,(1-p)/2},\ q_{i,1-(1-p)/2}]\big\}, \tag{4.22}$$

where y_i is the true observation, and $q_{i,\alpha}$ denotes the α-quantile of the predictive distribution. Table 4.1 reports the Oxford site results for both GP variants and inference methods.

Both ARGP and time-series GP achieve competitive point accuracy, with the latter showing consistently lower MAE and RMSE at Oxford. The stronger performance of the time-series GP could be due to kernel selection and the difference in input structures. In terms of calibration, all models align reasonably with nominal confidence levels. At the 68% interval, coverage is slightly above target (e.g. 76–78% for TS-GP), indicating conservative uncertainty in narrow bands. At the 95% and 99.7% levels, coverage falls a few percentage points short of nominal, suggesting mild overconfidence. Importantly, the NUTS variants reduce this gap relative to MLE, yielding broader intervals and more reliable coverage. Overall, the time-series GP with NUTS provides the best balance between accuracy and calibrated uncertainty at Oxford. Similar patterns are observed across the other UK and Hong Kong sites, with time-series GP outperforming ARGP in point accuracy and NUTS consistently improving calibration. Only time series GPs were used for subsequent modelling, as although ARGP models achieved comparable error metrics to time series GP models, their predictions remained close to mean

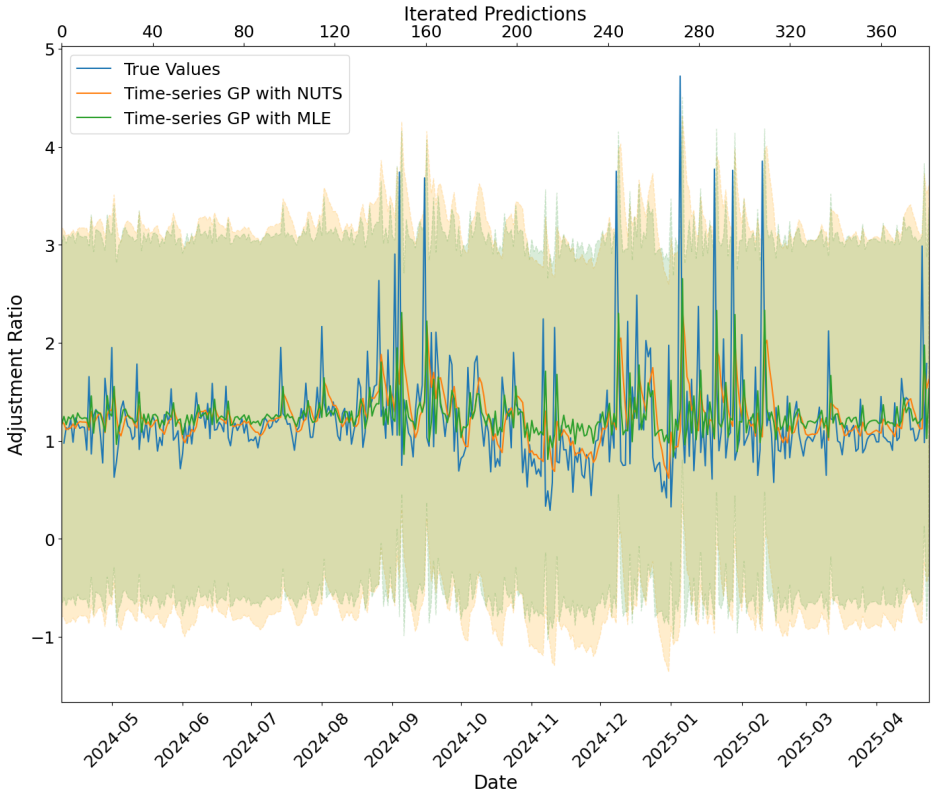

Figure 4.3: Time-series GP at the Oxford site: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

values and largely failed to capture the variability and temporal dynamics of PV generation. By contrast, time series GPs explicitly encode seasonal, diurnal, and trend structures in their kernels, producing forecasts that are not only statistically accurate but also physically meaningful across changing conditions.

As shown in Figure 4.3, the NUTS variant produces visibly broader credible intervals than MLE. This arises because NUTS integrates over hyperparameter uncertainty, inflating predictive variance through the law of total variance:

$$\mathrm{Var}(y^* \mid \mathcal{D}) = \mathbb{E}_{\theta|\mathcal{D}}[\mathrm{Var}(y^* \mid \theta, \mathcal{D})] \; + \; \mathrm{Var}_{\theta|\mathcal{D}}[\mathbb{E}(y^* \mid \theta, \mathcal{D})]\,.$$

The second term vanishes for MLE point estimates but is non-zero when hyper-parameters are marginalised under NUTS, leading to wider and better-calibrated credible bands. This trend is consistent across all other sites (see Appendix Figures A.21–A.31).

4.2 Benchmark Models

These benchmark models provide a diverse set of comparison points to assess the performance and robustness of the proposed GP-based frameworks.

4.2.1 Statistical Methods

In this thesis, we further consider benchmark autoregressive models for comparison. Specifically, we include classical statistical models such as ARIMA, as well as neural network-based autoregressive approaches such as DeepAR implemented via GluonTS [Alexandrov et al., 2020]. DeepAR combines RNNs, specifically LSTM networks (discussed in Section 4.2.2.1), with probabilistic forecasting by modelling the conditional distribution $p(y_t \mid y_{1:t-1})$. This allows for flexible handling of complex temporal dependencies and uncertainty quantification through sampling-based prediction rather than point estimates.

4.2.1.1 ARIMA

Similar to the AR models defined in Equation 4.12, ARIMA models incorporate additional integration and moving average (MA) components to handle more complex temporal patterns. The MA component accounts for past forecast errors by incorporating them directly into the prediction equation.

$$y_t = \mu + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}, \tag{4.23}$$

When combining AR and MA terms, we obtain an ARMA model, which can be expressed as:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t, \tag{4.24}$$

where p is the number of AR terms (L in Equation (4.12)), q is the number of MA terms, ϕ_i are the AR coefficients, θ_j are the MA coefficients, and ϵ_t represents white noise.

ARIMA models extend the ARMA framework by introducing an integration (I) component to handle non-stationarity. On this transformed (differenced) dataset, each predicted value is regressed on its past values (AR), while past forecast errors are also incorporated via the MA component. This formulation allows ARIMA to effectively model a wide range of time-series behaviours, including trends and temporary shocks, making it a more versatile tool than ARMA [Box and Jenkins, 1970]:

$$\nabla^d y_t = \mu + \sum_{i=1}^{p} \phi_i \nabla^d y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t, \tag{4.25}$$

where d is the order of differencing required to achieve stationarity.

This class of autoregressive models has a formal connection to time-series GPs through their equivalent state-space representations. In AR and ARMA models, the one-step-ahead predictive distribution for y_t given past observations can be written as:

$$p(y_t \mid \mathbf{y}_{1:t-1}) = \mathcal{N}(A(t, 1\!:\!t\!-\!1)\, \mathbf{y}_{1:t-1},\ M(t,t))\,, \tag{4.26}$$

where $A(t, 1\!:\!t\!-\!1)$ represents the vector of autoregressive coefficients and $M(t,t)$ denotes the one-step predictive variance. In the case of AR models, $M(t,t)$ corresponds to the white noise variance ϵ_t. For ARMA models, it also includes the MA components.

For ARIMA models, this predictive distribution is applied to the differenced series $\nabla^d y_t = (1 - B)^d y_t$:

$$p(\nabla^d y_t \mid \nabla^d \mathbf{y}_{1:t-1}) = \mathcal{N}\big(A(t, 1\!:\!t\!-\!1)\, \nabla^d \mathbf{y}_{1:t-1},\ M(t,t)\big)\,. \tag{4.27}$$

The predicted values of the original series y_t can then be obtained by recursively integrating the differenced forecasts.

ARIMA-style models provide a good basis for benchmarking GP models, especially for representing linear dynamics in autoregressive terms. For further details on the equivalence between autoregressive state-space models and GP formulations, see [Turner, 2011].

4.2.1.2 GluonTS

GluonTS is a probabilistic time-series modelling toolkit developed by [Alexandrov et al., 2020]. In this thesis, we adopt the DeepAR model, implemented via GluonTS, as a benchmark for the GP-based post-processing components.

DeepAR belongs to the class of neural network-based autoregressive models. It uses a RNN architecture, typically employing LSTM or gated recurrent unit (GRU) cells, to model the sequential structure of time-series data. At each time step t, DeepAR takes as input the past observations $y_{1:t-1}$ and any available covariates, and outputs the parameters of a predictive distribution for y_t.

Specifically, DeepAR parameterises a parametric likelihood function, most commonly Gaussian for continuous-valued series. The RNN produces the mean μ_t and standard deviation σ_t of this distribution at each time point:

$$p(y_t \mid y_{1:t-1}, \mathbf{x}_{1:t}) = \mathcal{N}(y_t \mid \mu_t, \sigma_t^2), \tag{4.28}$$

where $\mathbf{x}_{1:t}$ denotes any optional covariates.

After obtaining these parameters, DeepAR draws a sample y_t from the predicted distribution:

$$y_t \sim \mathcal{N}(\mu_t, \sigma_t^2), \tag{4.29}$$

and recursively uses this sampled value as input for predicting the next step y_{t+1}. This recursive sampling defines an autoregressive generative process that models the sequence of conditional distributions:

$$p(y_{t+1} \mid y_{1:t}) = \mathcal{N}(\mu_{t+1}, \sigma_{t+1}^2). \tag{4.30}$$

By iteratively sampling from these conditional distributions, DeepAR can generate comprehensive probabilistic trajectories over the forecast horizon. The predictive mean for each future time step can be approximated by averaging over multiple sampled paths:

$$\mathbb{E}[y_{t+1}] \approx \frac{1}{S} \sum_{s=1}^{S} y_{t+1}^{(s)}, \tag{4.31}$$

where S denotes the number of generated samples.

This sampling-based approach provides both point forecasts and quantification of forecast uncertainty through empirical quantiles. Compared to traditional AR models, DeepAR leverages the flexibility of neural networks to capture complex non-linear temporal dependencies and seasonality patterns, making it a strong benchmark for probabilistic forecasting.

4.2.2 Neural Network-based Approaches

The ability of ML approaches to rapidly uncover complex nonlinear relationships from datasets with high amounts of noise makes them a popular choice for fore-

casting techniques. Neural networks are among the most powerful ML models, as they automatically process complex information through a network of interconnected nodes to learn patterns with minimal information beyond the relatively large training datasets provided. Some common types include feed-forward neural networks (FNNs), CNNs, RNNs, and GANs. RNNs are typically used for sequential data handling, such as time series forecasts, because of their ability to store recent input as a short-term memory through feedback connections. LSTM RNNs are often chosen for PV forecasting due to their ability to recognise temporally distant patterns in data. Alternatively, transformer architectures are unique in their absence of reliance solely on attention mechanisms (unlike RNNs), typically achieving further improved long-range modelling and parallelism than other architectures. In this section, we review the LSTM and transformer models, along with their applications to PV forecasting tasks.

4.2.2.1 LSTM

RNNs are widely used for modelling sequential data due to their ability to maintain temporal dependencies. However, standard RNNs suffer from unstable error gradients when trained using backpropagation through time. These gradients either vanish, impeding long-term learning, or explode, leading to unstable predictions.

The backpropagated error at unit j at time t is given by:

$$\vartheta_j(t) = f_j'\left(\mathrm{net}_j(t)\right) \sum_i w_{ij}\, \vartheta_i(t+1), \tag{4.32}$$

where $f_j'\left(\mathrm{net}_j(t)\right)$ is the derivative of the activation function and w_{ij} is the weight

from unit j to unit i. If the product of activation derivatives and weights satisfies:

$$f_j'\left(\text{net}_j(t)\right) \sum_i w_{ij} < 1, \tag{4.33}$$

then the gradient decays exponentially and vanishes. Conversely, if

$$f_j'\left(\text{net}_j(t)\right) \sum_i w_{ij} > 1, \tag{4.34}$$

then the gradient can grow uncontrollably, leading to exploding gradients. To avoid this, most RNNs operate with gradients constrained below 1, which limits their ability to capture long-term dependencies.

To address this limitation, LSTM networks were proposed by Hochreiter and Schmidhuber [1997]. LSTMs introduce a memory cell equipped with gating mechanisms that regulate the flow of information and gradients over time. These gates include:

- **Input gate:** controls how much new information enters the cell.
- **Forget gate:** decides which information is discarded from the cell state.
- **Output gate:** determines what part of the cell state contributes to the output.

These gates enable the LSTM to maintain a constant error flow across many time steps, avoiding vanishing or exploding gradients.

This internal memory mechanism allows LSTMs to effectively capture both short- and long-term dependencies in sequential data, making them well-suited for PV energy forecasting [Zhong et al., 2024].

Extensions of LSTM, such as CNN-LSTM hybrids, enhance performance by combining convolutional layers for spatial feature extraction with LSTM units for tem-

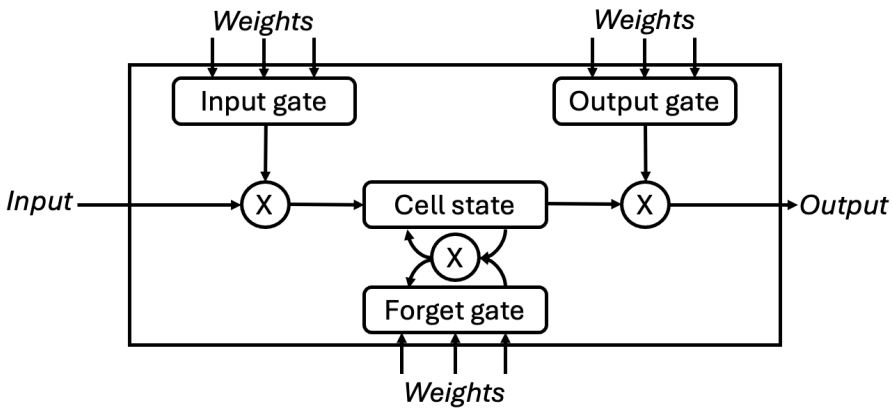

Figure 4.4: Schematic of a simplified LSTM cell, with weighted inputs modulated by input, output, and forget gates.

poral modelling. These models have demonstrated superior performance in PV forecasting tasks compared to standalone CNN, FNN, and RNN models [Fungtammasan and Koprinska, 2023].

In this thesis, LSTM networks are used in two contexts:

1. As post-processing components in the multiplicative approach, where the LSTM predicts the adjustment factor using recent observations in a one-step-ahead forecasting approach, rather than estimating the parameters of a probabilistic distribution as in DeepAR.

2. As direct predictors for next-day PV energy output, based on historical energy inputs.

4.2.2.2 Transformer

Despite the error gradient gating of LSTM, the sequential nature of RNNs limits their ability to model longer-range temporal dependencies. In contrast, the transformer architecture first proposed by [Vaswani et al., 2017] addresses these issues

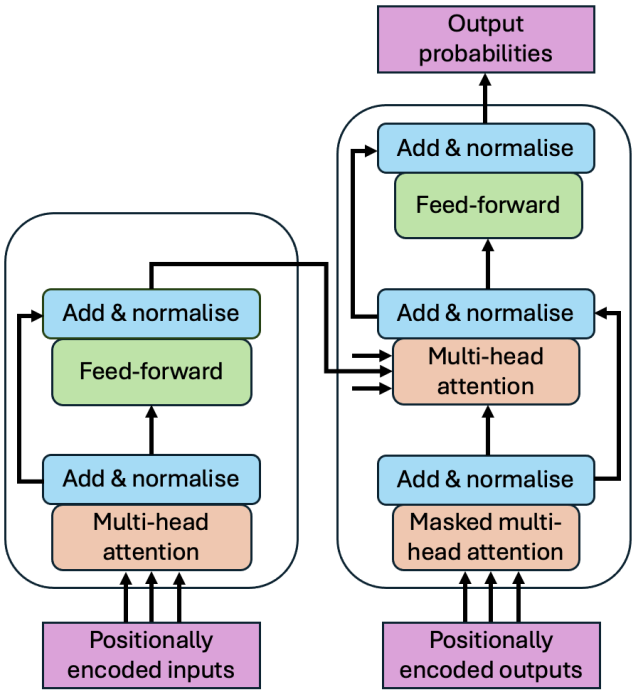

Figure 4.5: Schematic of simplified transformer model architecture. Outputs from encoder layers (left) feed into decoder layers (right) along with previous output data to generate a probability distribution.

by using attention mechanisms in place of any recurrence.

As shown in the figure above, all transformer models rely on tokenisers to convert text into discrete tokens. Since attention mechanisms are inherently permutation invariant, positional encodings are added to preserve sequence order. These encoded tokens are passed through encoder layers to produce contextualised representations. Decoder layers then transform these representations into probability distributions over the output vocabulary. As each new token is generated, it is appended to the sequence, and the process iterates autoregressively.

Transformers consist of multiple identical encoding layers, each containing a multi-head self-attention sublayer that receives all keys, values and queries from the previous layer. This allows each position in the encoder to attend to all posi-

tions in the previous layer of the encoder. Within the multi-head attention sublayer lies a scaled dot-product attention mechanism which generates a matrix of attention scores determined by the similarity between matrices of input keys K and queries Q. This is scaled relative to key vectors before conversion into a probability weight using a softmax function, then applied to value matrix V:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4.35}$$

Outputs from self-attention sublayers are each paired with an integrated FNN sublayer that further processes encoded tokens. At each sublayer, input values are added back to the output of the previous sublayer to reduce the risk of gradient exploding or vanishing, then normalised across features at each position to stabilise and accelerate learning. The final output from encoder layers act as sequence representations that encode contextual information from the attention mechanism and local transformations with positional context added by the FNN.

Multi-head self-attention mechanisms also exist in decoding layers, allowing each position in the decoder to attend to input queries from the previous decoder layer, as well as key-value pair outputs from the encoder. Consequently, every decoder layer can calculate values that factor in all positions in the input sequence.

Decoding layers consist of the same self-attention and FNN sublayers as encoding layers, with an additional multi-head cross-attention sublayer that leverages the output of all encoder layers. Additionally, a masked multi-head layer in decoders ensures that they only look at tokens in previous positions during time-series forecasting training, even when future data points would otherwise be available.

Transformer architectures have been adapted for use on various multivariate time-series datasets including for PV power predictions, often outperforming other RNN forecasting models such as LSTM and GRUs [Sherozbek et al., 2023]. How-

ever, these models require large amounts of data and careful tuning of temporal weighting; without these, they may be outperformed by simpler models [Su et al., 2025].

In this thesis, transformers are used in two contexts:

1. As post-processing components in the multiplicative approach, where the transformer predicts the adjustment factor using recent observations in a one-step-ahead forecasting approach.
2. As direct predictors for next-day PV energy output, based on historical energy inputs.

4.2.3 Tree-based Ensemble Methods

Supervised ML methods such as neural networks remain popular choices for forecasting due to their strong predictive performance and ability to model complex nonlinear relationships. However, they often suffer from challenges such as overfitting, high data requirements, and relatively limited generalisation outside the training range. Ensemble methods, which train two or more ML algorithms on a given dataset, address these limitations. By combining multiple supervised learning models with high bias and diversity, one can reduce overall variance and improve prediction stability through averaging effects. Tree-based ensemble methods, in particular, offer a balance between accuracy and interpretability while still being able to handle multidimensional data. This makes them attractive options as time-series forecasting models, including next-day PV power forecasting. The following section discusses the theoretical basis of two popular tree ensemble methods, random forest and XGBoost, as well as their relative strengths or weaknesses and applications in forecasting.

4.2.3.1 Random Forests

Random forests are ensembles of decision trees combined into a single, more flexible model with improved accuracy while still significantly less computationally expensive than neural network models [Heath et al., 1993]. For each tree, random subsets of classification or regression training data are sampled and used to train each weak base model. At each splitting node of trees, the training data subset is further sampled and the process is repeated until final predictions are made for voting or averaging. This sampling and training with high-variance models limits the chances of overfitting and learning noise [Ho, 1998].

Although random forests are typically applied to regression and classification data, they can be adapted to make sequential predictions as well. By using a sliding window of earlier observations as input features to train the model and future observations as the target variable, full decision trees using base models can be made in parallel that grow according to defined parameters, then generate predictions for future values with high bias and variance. The final outputs of each tree are then averaged to give a final prediction value. This averaging effect of base models further improves error and accuracy with each tree included.

Random forests have been used in forecasting and are highly efficient on one-step time series data with sufficient optimisation [Tyralis and Papacharalampous, 2017]. This extends to short-term PV forecasting, with [Singh et al., 2023] reporting improved performance over neural networks used on a rooftop PV dataset. However, their parallel construction of trees limits their potential for learning or further reduction of errors.

4.2.3.2 XGBoost

In contrast to random forests, the XGBoost algorithm introduced by [Chen and Guestrin, 2016] constructs trees sequentially and relies on the ML boost method to sequentially correct the residual error from previous trees, iteratively improving the prediction accuracy. This approach greatly reduces underfitting, as each subsequent tree acts to correct the inaccuracies of previous predictions. For each tree iteration t, the predicted value y for sample i can be calculated:

$$\hat{y}^i(t) = \hat{y}^i(t-1) + \eta f_t(x_i) \tag{4.36}$$

where $f_t(x_i)$ represents a newly generated tree at that iteration, trained to reduce the residual error. A learning rate η is applied to prevent over-correction. The partial first derivative

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \tag{4.37}$$

represents the rate of change of the loss function $l(y_i, \hat{y}_i^{(t-1)})$, which the model aims to minimise using iterative adjustments with each subsequent model. The partial second derivative

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} \tag{4.38}$$

represents the rate of change of the gradient. Through the second-order Taylor approximation of loss, XGBoost can draw confidence values in adjustments made by each model to accelerate convergence and minimise the number of trees that need to be made.

Consequently, XGBoost offers improvements over other decision trees like random forests in prediction accuracy by successively reducing bias and capturing more subtle trends in data through its reliance on residuals for modelling rather

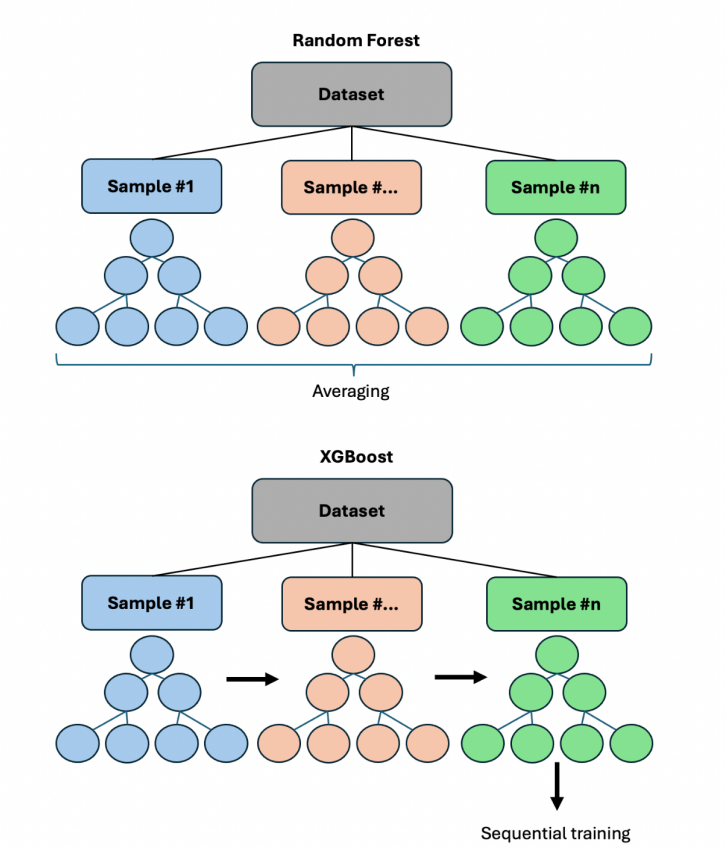

Figure 4.6: Schematic of simplified random forest and XGBoost decision trees. Whereas random forests (top) build complete decision trees in parallel to one another and average their outputs to generate final predictions, XGBoost (bottom) iteratively trains decision trees, using the error residuals of each previous model to fit the next model. The final prediction is a weighted sum of all tree outputs, optimised via gradient descent to minimise the loss function.

than averaging effects. These differences are particularly applicable to time series forecasting when using noisy or incomplete datasets, with [Lari et al., 2025] reporting XGBoost to display greater robustness and predictive performance than other decision trees.

Commercial Model: Quartz Quartz is a solar power forecasting model developed by [Open Climate Fix, 2024]. It employs gradient-boosted decision trees (via XGBoost) trained on over 25,000 PV sites and approximately five years of historical data. The model integrates nine NWP variables (such as GHI, temperature, wind, cloud cover, etc.) and generates forecasts up to 48 hours ahead at 15-minute intervals, making it adaptable for next-day (24-hour) prediction benchmarks. In this thesis, Quartz is used as a direct predictive baseline: given a PV site's latitude, longitude, and capacity, it generates next-day PV power forecasts at the corresponding resolution of the power output data, as described in Table 2.1.

4.3 Results

4.3.1 Cross-validation

Cross-validation in ML typically involves selecting a subset of data as a validation set to estimate model accuracy. For time-series data, specific methods like forward chaining and sliding-window validation are commonly used to preserve the sequential nature of the data. In the forward-chaining validation method, the training set size is incrementally expanded by adding new data points from the test set after each prediction. In contrast, the sliding-window validation method shifts the training and testing datasets forward over time, maintaining a fixed size for both windows. Therefore, sliding window validation is used when adapting to recent trends is required, while forward chaining validation is typically used to

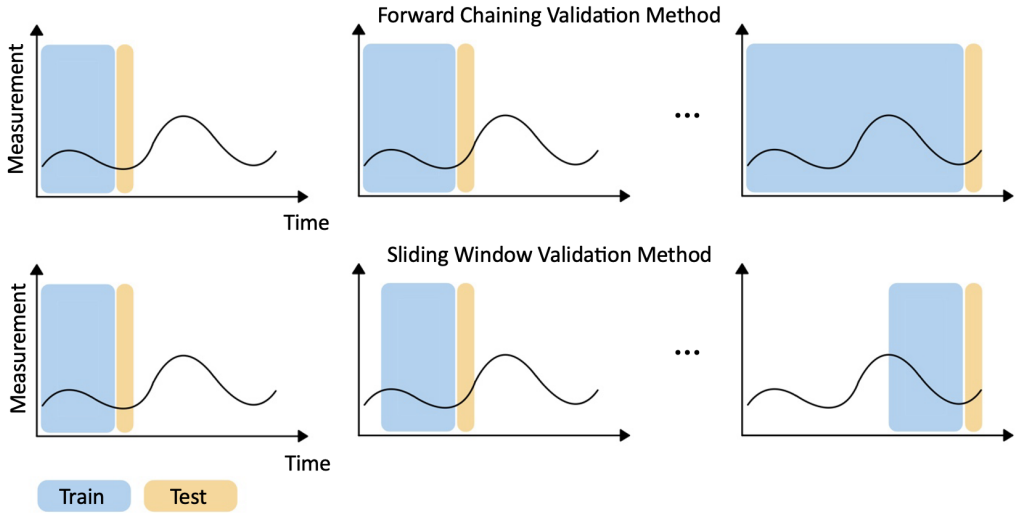

Figure 4.7: Time-series cross-validation strategies. Forward chaining: the training set expands over time by including new data after each step. Sliding-window: both training and testing windows move forward in time with fixed lengths, discarding older data.

capture long-term dependencies. A schematic diagram of these methods is shown in Figure 4.7.

The forward chaining validation method was used in this thesis to validate the accuracy of post-processing models due to the long-term temporal dependencies in the data. The initial training set consisted of the first 50% of the dataset. Testing was then performed in a one-day-ahead forecasting setting across typically two years of data (over 700 samples). This approach not only validates the one-step-ahead GP method but also allows other benchmarking statistical models to be evaluated using the same framework.

4.3.2 Benchmarking Results

In this section, we present the final power and energy results obtained using the hybrid multiplicative modelling framework proposed in this thesis. Model performance is evaluated using single-point error metrics including MAE, RMSE,

MAPE, and nRMSE across 12 sites in the UK and Hong Kong.

Two categories of models are assessed: post-processing models and direct forecasters. In the post-processing setting, a deterministic physical model ('model chain (AC)') is first applied. This chain accounts for weather data, transposition, module characteristics, and clipping, and serves as a baseline or 'first-try' estimate. The residual daily discrepancies are then post-processed using time-series GPs and a range of machine-learning benchmarks. These benchmarks include statistical approaches such as ARIMA and GluonTS, neural-'network architectures such as LSTM and transformer models, and ensemble-tree based methods such as XGBoost and random forest.

In contrast, the direct forecasting setting bypasses residual correction altogether, with models mapping historical observations directly to future energy values. In this case, direct predictors including GluonTS, LSTM, and transformer networks are implemented as machine learning forecasters in an autoregressive framework. Finally, Quartz, a commercial forecasting model developed by Open Climate Fix, is included as an external benchmark.

The following subsections report detailed results for UK sites, Hong Kong sites, and an overall performance comparison across all locations.

4.3.2.1 UK sites

Table 4.2 shows the errors and computational times associated with each model used for PV power and energy prediction at the Oxford site.

As observed, post-processing models consistently outperform direct forecasters, confirming the value of correcting residuals from the deterministic AC 'model chain'. Within this category, the time-series GP achieves the lowest error, with the NUTS variant marginally improving over MLE due to its broader uncertainty

bands from hyperparameter marginalisation, though at higher computational cost. GluonTS provides competitive accuracy but does not fully capture the heavy-tailed residual distribution of PV adjustment factors, due to its reliance on parametric likelihood approximation. Although ARIMA is the lowest-performing of the statistical post-processing models, its accuracy is broadly comparable to GluonTS and only around 10-15% worse than the best-performing model. This suggests that despite its linear autoregressive structure (Equation 4.25), ARIMA can still capture some of the underlying dynamics of PV time series. Neural networks (LSTM, transformer) deliver moderate accuracy on power predictions, but their data requirements and sensitivity to training length limit gains at site level. Ensemble tree methods (random forest, XGBoost) perform weakest, likely reflecting overfitting to the short sliding-window inputs and poor extrapolation on extreme days.

For direct forecasters, the AC model underperforms relative to the post-processed approaches. Neural network forecasters and GluonTS, when used in direct mode, also perform weakly compared to their post-processing counterparts, likely due to their limitations in a purely univariate setup and the relatively short historical training sequences available at site level. The commercial Quartz model achieves accuracy broadly comparable to the direct neural forecasters, though with slightly higher error magnitudes at Oxford. Similar patterns are observed across the other UK sites (see Table A.1 to Table A.5 in Appendix).

The error distribution plots (Figure 4.8) provide further insight. The relative prediction error time series shows that GluonTS tends to overpredict on high-irradiance days, while the AC model underpredicts during low-output periods. Both GP variants (MLE and NUTS) display narrower fluctuations, with errors more tightly centred around zero. These patterns highlight the advantage of GP post-processing in correcting systematic biases from the physical model. Across

Table 4.2: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the Oxford site. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s][3]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.34	0.56	23.00	247.17	552.88	11.05	2.03	2.98	24.08
Time series GP (NUTS)	0.34	0.55	286.00	239.44	547.79	10.95	1.98	2.94	22.93
GluonTS	0.35	0.57	11.00	277.55	581.19	11.61	2.136	3.11	26.20
ARIMA	0.35	0.571	24.33	277.18	578.98	11.57	2.17	3.06	27.9
LSTM (concurrent)[1]	0.36	0.58	168.30	268.75	575.13	11.49	2.83	3.70	31.99
LSTM (recurrent)[1]	0.35	0.58	190.30	290.79	586.70	11.72	3.53	4.29	38.05
Transformer	0.36	0.57	7.88	280.98	583.22	11.65	2.477	3.71	34.21
Random Forest	0.39	0.60	445.82	320.49	716.79	14.32	3.29	5.79	44.29
XGBoost	0.461	0.72	473.202	323.30	779.44	15.57	3.39	6.94	46.95
Direct Predictors									
Model Chain (AC)[2]	–	–	–	312.50	682.42	13.64	3.88	5.42	41.49
GluonTS	–	–	–	–	–	–	4.63	5.96	48.77
LSTM (concurrent)	–	–	–	–	–	–	6.97	8.90	59.92
LSTM (recurrent)	–	–	–	–	–	–	7.03	9.18	67.23
Transformer	–	–	–	–	–	–	7.08	9.02	65.59
On-the-market Model									
Quartz (OCF)	–	–	–	349.94	701.45	14.01	5.45	6.82	54.68

[1] LSTM (recurrent) corresponds to the standard recurrent architecture with hidden states carried forward (recurrent=True), whereas LSTM (feed-forward) corresponds to the non-recurrent variant (recurrent=False) that processes autoregressive inputs in a single pass.
[2] "Model Chain (AC)" refers to the deterministic physical model with clipping. It is evaluated over the test set to provide an apples-to-apples comparison with machine learning benchmarks.
[3] 'Computational time' here includes both training and inference time. For neural network models, this metric therefore combines long training times with relatively short inference times, whereas for Gaussian process models and other statistical approaches the distinction is less relevant.

the other UK sites, GP variants consistently concentrate most of their mass around the low-error region, while GluonTS exhibits a wider spread and the AC model alternates between under- and overprediction. The AC model behaviour is strongly linked to clipping losses and site-specific inverter and module specifications (see Figures A.1 to Figure A.5 in the Appendix). Across the 6 UK sites measured, the Quartz model exhibits error values comparable to GLuonTS direct predictions and consistently outperforms all other direct predictors.

4.3.2.2 HK sites

Table 4.3 shows the errors and computational times associated with each model used for PV power and energy prediction at the HK site A.

For HK site A, statistical post-processing models out-perform both NN-based and tree-based ensemble methods across error metrics for PV power predictions. Similar to the UK sites, the time-series GP NUTS model has the lowest MAE and RMSE for adjustment factor predictions, though greater cost of computational time. Similar trends can be observed in both small and large HK sites (see Appendix A.1.2). Direct model predictions have comparable PV power prediction error values to statistical post-processing models and only slightly worse cumulative PV energy prediction errors at HK site A. The AC model prediction performs similarly well at another small HK site C (see Table A.7 in Appendix). For HK small site B and large sites D-F, GP MLE and NUTS models display lower power and energy prediction errors across metrics in comparison to the AC model and other tested models at the same site (Tables A.6, A.8-A.10 in Appendix). The Quartz model generally outperforms NN and ensemble-based direct predictor models but is consistently worse than GluonTS direct energy predictions, possibly reflecting the greater heterogeneity in HK site data.

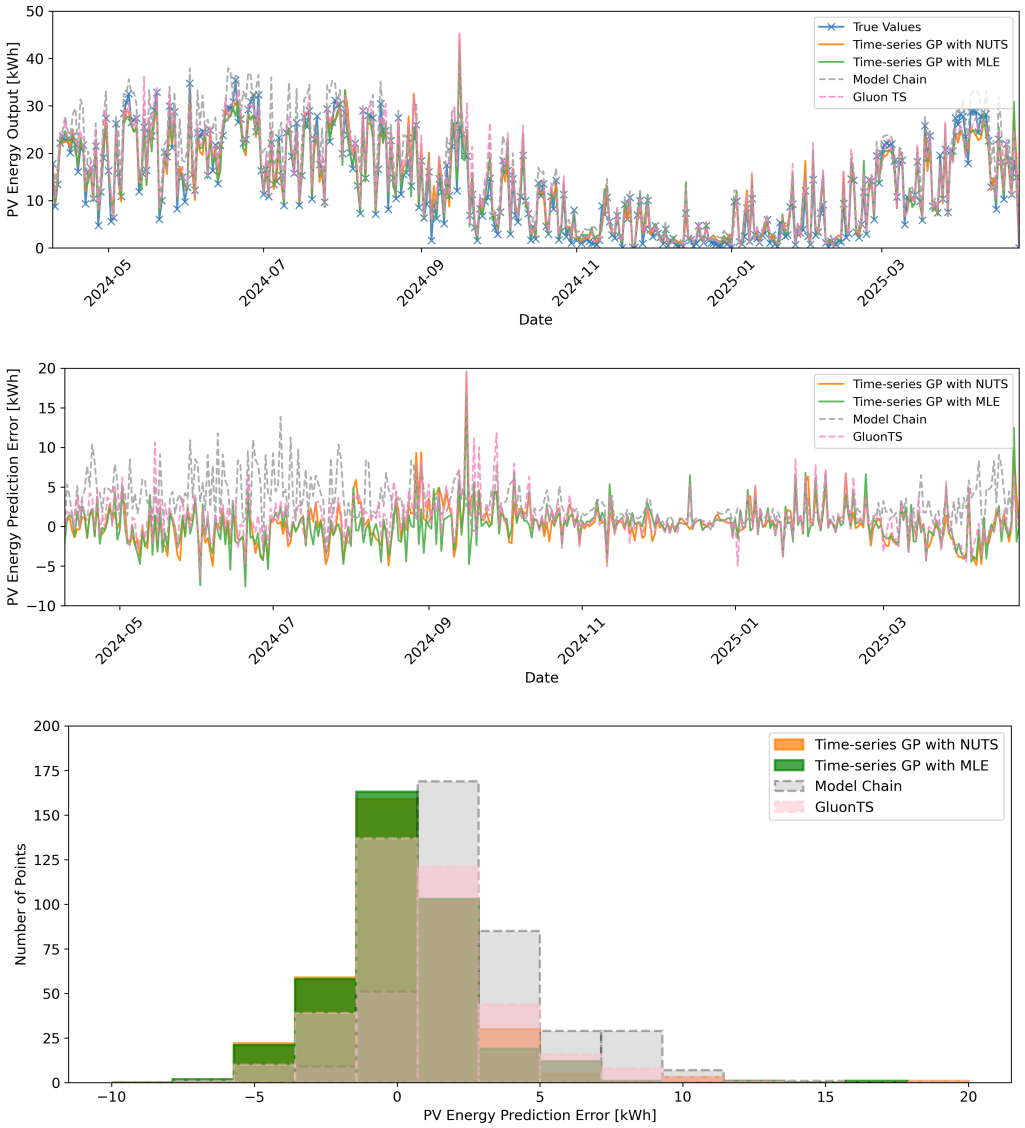

Figure 4.8: Oxford UK Site: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 5.005 kW (DC) while the peak AC capacity is approximately 3.96 kW. GluonTS predictions are shown in pink, GP MLE in green, GP NUTS in orange, the AC model in grey, and true values in blue.

Table 4.3: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for HK site A. Statistical models are highlighted in green, NN-based models in red and tree-based ensemble methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.11	0.24	27.46	76.34	156.68	6.53	0.97	1.16	22.97
Time series GP (NUTS)	0.10	0.15	299.00	76.22	151.63	6.32	1.02	1.22	24.00
GluonTS	0.11	0.17	5.11	79.48	159.53	6.65	1.09	1.48	30.12
ARIMA	0.11	0.15	27.73	514.69	722.77	32.78	1.40	2.09	59.10
LSTM (concurrent)	0.25	0.31	133.86	497.44	696.77	29.02	0.85	1.58	46.32
LSTM (recurrent)	0.40	0.66	145.62	497.17	696.41	29.01	0.85	1.58	46.32
Transformer	0.12	0.18	2.48	505.99	707.52	29.48	1.37	2.04	54.93
Random Forest	0.11	0.16	2710.14	518.16	728.56	30.35	1.48	2.17	59.97
XGBoost	0.11	0.16	557.23	492.81	693.56	28.90	1.30	1.90	53.03
Direct Predictors									
Deterministic Model (AC)	–	–	–	79.88	160.56	6.69	1.12	1.52	32.95
GluonTS	–	–	–	–	–	–	2.46	3.38	52.33
LSTM (concurrent)	–	–	–	–	–	–	2.85	3.70	75.13
LSTM (recurrent)	–	–	–	–	–	–	2.74	3.59	72.88
Transformer	–	–	–	–	–	–	2.65	3.40	70.42
On-the-market Model									
Quartz (OCF)	–	–	–	489.49	663.80	27.66	2.45	3.11	68.06

Figure 4.9 shows a comparison between the predicted energy outputs and errors of the AC model, GluonTS model, and time-series GP MLE and NUTS models for HK site A. All models closely align with observed PV energy output, with few isolated instances of extreme over- and under-prediction. Time-series GP cumulative PV energy error distributions are normally distributed around 0 with slightly more under-prediction, whereas the AC model and GluonTS error values are slightly left-skewed due to more consistent over-prediction. The same pattern is observed for all other HK sites (as seen in Appendix A.1.2), where GP model prediction errors are more normally distributed and closer to 0 relative to the GluonTS model and AC model.

4.3.2.3 Overall Performance

Figure 4.10 shows the overall power and energy prediction errors of each model, combining UK and HK (all sites) datasets. GP MLE and NUTS consistently outperform all other models in both power and energy prediction accuracy, highlighting their suitability for PV datasets. In general, statistical models exhibit the

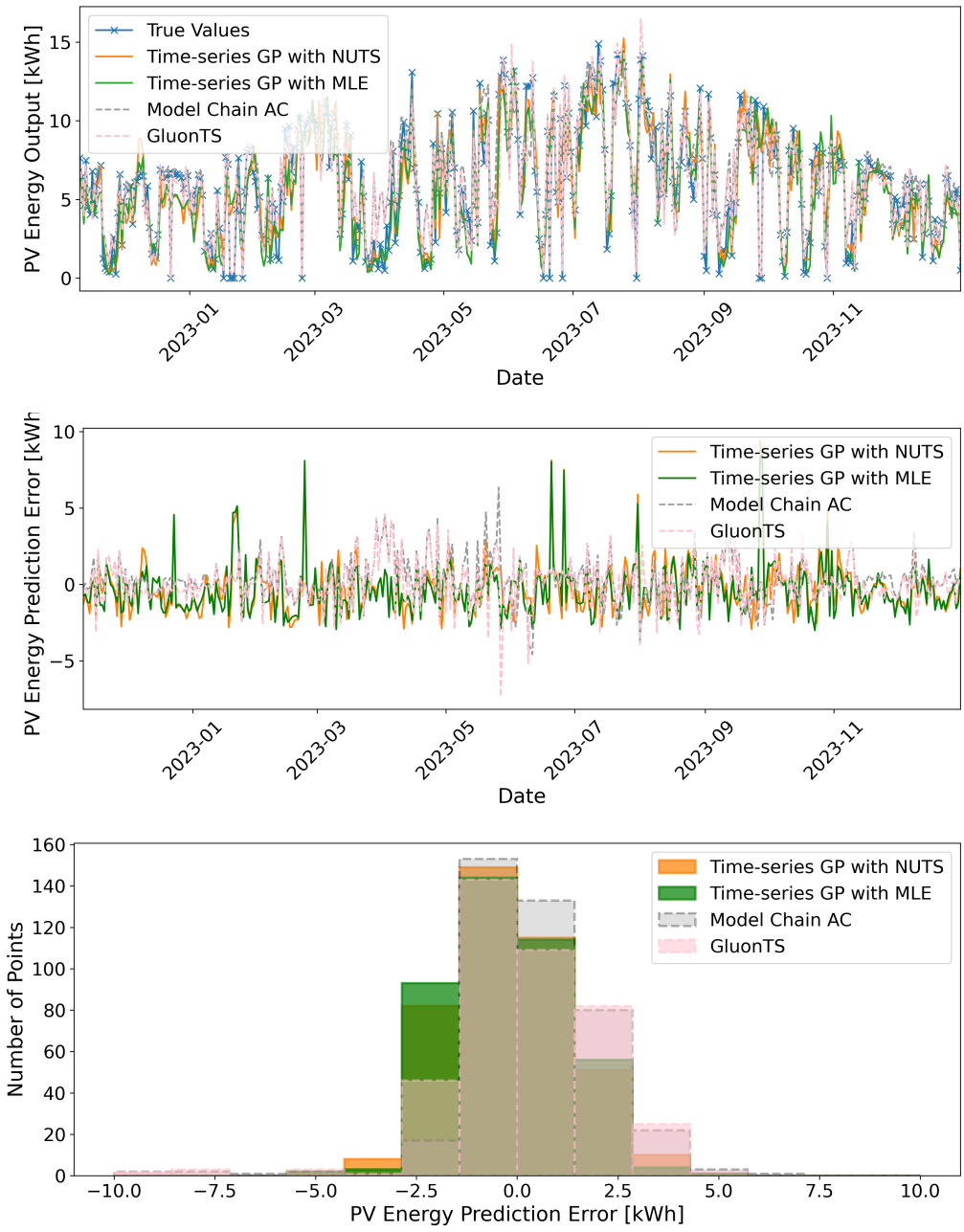

Figure 4.9: HK Site A: (Top) Comparison of observed PV day-ahead energy values with predictions made by the GluonTS, GP MLE and GP NUTS models, as well as the determinsitic AC model. (Middle) Day-ahead PV energy prediction errors for GLuonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead relative PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 2.40 kW while the peak capacity is approximately 2.04 kW. True observed values are shown in blue, GP NUTS in orange, GP MLE in green, deterministic AC model chain in grey, and GluonTS in pink.

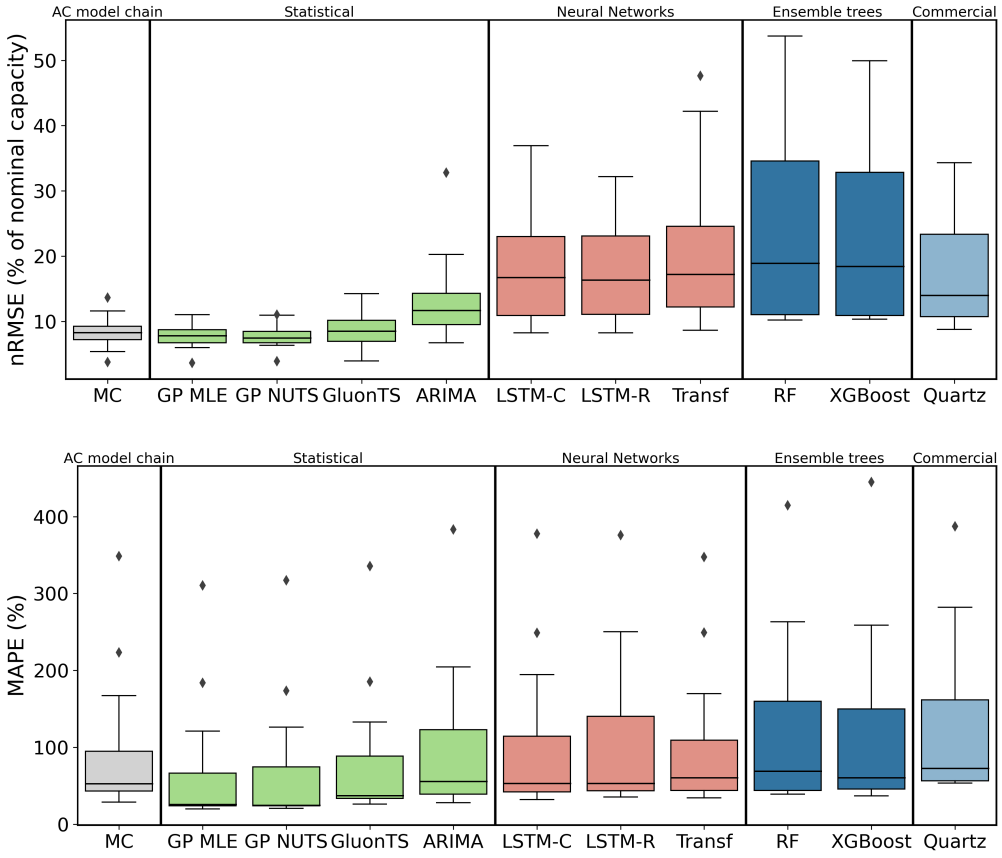

Figure 4.10: Overall performance comparison: (Top) Whisker plots of predicted 10-min horizon PV power nRMSE distribution across all sites for deterministic AC 'model chain', GP models, and all benchmarking models. (Bottom) Whisker plots of combined MAPE for day-ahead PV energy prediction at all sites, for deterministic AC 'model chain', GP models, and all benchmarking models. MC = AC 'model chain'. Statistical models shown in green, NN models in red (LSTM-C = concurrent, LSTM-R = recurrent, transf = transformer), ensemble tree methods in blue (RF = random forest), and Quartz model in cyan.

lowest medians and interquartile ranges, whereas LSTM and ensemble tree models have the largest spread in error. These high distributions represent the poor predictive stability of these models, likely caused by factors such as over-fitting or insufficient training at each site.

The OCF Quartz model achieves competitive power nRMSE relative to the post-processing models and consistently outperforms ensemble trees, indicating strong instantaneous prediction capability. However, its median energy MAPE is higher than other models, likely reflecting systematic daily biases that accumulate over time and are amplified by the shorter 15-minute prediction intervals. Quartz is trained on an XGBoost backbone using 9 NWP variables across over 25,000 PV sites, which strengthens its ability to generalise and adapt to different environments but limits fine-tuning at individual sites. As such, it appears optimised for short-term accuracy, whereas statistical post-processing models more effectively handle long-horizon bias corrections.

NN model prediction accuracy is commonly affected by irregularities in trends, for example due to sudden changes in cloud shading or other factors. Accuracy could be improved by combining a probabilistic NN that predicts distribution parameters at each time step with variable uncertainty, with an ensemble tree method to reduce variance. Similarly, the high prediction error of ensemble trees is often caused by residual bias and limited extrapolation when covariate distributions shift. These models could be combined with multivariate post-processing to learn any residual patterns not initially picked up.

4.4 Conclusion

This chapter evaluated GPs as Bayesian post-processors within a hybrid PV-forecasting pipeline and contrasted them with statistical, neural, and tree-based benchmarks in

both post-processing and direct-forecasting modes. ARGPs are particularly strong in capturing non-linear dynamics compared with classical AR models, owing to Bayesian uncertainty quantification and flexible kernel composition. However, time-series GPs outperformed ARGPs as the seasonal, long-term data was more naturally modelled as a function of time with composite kernels, whereas ARGPs tended to shrink toward the mean and under-represent variability given short site-level histories. Comparison between kernels found a combination of affine linear and Matern kernels was most effective for the PV datasets used.

Benchmarking GP performance against statistical, NN, and ensemble tree-based baselines across UK/HK sites revealed a clear hierarchy for both power and energy prediction: statistical post-processors (especially GPs) achieved the lowest errors, followed by neural networks, then ensemble trees. In particular, time-series GPs were the most reliable post-processors, delivering the best point accuracy and well-calibrated uncertainty, while ARIMA and GluonTS formed strong but inferior baselines and direct forecasts from the deterministic AC model were consistently worse.

Between Bayesian GP variants, NUTS typically yielded slightly lower errors than MLE, reflecting the benefits of full posterior exploration and hyperparameter uncertainty marginalisation. However, MLE outperformed NUTS at some sites, possibly due to insufficient site-level data making lower-variance point estimates advantageous in some cases.

5 | Conclusion

5.1 Summary of Findings

This thesis proposed and evaluated a multiplicative hybrid approach for next-day PV power and energy forecasting that combines a deterministic physics-based model with a GP post-processor that corrects residual site-specific errors using probabilistic and kernel-based learning. Pairing mechanistic priors with data-driven flexibility in this manner overcomes common challenges associated with PV datasets outlined in Chapter 1, including seasonality, gaps in the datasets, and sudden weather-induced irregularities.

The application of both time-series GPs and ARGPs to deterministic physical-based model was explored, along with kernel composition and hyperparameter choice. Across variants, the time-series GP more faithfully captured trends than the ARGP, which tended to regress toward the mean in highly variable regimes. Comprehensive benchmarking was then performed against multiple statistical, NN-driven, and ensemble tree-based models on a total of 12 datasets varying in size and location between HK and the UK. Diverse power and energy prediction error metrics revealed GP post-processing to consistently outperform all other post-processing models in addition to direct predictions by the deterministic AC model, effectively adapting to site heterogeneity and outliers. Relative to commercial baselines, Quartz showed strong instantaneous power predictive accuracy but weaker day-ahead energy predictive accuracy, reinforcing the strength of post-processing in correcting biases at daily aggregation steps. Additionally, the NUTS model with full Bayesian inference generally yielded lower errors than MLE at the cost of greater computational time due to hyperparameter uncertainty marginalisation. However, MLE outperformed all other models, including NUTS,

at sites with limited data or weak priors where lower-variance point estimates were preferable.

Overall, the hybrid approach proposed shows promise for real-world applications due to its interpretability, ease of implementation and modest data requirements in addition to high performance. By combining physically grounded priors with probabilistic residual learning in a multiplicative form, the proposed multiplicative GP-augmented approach delivers state-of-the-art performance across diverse sites while laying a clear path toward scalable, uncertainty-aware forecasting in real-world operations.

5.2 Future Directions

5.2.1 Using Additional Weather Data

Ensemble NWP inputs, such as GHI, temperature, wind, or cloud cover, can be incorporated into the hybrid model to improve conditioning. Ensembles provide both a central tendency and a spread that reflect meteorological uncertainty; conditioning on this information can reduce systematic bias and improve uncertainty quantification.

This may be achieved by applying ensemble post-processing to NWP irradiance upstream of the physical model for calibrated probabilistic irradiance, which can be propagated through the model chain to generate an ensemble of PV power trajectories. GPs may post-process each trajectory before aggregating the resulting distributions or post-processing ensemble summaries of trajectories. Alternatively, the summary features of the ensemble may be directly fed as covariates into the GP for the multiplicative adjustment model to enable state-dependent corrections. For example, a high ensemble spread or the presence of specific regime

indicators would cause the GP to increase expected adjustment or predictive variance.

Conditioning models in this way reduces the impact of systematic NWP biases, stabilising and shrinking residuals, thus increasing the stability of kernel hyperparameters. It may also yield better-calibrated forecasts by improving model flexibility, as ensemble diagnostics may act as indicators of environmental factors such as cloud cover.

5.2.2 Multivariate Residual Learning

Instead of modelling a single univariate adjustment factor solely as a function of time, the residual model can be extended to incorporate multiple inputs and outputs. By introducing a feature vector into the GP, additional information such as ensemble statistics, physical-model diagnostics, site metadata, calendar effects, and recent residuals can be included. This enables the model to learn state-dependent corrections — for example, applying larger multiplicative adjustments on days with high ensemble spread or during periods of inverter clipping.

Conditioning the residual learner on a rich set of covariates helps stabilize and shrink residuals, thus enhancing robustness to regime shifts such as seasonal amplitude changes or persistent clipping. Multi-task sharing across sites can further reduce the data requirements for each location and improve imputation on days with missing sensor data. Combined with sparse or variational multi-output GP implementations, this approach remains computationally scalable while providing more accurate and physically interpretable corrections.

5.2.3 GP Approximations

Current Gaussian process (GP) inference scales as $\mathcal{O}(n^3)$ with respect to the number of training points n, due to the Cholesky factorisation of the covariance matrix. This cubic complexity increases both computational time and memory cost, potentially leading to slower end-to-end predictions compared to lighter models.

Multiple principled approximations exist that can reduce complexity, including but not limited to:

State-space GPs: For one-dimensional time series and kernels that admit a state-space (SDE) representation, such as the Matérn family with $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$, the GP can be expressed as a linear dynamical system and inferred using Kalman filtering and smoothing [Svensson et al., 2015]. This reduces both time and memory complexity to $\mathcal{O}(n)$, supports online updates, and is particularly well-suited to long histories or high-frequency (intra-day) data.

Sparse or variational GPs: By introducing a small set of m inducing points to summarise the process, the full covariance structure can be approximated using an inducing kernel and corresponding cross-covariances. The model parameters are typically learned by maximising a variational evidence lower bound (ELBO) in mini-batches, resulting in training complexity $\mathcal{O}(nm^2)$ and prediction cost $\mathcal{O}(m^2)$. This approach is particularly suitable for day-ahead energy forecasting with moderate amounts of historical data (hundreds to thousands of daily samples), preserving predictive accuracy while significantly reducing computational and memory demands.

Structured kernel interpolation (SKI): Inducing points are placed on a regular grid and interpolated to true inputs using sparse weighted interpolation. This

structure allows the use of iterative solvers with near-linear complexity between $\mathcal{O}(n)$ and $\mathcal{O}(nm^2)$, depending on grid density and input dimensionality. SKI performs best for stationary kernels and low-dimensional input spaces.

Stochastic variational inference (SVI): Replacing or complementing MLE or NUTS-based inference with SVI enables approximate posterior inference over hyperparameters at greatly reduced cost. The variational posterior is optimised using stochastic gradient updates over mini-batches, retaining meaningful uncertainty estimates while enabling scalability to larger datasets.

The use of sparse or variational GP approximations tailored to time-series data would substantially improve computational efficiency, reducing latency and energy consumption while maintaining accuracy. These methods enhance scalability across larger datasets and enable the use of more complex models for improved predictive performance. However, care must be taken, as aggressive approximation may lead to underestimation of predictive variance.

Bibliography

Ahmed S. Abbas, Ragab A. El-Sehiemy, Adel Abou El-Ela, Eman S. Ali, Karar Mahmoud, Matti Lehtonen, and Mohamed M. F. Darwish. Optimal harmonic mitigation in distribution systems with inverter based distributed generation. *Applied Sciences*, 11(2):774, 2021. 10.3390/app11020774.

Sameer Al-Dahidi, Muthusamy Madhiarasan, Laith Al-Ghussain, Ahmed Abubaker, Adnan Ahmad, Mohammed Alrbai, Masoud Aghaei, Hussein Alahmer, Ali Alahmer, Paolo Baraldi, et al. Forecasting solar photovoltaic power production: A comprehensive review and innovative data-driven modeling framework. *Energies*, 17(16):4145, 2024. 10.3390/en17164145.

A. S. Al-Ezzi and M. N. M. Ansari. Photovoltaic solar cells: A review. *Applied System Innovation*, 5(4):67, 2022. ISSN 2571-5577. 10.3390/asi5040067.

Idris Al Siyabi, Sourav Khanna, Senthilarasu Sundaram, and Tapas Mallick. Experimental and numerical thermal analysis of multi-layered microchannel heat sink for concentrating photovoltaic application. *Energies*, 12(1):122, 2019. 10.3390/en12010122.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. Gluonts: Probabilistic time series models in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

Mauricio Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4 (3):195–266, 2012.

A. Amiri, A. Chouder, H. Oudira, S. Silvestre, and S. Kichou. Improving photovoltaic power prediction: Insights through computational modeling and feature selection. *Energies*, 17:3078, 2024.

J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de Pison, and F. Antonanzas-Torres. Review of photovoltaic power forecasting. *Solar Energy*, 136:78–111, October 2016. 10.1016/j.solener.2016.06.069.

J. Antonanzas, D. Pozo-Vázquez, L. A. Fernandez-Jimenez, and F. J. Martinez-de Pison. The value of day-ahead forecasting for photovoltaics in the spanish electricity market. *Solar Energy*, 158:140–146, 2017. ISSN 0038-092X. 10 .1016/j.solener.2017.09.016.

J. Benchimol et al. The uk solar energy resource and the impact of climate change. *Renewable Energy*, 64:203–213, 2014.

John Boland. Time series modeling of solar radiation. In *Modeling Solar Radiation at the Earth's Surface*, pages 189–224. Springer, 2015. 10.1007/978-3-319-18344-3_8.

John Boland. Characterising seasonality of solar radiation and solar farm output. *Energies*, 13(2):471, 2020. 10.3390/en13020471. URL https://doi.org/10.3390/en13020471.

George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1970.

S. Chahboun and M. Maaroufi. Ensemble methods comparison to predict the power produced by photovoltaic panels. In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning – Volume 1: BML*, pages 474–478. SciTePress, 2021. ISBN 978-989-758-559-3. 10.5220/0010736800003101.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 2016. ACM. 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

X. Chen, H. Yue, and Y. Sato. The impact of data augmentation on ai-driven predictive algorithms. *Processes*, 13(4):1195, 2024. 10.3390/pr13041195.

Y. Chen and X. Zhao. A review of photovoltaic power forecasting: Models, challenges and future directions. *Journal of Big Data*, 10(1):1–25, 2023. 10.1186/s40537-023-00705-8.

G. Chicco, V. Cocina, P. Di Leo, F. Spertino, and A. Massi Pavan. Error assessment of solar irradiance forecasts and ac power from energy conversion model in grid-connected photovoltaic systems. *Energies*, 9(1):8, 2016.

Janae Csavina, Jason Field, Omar Félix, Alba Y. Corral-Avitia, A. Eduardo Sáez, and Eric A. Betterton. Effect of wind speed and relative humidity on atmospheric dust concentrations in semi-arid climates. *Science of the Total Environment*, 487:82–90, 2014. 10.1016/j.scitotenv.2014.03.138. PMCID: PMC4072227.

Pietro Di Leo, Gianfranco Chicco, Vito Cocina, Fabio Spertino, and Alfonso Massi Pavan. Photovoltaic power forecasting using numerical weather prediction and satellite data. *Sensors*, 21(20):6840, 2021. 10 .3390/s21206840. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8531863/.

European Centre for Medium-Range Weather Forecasts. ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global cli-

mate. https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, 2017. Accessed 2025-06-09.

European Centre for Medium-Range Weather Forecasts. Integrated forecasting system (ifs). https://www.ecmwf.int/en/forecasts/documentation-and-support/medium-range-forecasts, 2023. Accessed 2025-06-09.

J. Fang, L. Zheng, and C. Liu. A novel method for missing data reconstruction in smart grid using generative adversarial networks. *IEEE Transactions on Industrial Informatics*, 20(3):4408–4417, 2023. 10.1109/TII.2023.3241234.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991. 10.1214/aos/1176347963.

G. Fungtammasan and I. Koprinska. Convolutional and lstm neural networks for solar power forecasting. In *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, June 2023. 10 .1109/IJCNN54540.2023.10191706.

Carina L. Gargalo, Alina A. Malanca, Adem R. N. Aouichaoui, Jakob K. Huusom, and Krist V. Gernaey. Navigating industry 4.0 and 5.0: the role of hybrid modelling in (bio)chemical engineering's digital transition. *Frontiers in Chemical Engineering*, 6:1494244, 2024. 10.3389/fceng.2024.1494244.

V. Gayathry, Deepa Kaliyaperumal, and Surender Reddy Salkuti. Seasonal solar irradiance forecasting using artificial intelligence techniques with uncertainty analysis. *Scientific Reports*, 14(1):17945, 2024. 10.1038/s41598-024-56253-3.

W. Glassley, J. Kleissl, C. van Dam, H. Shiu, J. Huang, G. Braun, and R. Holland. California renewable energy forecasting, resource data and mapping. Technical

Report CEC-500-2014-026, California Energy Commission, 2011. Accessed: 2025-06-04.

Tilmann Gneiting, Adrian E Raftery, Anton H Westveld III, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

Jonatan Hartikainen and Simo Särkkä. Kalman filtering and smoothing solutions to temporal gaussian process regression models. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 379–384. IEEE, 2010.

John E Hay and James A Davies. Calculations of the solar radiation incident on an inclined surface. *Proceedings of First Canadian Solar Radiation Data Workshop*, pages 59–72, 1980.

David Heath, Simon Kasif, and Steven Salzberg. k-dt: A multi-tree learning method. In *Proceedings of the Second International Workshop on Multistrategy Learning*, pages 138–149, 1993.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. 10.1109/34.709601. URL https://ieeexplore.ieee.org/document/709601.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. 10.1162/neco.1997.9.8.1735.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively

setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

William F Holmgren, Clifford W Hansen, and Mark A Mikofski. pvlib python: A python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, 2018. 10.21105/joss.00884.

Nina Horat, Sina Klerings, and Sebastian Lerch. Improving model chain approaches for probabilistic solar energy forecasting through post-processing and machine learning. *Advances in Atmospheric Sciences*, 41(6):1234–1256, 2024. 10.1007/s00376-024-4219-2.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, March 1964. 10.1214/aoms/1177703732.

Pierre Ineichen and Richard Perez. A new airmass independent formulation for the linke turbidity coefficient. *Solar Energy*, 73(3):151–157, 2002. ISSN 0038-092X. 10.1016/S0038-092X(02)00045-2.

R.H. Inman, H.T.C. Pedro, and C.F.M. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39 (6):535–576, 2013. 10.1016/j.pecs.2013.06.002. URL https://doi.org/10.1016/j.pecs.2013.06.002.

Md. Samin Safayat Islam, Puja Ghosh, Md.Õmer Faruque, Md.Ñashidul Islam, Md.Ãlamgir Hossain, Md.Šhafiul Alam, and Md.Ñafiqul Islam Sheikh. Optimizing short-term photovoltaic power forecasting: A novel approach with gaussian process regression and bayesian hyperparameter tuning. *Processes*, 12(3): 546, 2024. 10.3390/pr12030546. URL https://doi.org/10.3390/pr12030546.

Arnulf Jäger-Waldau. Snapshot of photovoltaics – may 2023. *EPJ Photovoltaics*,

14:23, 2023. 10.1051/epjpv/2023016. Viewpoint article. Published online 25 July 2023.

Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90(12):1–37, 2019. 10.18637/jss.v090.i12.

Hussein A. Kazem and Miqdam Tariq Chaichan. Effect of humidity on photovoltaic performance based on experimental study. *International Journal of Applied Engineering Research*, 10(23):43572–43577, 2015. Experimental study conducted in Sohar, Oman (July–September 2015).

Ali Jassim Lari, Antonio P. Sanfilippo, Dunia Bachour, and Daniel Perez-Astudillo. Using machine learning algorithms to forecast solar energy power output. *Electronics*, 14(5):866, 2025. 10.3390/electronics14050866. URL https://www.mdpi.com/2079-9292/14/5/866.

Andi A. H. Lateko, Hong-Tzer Yang, and Chao-Ming Huang. Short-term pv power forecasting using a regression-based ensemble method. *Energies*, 15 (11):4171, 2022. 10.3390/en15114171.

Philippe Lauret, Rodrigo Alonso-Suárez, Josselin Le Gal La Salle, and Mathieu David. Solar forecasts based on the clear sky index or the clearness index: Which is better? *Solar*, 2(4):432–444, 2022. 10.3390/solar2040026. URL https://doi.org/10.3390/solar2040026.

Paolo Di Leo, Alessandro Ciocia, Gabriele Malgaroli, and Filippo Spertino. Advancements and challenges in photovoltaic power forecasting: A comprehensive review. *Energies*, 18(8):2108, 2025. 10.3390/en18082108.

Q. Li, Y. Xu, B. S. H. Chew, H. Ding, and G. Zhao. An integrated missing-data tolerant model for probabilistic pv power generation forecasting. *IEEE*

Transactions on Power Systems, 37:4447–4459, 2022. 10.1109/TPWRS.2022.3146989.

Zinan Lin, Qi Zhou, Zhe Wang, Ce Wang, Davis Boyd Bookhart, and Marcus Leung-Shea. A high-resolution three-year dataset supporting rooftop photovoltaics (pv) generation analytics. Dryad Digital Repository, 2024. URL https://doi.org/10.5061/dryad.m37pvmd99.

Dong C. Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 6(1):1–39, 2012. 10.1145/2133360.2133363.

Wei Liu, Chao Ren, and Yan Xu. Pv generation forecasting with missing input data: A super-resolution perception approach. *IEEE Transactions on Sustainable Energy*, 12(2):1493–1496, 2021. 10.1109/TSTE.2020.3029731.

P. Löper, D. Pysch, A. Richter, M. Hermle, S. Janz, M. Zacharias, and S. W. Glunz. Analysis of the temperature dependence of the open-circuit voltage. *Energy Procedia*, 27:135–142, 2012. 10.1016/j.egypro.2012.07.041. Available online 25 August 2012.

Javier López Gómez, Ana Ogando Martínez, Francisco Troncoso Pastoriza, Lara Febrero Garrido, Enrique Granada Álvarez, and José Antonio Orosa García. Photovoltaic power prediction using artificial neural networks and numerical weather data. *Sustainability*, 12(24):10295, 2020. 10.3390/su122410295.

Elke Lorenz, Thomas Scheidsteger, Jens Hurka, Detlev Heinemann, and Christian Kurz. Regional pv power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, 19(7):757–771, November 2011.

Elke Lorenz, Jan Kuehnert, and Detlev Heinemann. Short term forecasting of solar irradiance by combining satellite data and numerical weather predictions. In *Proceedings of the 27th European Photovoltaic Solar Energy Conference and Exhibition*. WIP-Munich, 2012.

B. Marion, J. Adelstein, and K. Boyle. Performance parameters for grid-connected PV systems. In *Proc. of the 31st IEEE Photovoltaic Specialists Conference*, pages 1601–1606, Lake Buena Vista, FL, USA, 2005. IEEE. 10.1109/PVSC.2005.1488451.

M.J. Mayer and G. Gróf. Extensive comparison of physical models for photovoltaic power forecasting. *Applied Energy*, 283:116239, February 2021. 10.1016/j.apenergy.2020.116239.

M.J. Mayer and D. Yang. Probabilistic photovoltaic power forecasting using a calibrated ensemble of model chains. *Renewable and Sustainable Energy Reviews*, 168:112821, October 2022. 10.1016/j.rser.2022.112821.

MJ Mayer and D Yang. Pairing ensemble numerical weather prediction with ensemble physical model chain for probabilistic photovoltaic power forecasting. *Renewable and Sustainable Energy Reviews*, 175:113171, Apr 2023. 10.1016/j.rser.2023.113171.

M.J. Mayer et al. Improving model chain approaches for probabilistic solar energy forecasting through post-processing and machine learning. arXiv preprint arXiv:2406.04424, 2024.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Karim Menoufi. Dust accumulation on the surface of photovoltaic panels: In-

troducing the photovoltaic soiling index (pvsi). *Sustainability*, 9(6):963, 2017. 10.3390/su9060963.

Leonardo Micheli, Matthew Muller, Marios Theristis, Greg P. Smestad, Florencia Almonacid, and Eduardo F. Fernandez. Quantifying the impact of inverter clipping on photovoltaic performance and soiling losses. *Renewable Energy*, 225: 120317, 2024. 10.1016/j.renene.2024.120317.

Brian Mirletz and D. Guittet. Impacts of dispatch strategies and forecast errors on the economic value of pv-plus-battery systems. Technical Report NREL/TP-6A20-86194, National Renewable Energy Laboratory, 2023. URL https://www.nrel.gov/docs/fy23osti/86194.pdf. Accessed: 2025-06-04.

National Centers for Environmental Prediction (NCEP). Global Forecast System (GFS). https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast, 2020. Accessed 2025-06-09.

U.S. Department of Energy. Solar forecasting: Maximizing its value for grid integration. Technical report, U.S. Department of Energy, July 2016. URL https://www.energy.gov/sites/prod/files/2016/08/f33/Solar%20Forecasting%20White%20Paper_SunShot_2016.pdf. Accessed: 2025-06-04.

E. Ogliari, A. Dolara, G. Manzolini, and S. Leva. Physical and hybrid methods comparison for the day ahead pv output power forecast. *Renewable Energy*, 113:11–21, 2017a. 10.1016/j.renene.2017.05.034.

Emanuele Ogliari, Alberto Dolara, Giampaolo Manzolini, and Sonia Leva. Physical and hybrid methods comparison for the day ahead pv output power forecast. *Renewable Energy*, 113:11–21, December 2017b. ISSN 0960-1481. 10.1016/j.renene.2017.05.046.

Open Climate Fix. Uk pv power forecasting dataset. https://huggingface.co/datasets/openclimatefix/uk_pv, 2022. Accessed 2025-06-09.

Open Climate Fix. Open-source quartz solar forecast. https://github.com/openclimatefix/open-source-quartz-solar-forecast, 2024. Accessed 2025-07-21.

Michael A. Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimization and Quadrature*. PhD thesis, University of Oxford, 2010. URL https://www.robots.ox.ac.uk/~mosb/public/pdf/2160/full_thesis.pdf.

Michael A Osborne, Roman Garnett, and Stephen J Roberts. Active data selection for sensor networks with faults and changepoints. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 533–540. IEEE, 2010. 10.1109/AINA.2010.107.

Justyna Pastuszak and Paweł Węgierek. Photovoltaic cell generations and current research directions for their development. *Materials*, 15(16):5542, 2022. 10.3390/ma15165542.

Hamed H. Pourasl, Reza Vatankhah Barenji, and Vahid M. Khojastehnezhad. Solar energy status in the world: A comprehensive review. *Energy Reports*, 10:3474–3493, 2023. 10.1016/j.egyr.2023.10.022. URL https://www.sciencedirect.com/science/article/pii/S2352484723014579.

Joaquin Quinonero-Candela, Agathe Girard, Jan Larsen, and Carl Edward Rasmussen. Propagation of uncertainty in bayesian kernel models—application to multiple-step ahead forecasting. In *IEEE International Conference on Acous-*

tics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., volume 2, pages II–701. IEEE, 2003.

Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. URL https://gaussianprocess.org/gpml/.

Ibrahim Reda and Afshin Andreas. Solar position algorithm for solar radiation applications. *Solar Energy*, 76(5):577–589, 2004. 10.1016/j.solener.2003.12.003.

Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems 30: Proceedings of NIPS 2017*, volume 30 of *Advances in Neural Information Processing Systems*, pages 4645–4654. Curran Associates Inc., 2017. URL https://papers.nips.cc/paper/7050-non-stationary-spectral-kernels.pdf. Available at https://arxiv.org/abs/1705.08736.

Stephen J Roberts, Michael A Osborne, Mark Ebden, Steve Reece, Neil Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.

Maja Rudolph, Stefan Kurz, and Barbara Rakitsch. Hybrid modeling design patterns. *Journal of Mathematics in Industry*, 14:3, 2024. 10.1186/s13362-024-00141-0.

M. Schmelas, T. Feldmann, J. da Costa Fernandes, and E. Bollin. Photovoltaics energy prediction under complex conditions for a predictive energy management system. *Journal of Solar Energy Engineering*, 137(3):031015, 2015. 10.1115/1.4029662.

Lina M. Shaker, Ahmed A. Al-Amiery, and Mahdi M. Hanoon. Examining the influence of thermal effects on solar cells: A comprehensive review. *Sustainable Energy Research*, 11:1–18, 2024. 10.1186/s40807-024-00100-8.

Jumaboev Sherozbek, Jaewoo Park, Mohammad Shaheer Akhtar, and O-Bong Yang. Transformers-based encoder model for forecasting hourly power output of transparent photovoltaic module systems. *Energies*, 16(3):1353, 2023. 10 .3390/en16031353.

Talha Ahmad Siddiqui, Samarth Bharadwaj, and Shivkumar Kalyanaraman. A deep learning approach to solar-irradiance forecasting in sky-videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2272–2280. IEEE, 2019. 10.1109/WACV.2019.00234.

Upma Singh, Shekhar Singh, Saket Gupta, Majed A. Alotaibi, and Hasmat Malik. Forecasting rooftop photovoltaic solar power using machine learning techniques. *Sustainable Energy Technologies and Assessments*, 58:103222, 2023. 10.1016/j.seta.2023.103222. URL https://www.sciencedirect.com/science/article/pii/S2213138823001670.

Solcast. Solcast api toolkit. https://solcast.com/, 2022. Accessed 2025-06-09.

Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58(3):80, 2025. 10.1007/s10462-024-11044-2.

Andreas Svensson, Arno Solin, Simo Särkkä, and Thomas B. Schön. Computationally efficient bayesian learning of gaussian process state space models. *arXiv preprint*, 2015.

S. Theocharides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, and G.E. Georghiou. Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Applied Energy*, 268: 115023, 2020. 10.1016/j.apenergy.2020.115023.

Spyros Theocharides, George Makrides, and George E. Georghiou. Pv generation forecasting utilizing a classification-only approach. *EPJ Photovoltaics*, 15:12, 2024. 10.1051/epjpv/2024011. URL https://www.epj-pv.org/articles/epjpv/full_html/2024/01/pv230028/pv230028.html. Accessed: 2025-06-04.

Richard E. Turner. *Statistical Models for Natural Sounds*. PhD thesis, University of Cambridge, 2011.

Hristos Tyralis and Georgia Papacharalampous. Variable selection in time series forecasting using random forests. *Algorithms*, 10(4):114, 2017. 10.3390/a10040114. URL https://www.mdpi.com/1999-4893/10/4/114.

R. Ulbricht, U. Fischer, W. Lehner, and H. Donker. First steps towards a systematical optimized strategy for solar energy supply forecasting. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 1–12, 2013. URL https://www.mdpi.com/1996-1073/15/9/3320. Accessed: 2025-06-04.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

K. Wang, X. Qi, and H. Liu. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. *Applied Energy*, 251:113315, 2019. 10.1016/j.apenergy.2019.113315.

Wei Wang, Da Yang, Tao Hong, and Jan Kleissl. An archived dataset from the ecmwf ensemble prediction system for probabilistic solar power forecasting. *Solar Energy*, 248:64–75, December 2022.

X. Xiang, X. Li, Y. Zhang, and J. Hu. A short-term forecasting method for photovoltaic power generation based on the tcn-ecanet-gru hybrid model. *Scientific Reports*, 14:6744, 2024. 10.1038/s41598-024-53728-6.

Da Yang, Eric Wu, and Jan Kleissl. Operational solar forecasting for the real-time market. *International Journal of Forecasting*, 35(4):1499–1519, September 2019.

G.P. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003. 10.1016/S0925-2312(01)00702-0.

Yuanchang Zhong, Tengfei He, and Zhongyuan Mao. Enhanced solar power prediction using attention-based dipls-bilstm model. *Electronics*, 13(23):4815, December 2024. ISSN 2079-9292. 10.3390/electronics13234815.

A | Appendix

A.1 PV Power and Energy Forecast Results

A.1.1 UK sites deterministic model results

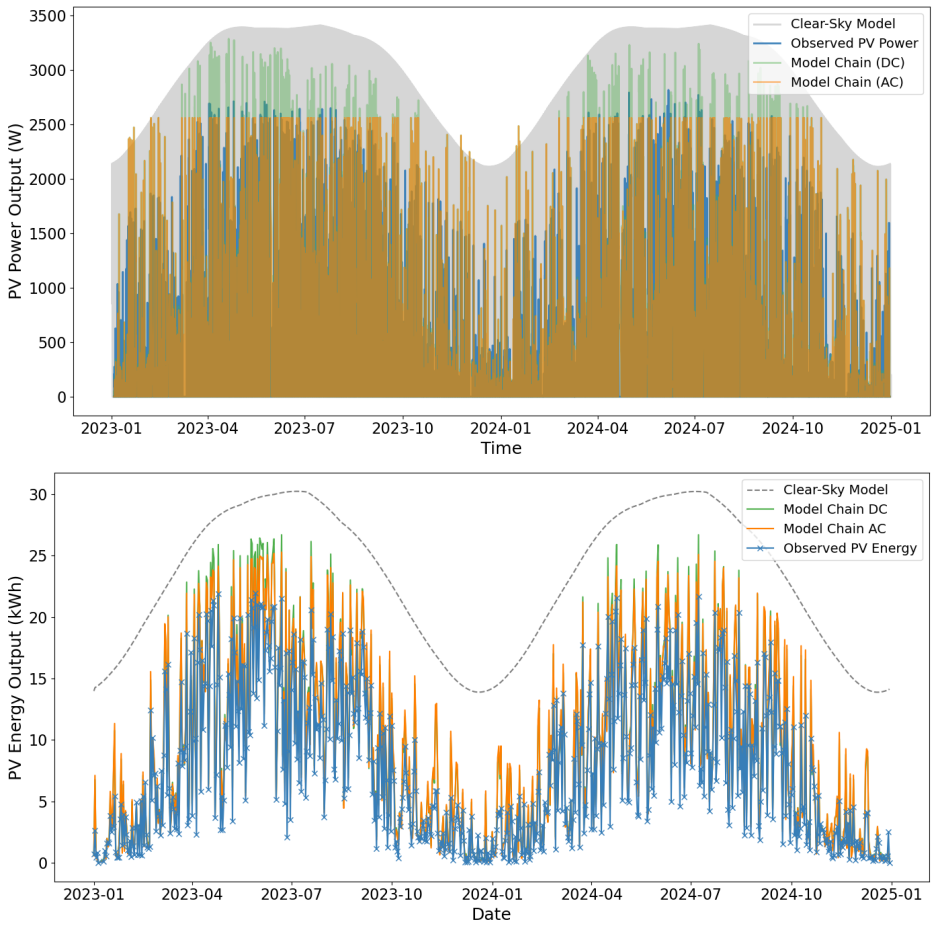

Figure A.1: UK Site 1: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. The nominal capacity is 3.36 kW while the peak capacity is approximately 2.81 kW.

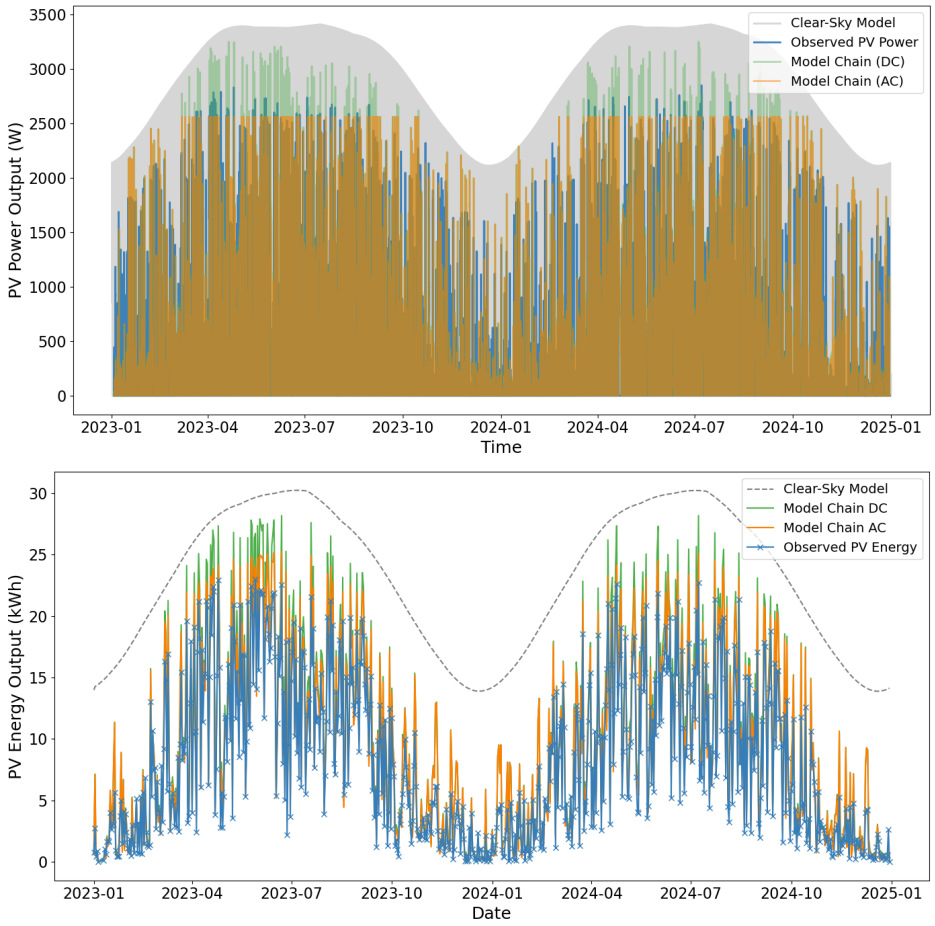

Figure A.2: UK Site 2: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. The nominal capacity is 3.36 kW while the peak capacity is approximately 2.85 kW.

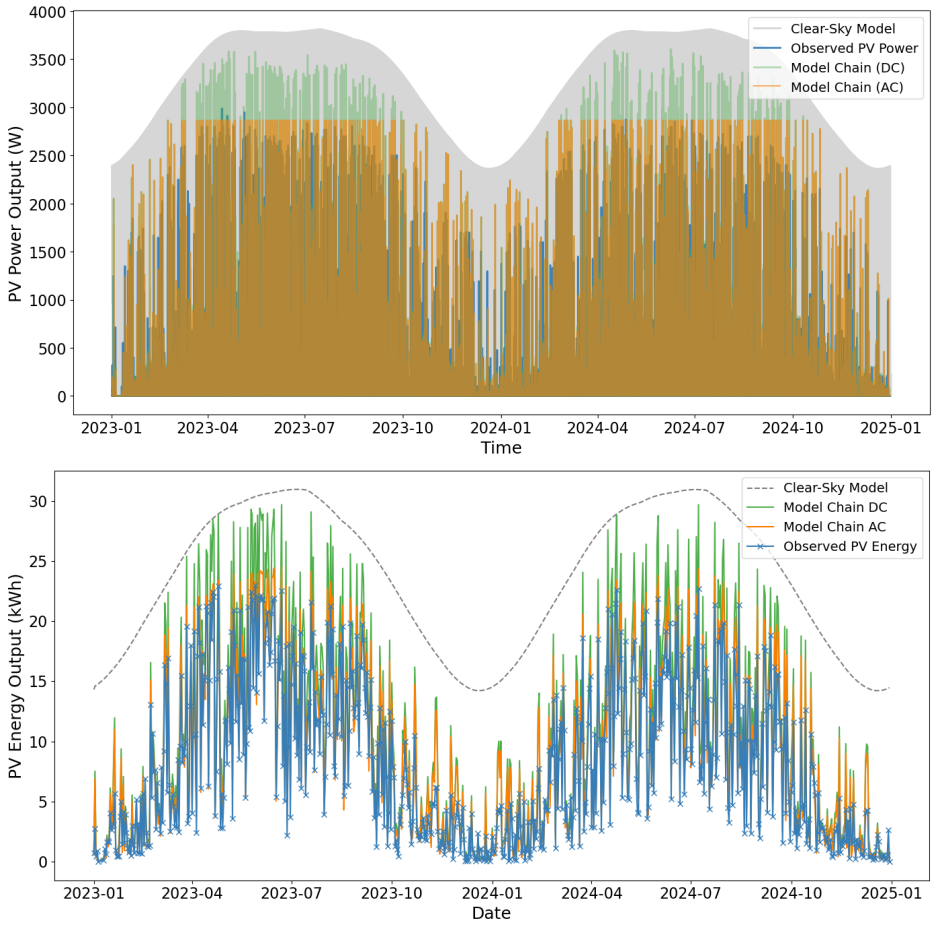

Figure A.3: UK Site 3: (Top) Comparison of observed PV power and energy with a clear-sky model, as well as deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV energy data alongside deterministic model predictions with and without inverter clipping. The nominal capacity is 3.76 kW while the peak capacity is approximately 2.98 kW.

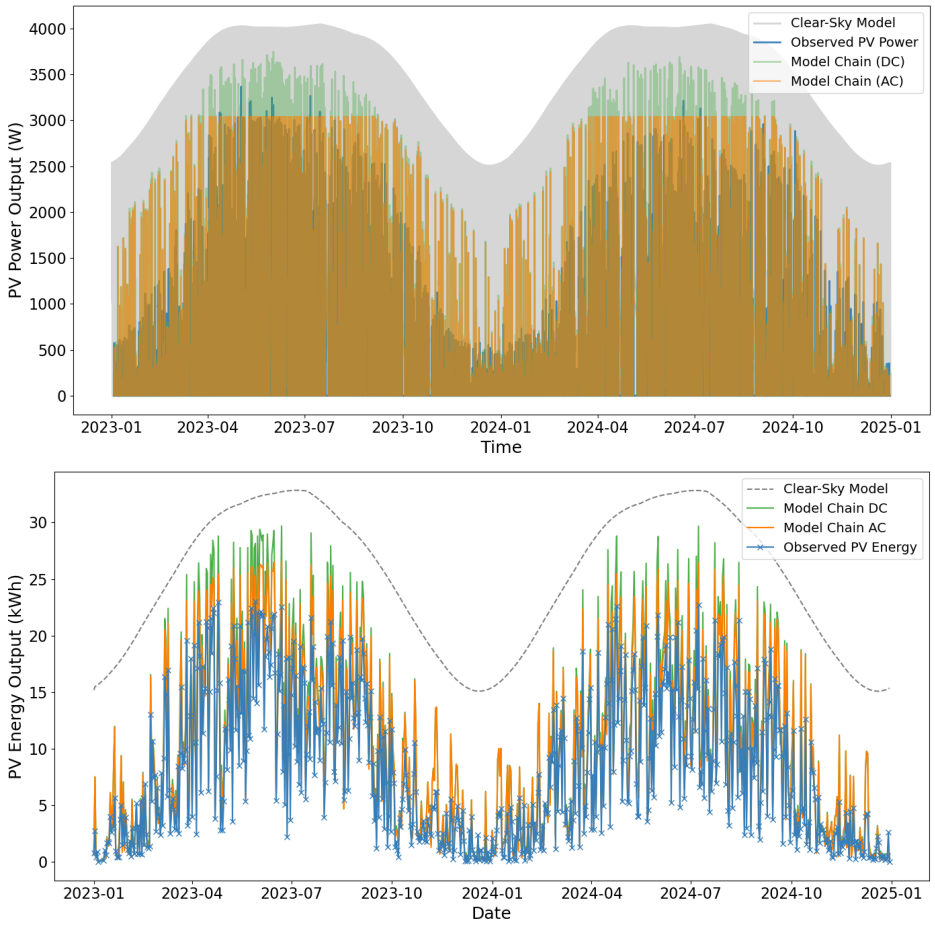

Figure A.4: UK Site 4: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV energy data alongside model predictions with and without inverter clipping. The nominal capacity is 3.99 kW while the peak capacity is approximately 3.36 kW.

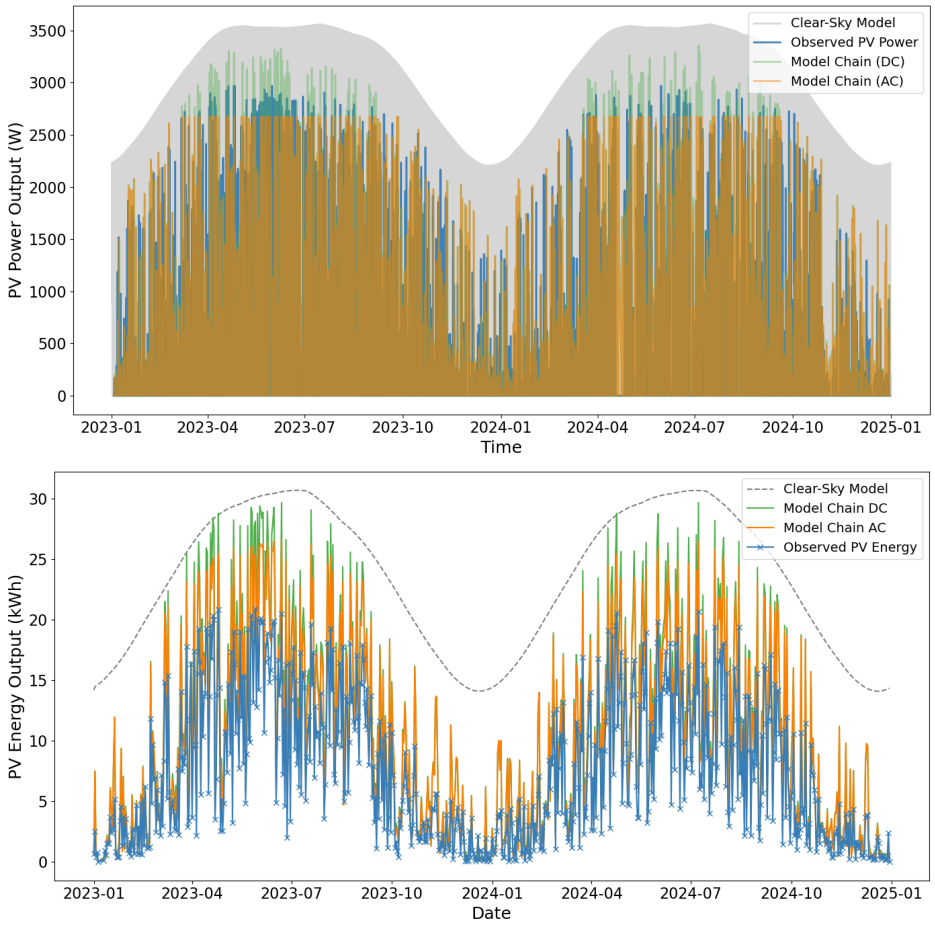

Figure A.5: UK site 5: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. The nominal capacity is 3.51 kW while the peak capacity is approximately 2.97 kW.

A.1.2 HK sites deterministic model results

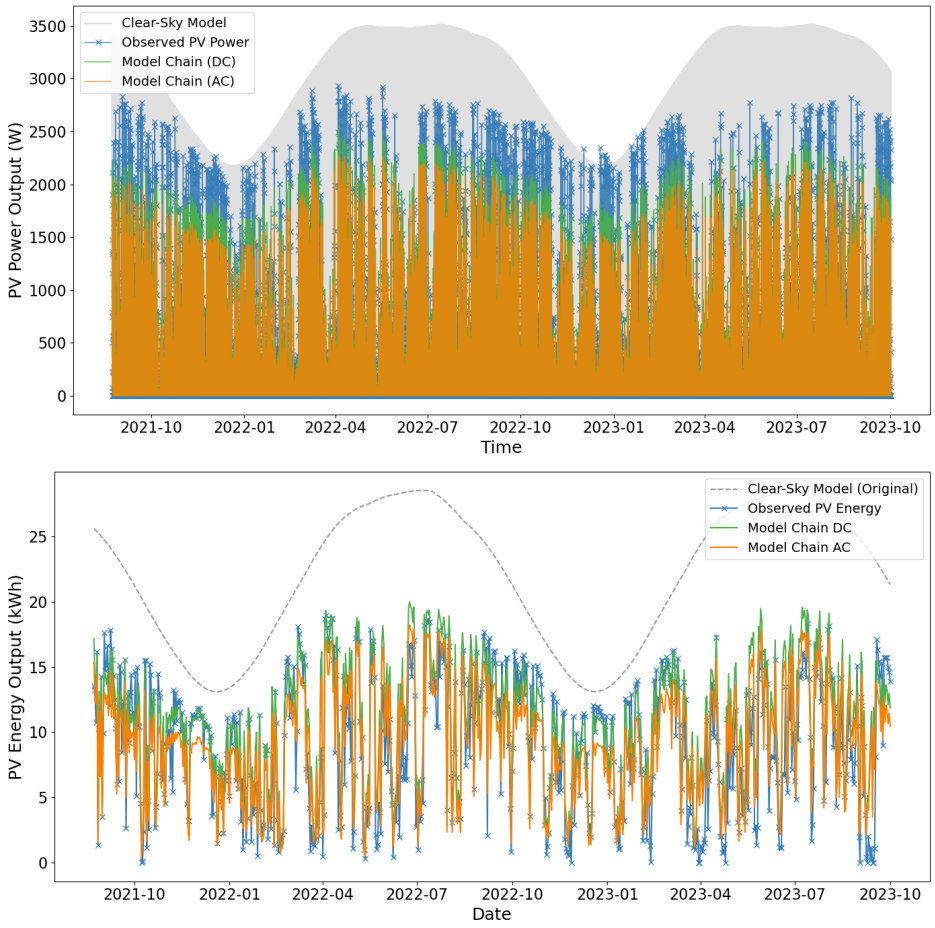

Figure A.6: HK Site B: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data at HK site B alongside deterministic model predictions with and without inverter clipping. Nominal capacity is estimated to be 3.40 kW, while the peak capacity is approximately 2.93 kW.

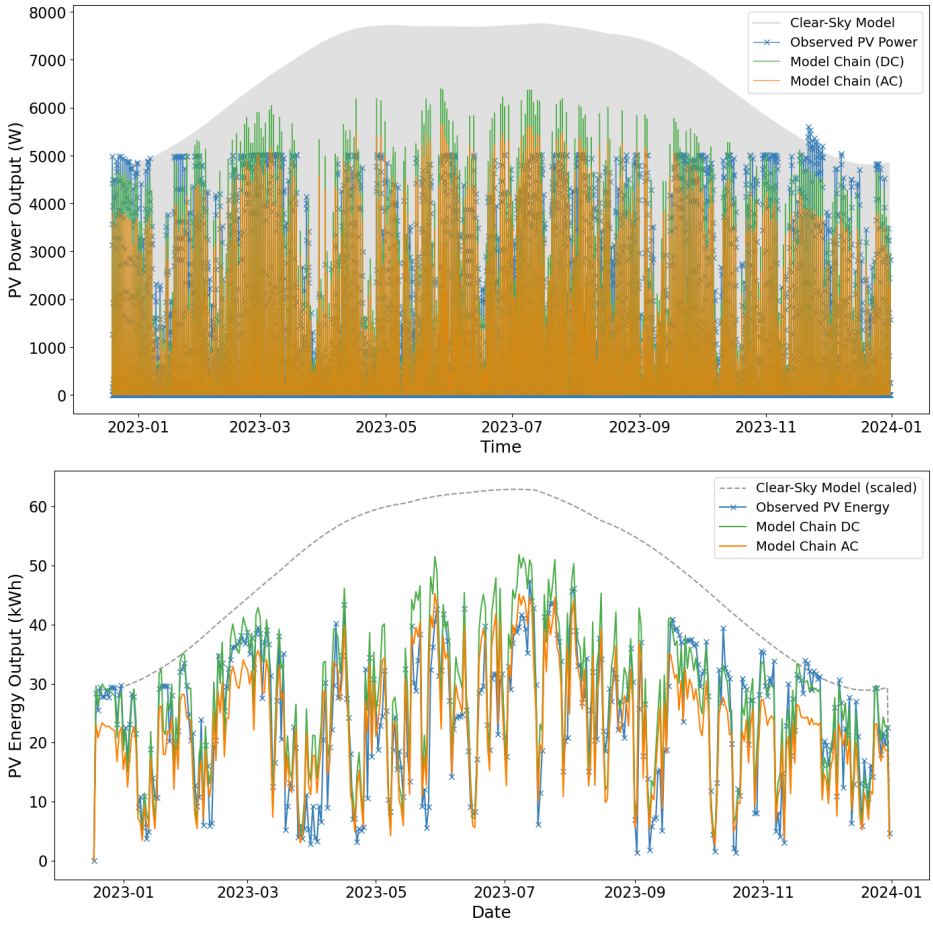

Figure A.7: HK site C: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. Nominal capacity is estimated to be 5.88 kW, while the peak capacity is approximately 5.00 kW.

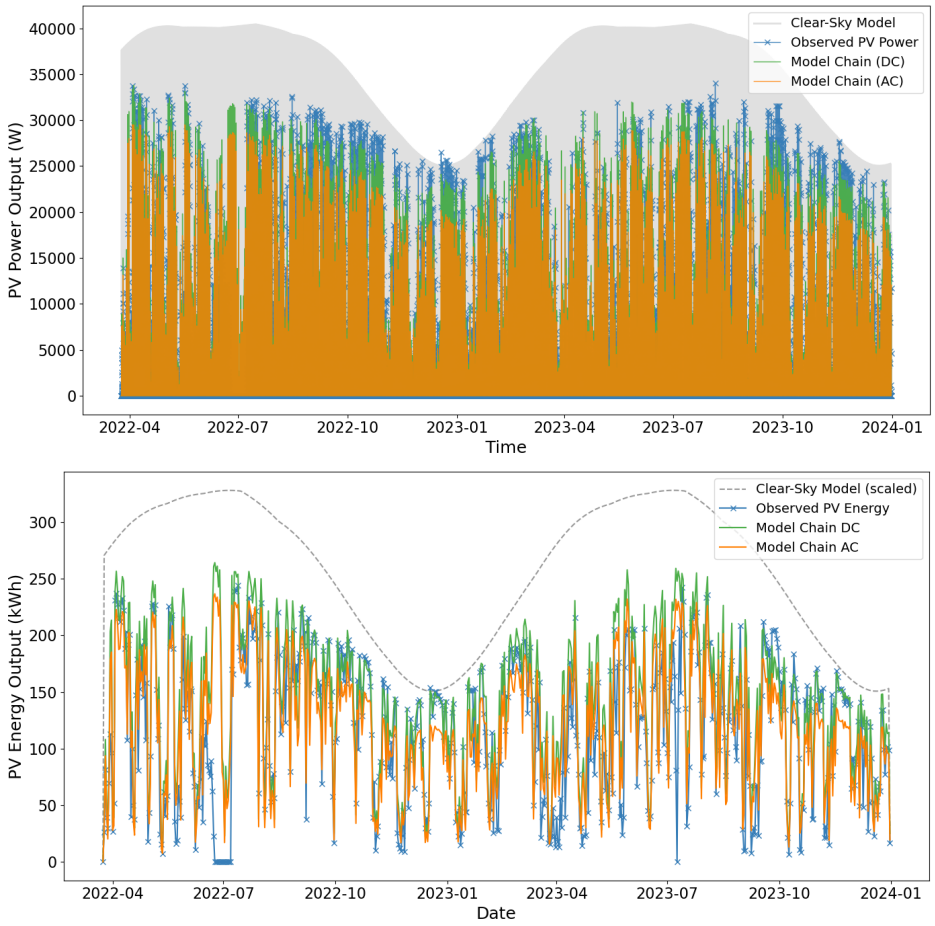

Figure A.8: HK Site D: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. Nominal capacity is estimated to be 34.10 kW, while the peak capacity is approximately 28.99 kW.

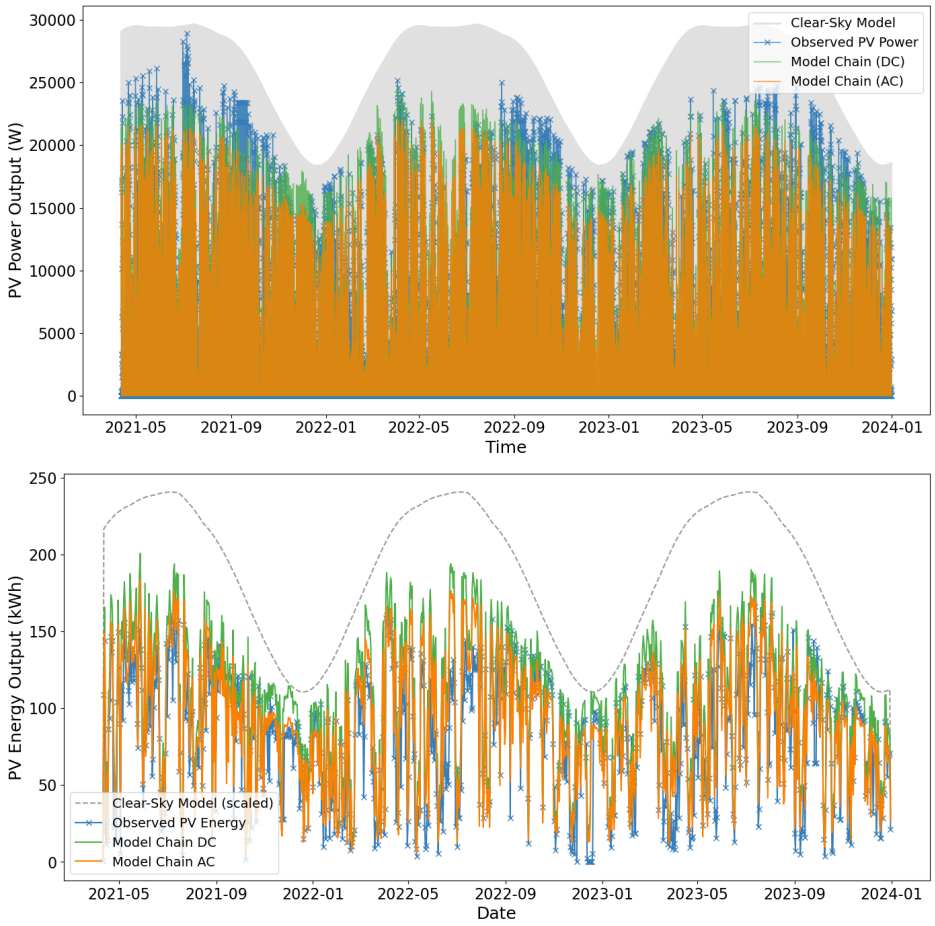

Figure A.9: HK Site E: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. Nominal capacity is estimated to be 34.06 kW, while the peak capacity is approximately 28.95 kW.

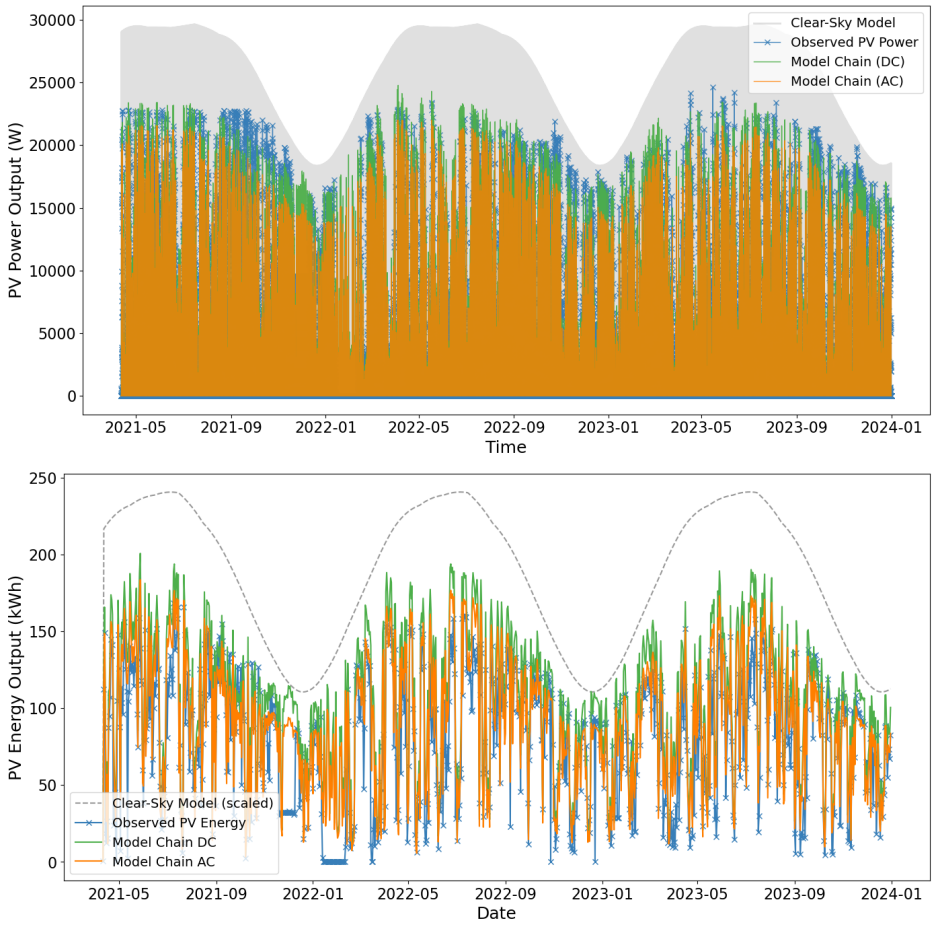

Figure A.10: HK Site F: (Top) Comparison of observed PV power with a clear-sky model, as well as with deterministic physical model predictions with and without inverter clipping. (Bottom) Comparison of observed PV day-ahead cumulative energy data alongside deterministic model predictions with and without inverter clipping. Nominal capacity is estimated to be 29.01 kW, while the peak capacity is approximately 24.66 kW.

A.1.3 UK sites benchmarking results

Table A.1: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the UK site 1. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.45	0.36	36.2	97.59	231.29	6.88	1.11	1.39	24.05
Time series GP (NUTS)	0.46	0.37	371.00	107.87	250.36	7.45	1.33	1.67	24.49
GluonTS	0.46	0.41	4.89	121.33	272.19	8.10	2.09	2.86	38.38
ARIMA	0.46	0.62	19.62	137.28	313.52	9.33	2.94	4.61	47.45
LSTM (concurrent)	0.3890	0.4623	248.00	137.34	300.13	8.93	2.25	3.13	45.90
LSTM (recurrent)	0.39	0.4622	26.00	137.14	300.00	8.93	2.24	3.12	45.83
Transformer	0.49	0.43	2.75	170.76	282.04	8.39	3.31	5.93	47.05
Random forest	0.54	0.71	885.73	151.25	364.27	10.84	2.74	4.51	45.70
XGBoost	0.52	0.69	139.30	156.30	363.57	10.82	2.93	4.54	46.95
Direct Predictors									
Deterministic model (AC)	–	–	–	127.13	278.68	8.29	2.14	2.94	54.20
GluonTS	–	–	–	–	–	–	3.71	4.90	61.06
LSTM (concurrent)	–	–	–	–	–	–	6.14	7.11	195.06
LSTM (recurrent)	–	–	–	–	–	–	5.82	6.91	195.00
Transformer	–	–	–	–	–	–	5.77	6.82	185.97
On-the-market Model									
Quartz (OCF)	–	–	–	163.50	333.56	9.93	5.51	6.50	81.62

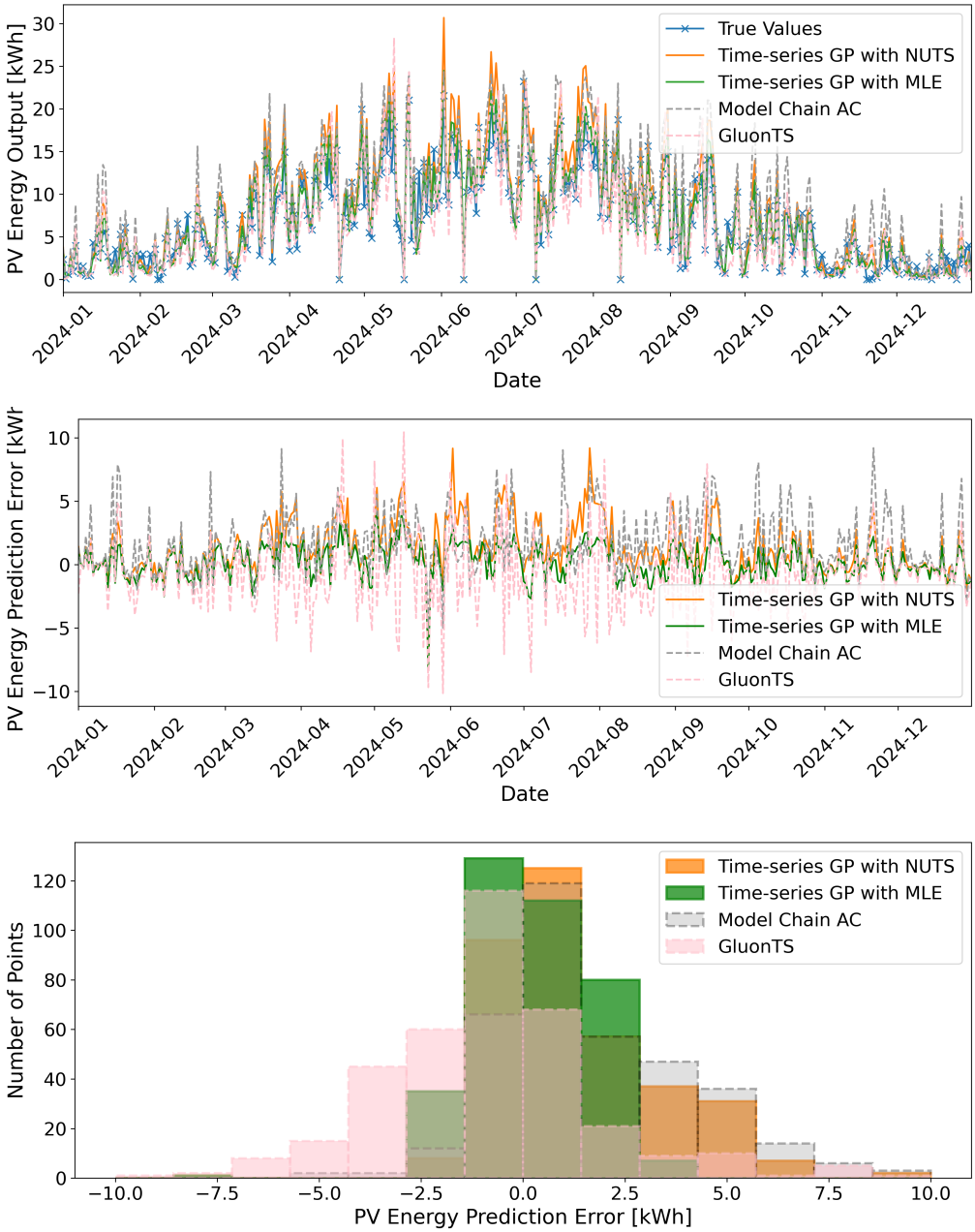

Figure A.11: UK Site 1: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 3.36 kW while the peak capacity is approximately 2.81 kW.

Table A.2: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the UK site 2. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.45	0.34	41.02	117.44	297.36	8.85	1.19	1.39	19.78
Time series GP (NUTS)	0.44	0.35	392.11	123.14	302.73	9.01	1.28	1.49	20.66
GluonTS	0.46	0.37	4.74	137.36	293.62	8.74	2.26	3.18	36.31
ARIMA	0.46	0.60	20.03	166.44	375.06	11.16	2.79	4.18	39.27
LSTM (concurrent)	0.36	0.44	302.73	147.66	317.16	9.44	2.06	2.85	35.21
LSTM (recurrent)	0.36	0.44	318.07	148.35	318.67	9.48	2.08	2.88	35.43
Transformer	0.52	0.46	2.56	174.76	306.53	9.12	3.15	5.128	43.42
Random forest	0.551	0.726	782.58	170.81	414.29	11.02	2.92	4.86	42.55
XGBoost	0.521	0.684	491.64	171.24	410.44	10.92	2.96	4.94	42.17
Direct Predictors									
Deterministic model (AC)	–	–	–	144.02	304.85	9.07	2.14	2.94	44.20
GluonTS	–	–	–	–	–	–	3.96	5.12	51.02
LSTM (concurrent)	–	–	–	–	–	–	5.67	6.84	168.55
LSTM (recurrent)	–	–	–	–	–	–	5.49	6.57	165.27
Transformer	–	–	–	–	–	–	5.33	6.50	160.89
On-the-market Model									
Quartz (OCF)	–	–	–	182.81	373.03	11.10	5.06	6.20	56.61

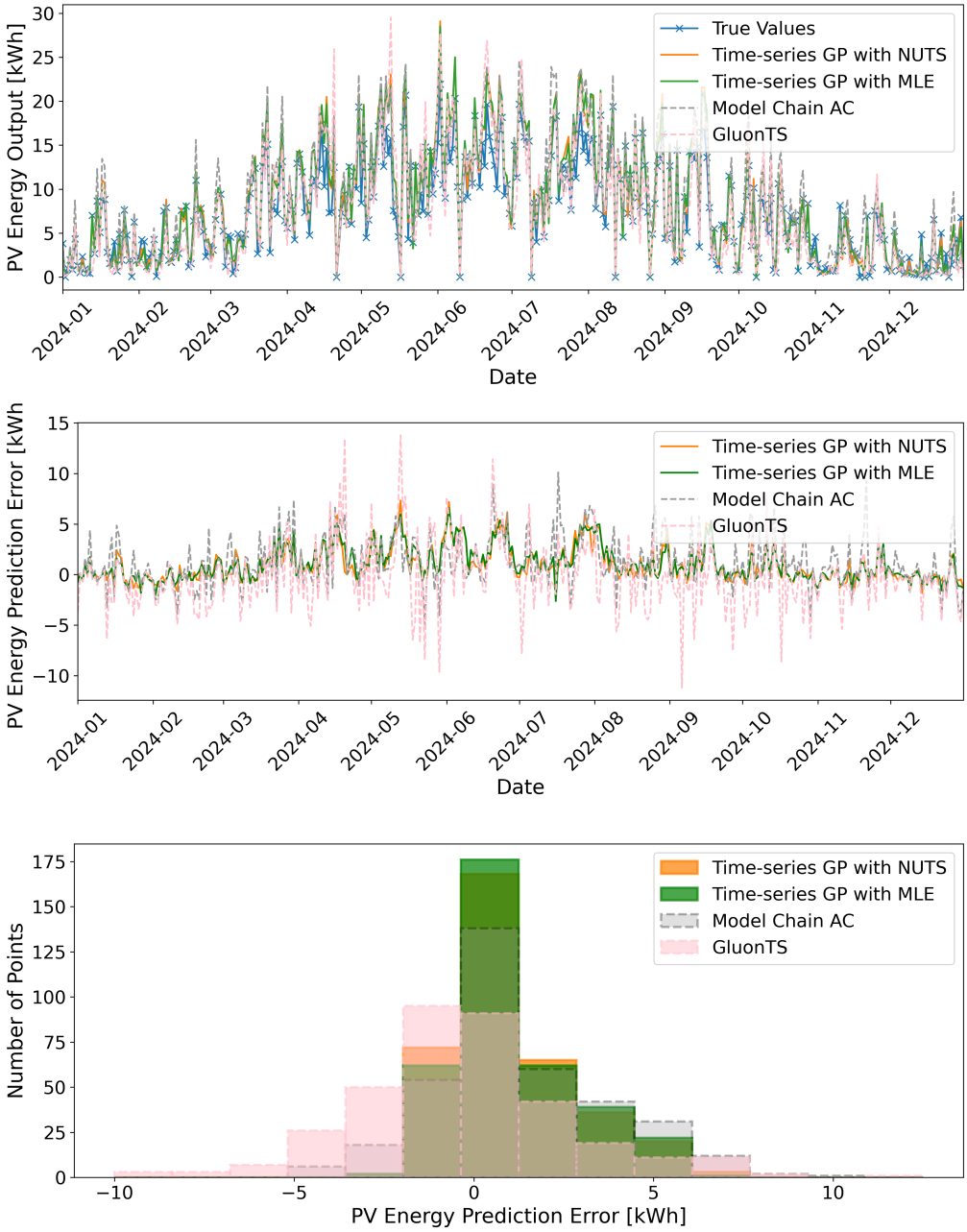

Figure A.12: UK Site 2: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 3.36 kW while the peak capacity is approximately 2.85 kW.

Table A.3: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the UK site 3. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.31	0.16	34.22	112.49	286.82	7.15	1.39	1.83	24.54
Time series GP (NUTS)	0.32	0.17	287.64	112.31	287.52	7.65	1.39	1.76	23.55
GluonTS	0.33	0.17	4.42	116.68	289.05	7.69	1.86	2.64	33.90
ARIMA	0.31	0.41	22.38	125.79	299.55	7.97	1.88	2.63	33.59
LSTM (concurrent)	0.22	0.34	232.10	123.80	289.57	7.70	2.26	3.03	39.55
LSTM (recurrent)	0.21	0.33	228.85	124.09	290.09	7.72	2.27	3.04	39.71
Transformer	0.34	0.17	2.76	128.346	270.33	7.18	2.21	3.29	36.65
Random forest	0.45	0.60	2061.37	143.30	406.86	10.82	2.62	5.28	38.94
XGBoost	0.386	0.483	1585.99	148.19	427.11	11.36	2.77	5.62	41.11
Direct Predictors									
Deterministic model (AC)	–	–	–	131.12	301.90	8.03	2.53	3.36	52.26
GluonTS	–	–	–	–	–	–	4.45	5.96	63.99
LSTM (concurrent)	–	–	–	–	–	–	6.85	8.66	190.12
LSTM (recurrent)	–	–	–	–	–	–	6.66	8.40	185.53
Transformer	–	–	–	–	–	–	6.53	8.29	181.31
On-the-market Model									
Quartz (OCF)	–	–	–	197.32	390.69	10.39	6.22	7.93	77

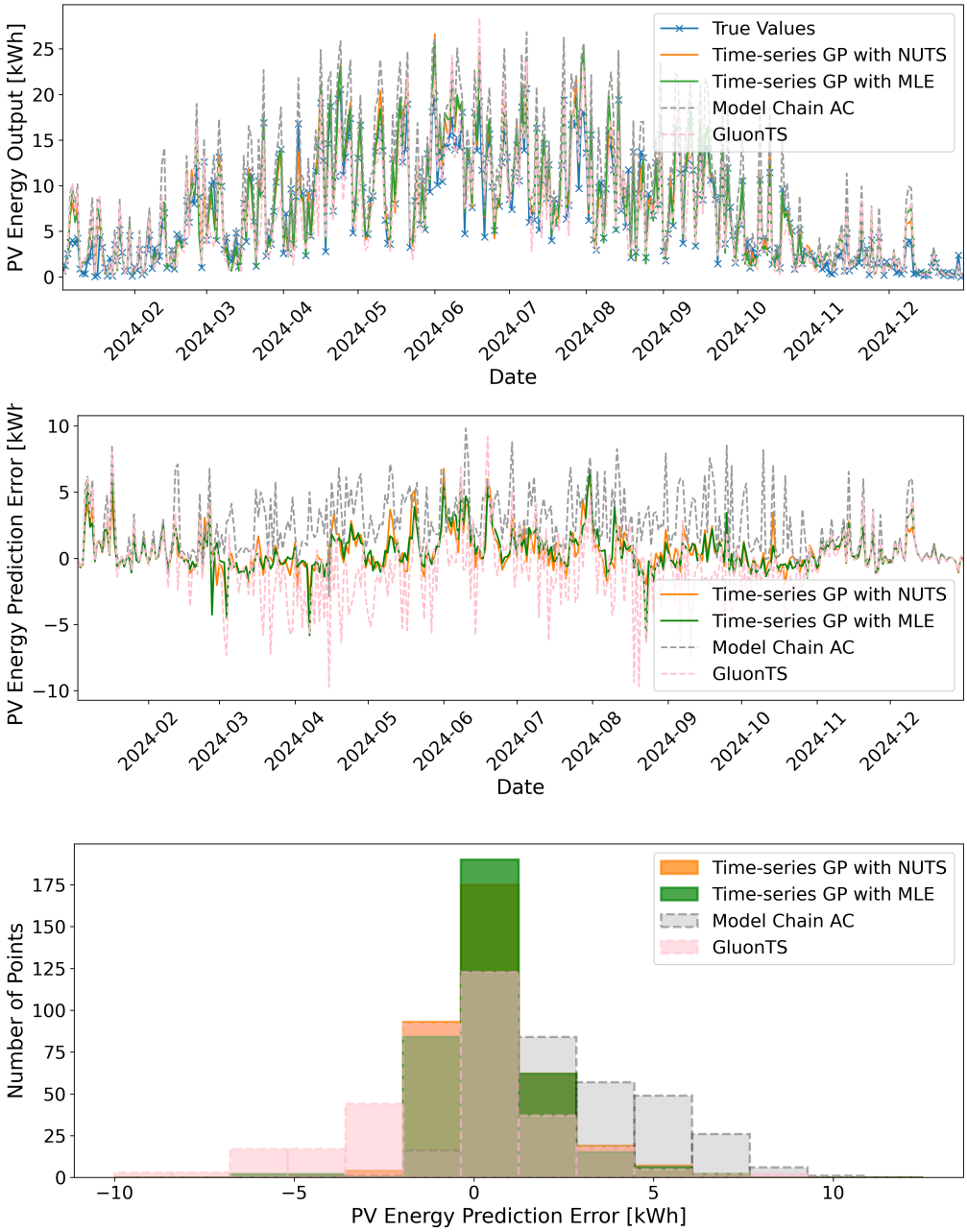

Figure A.13: UK Site 3: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 3.76 kW while the peak capacity is approximately 2.98 kW.

Table A.4: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the UK site 4. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.42	0.31	36.27	131.24	324.95	8.14	1.85	2.49	24.00
Time series GP (NUTS)	0.42	0.33	351.44	133.81	322.34	8.08	1.74	2.30	23.97
GluonTS	0.39	0.29	4.72	136.06	288.66	7.23	2.39	3.41	35.00
ARIMA	0.43	0.57	19.24	165.11	388.11	9.73	3.17	5.00	38.18
LSTM (concurrent)	0.33	0.42	297.88	210.70	461.12	11.56	3.02	4.23	42.98
LSTM (recurrent)	0.33	0.42	286.93	210.14	444.10	11.13	2.83	3.83	44.83
Transformer	0.49	0.39	2.80	182.50	445.50	11.17	3.51	5.85	43.90
Random forest	0.46	0.61	355.33	170.12	440.40	11.14	3.34	5.87	38.94
XGBoost	0.46	0.61	482.65	161.37	411.76	10.32	3.11	5.42	36.76
Direct Predictors									
Deterministic model (AC)	–	–	–	147.66	327.24	8.20	3.36	4.85	43.68
GluonTS	–	–	–	–	–	–	4.14	5.56	69.24
LSTM (concurrent)	–	–	–	–	–	–	6.77	8.00	82.48
LSTM (recurrent)	–	–	–	–	–	–	6.57	8.22	79.32
Transformer	–	–	–	–	–	–	6.46	7.61	75.20
On-the-market Model									
Quartz (OCF)	–	–	–	169.75	349.75	8.77	6.12	7.31	70.50

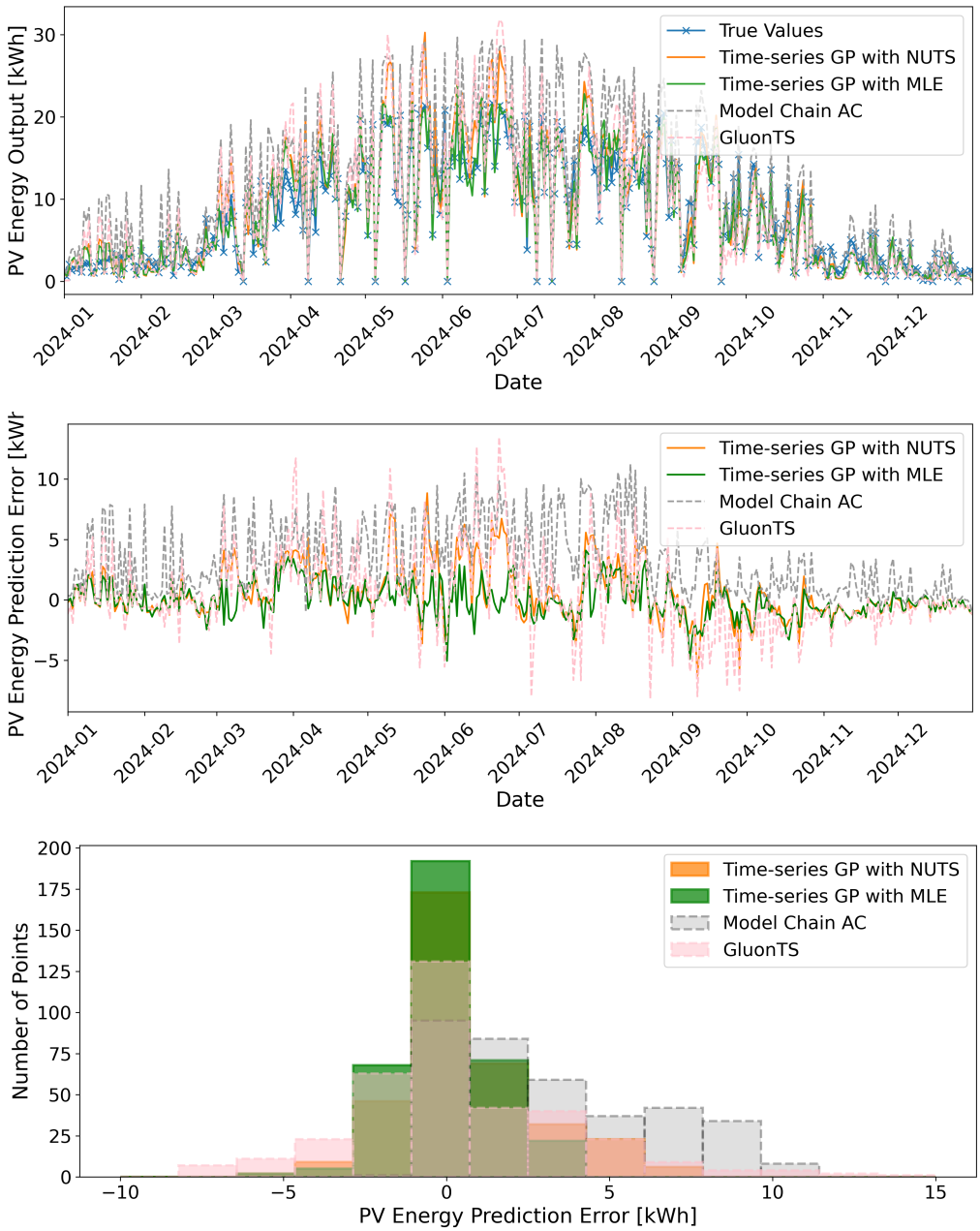

Figure A.14: UK Site 4: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 3.99 kW while the peak capacity is approximately 3.36 kW.

Table A.5: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the UK site 5. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.45	0.40	44.31	165.86	375.58	10.70	2.61	3.76	121.23
Time series GP (NUTS)	0.46	0.40	373.28	170.67	387.03	11.02	3.03	4.31	126.16
GluonTS	0.47	0.42	4.87	162.62	378.76	10.79	3.21	5.22	132.73
ARIMA	0.48	0.65	18.35	180.85	412.57	11.75	3.32	5.79	176.07
LSTM (concurrent)	0.3980	0.4286	276.698	202.42	449.12	12.80	3.55	5.56	194.64
LSTM (recurrent)	0.3981	0.4284	280.803	202.58	449.31	12.80	3.56	5.57	194.78
Transformer	0.527	0.523	2.91	221.81	430.79	12.27	4.39	6.77	169.69
Random forest	0.576	0.80	397.36	224.82	528.58	15.06	4.35	6.93	180.96
XGBoost	0.51	0.70	494.51	232.38	549.34	15.65	4.52	7.25	188.47
Direct Predictors									
Deterministic model (AC)	–	–	–	180.02	401.40	11.43	3.09	4.87	167.15
GluonTS	–	–	–	–	–	–	5.78	7.72	178.39
LSTM (concurrent)	–	–	–	–	–	–	6.93	8.80	195.53
LSTM (recurrent)	–	–	–	–	–	–	6.75	8.59	190.22
Transformer	–	–	–	–	–	–	6.58	8.38	183.62
On-the-market Model									
Quartz (OCF)	–	–	–	210.16	420.32	11.98	6.22	7.93	177

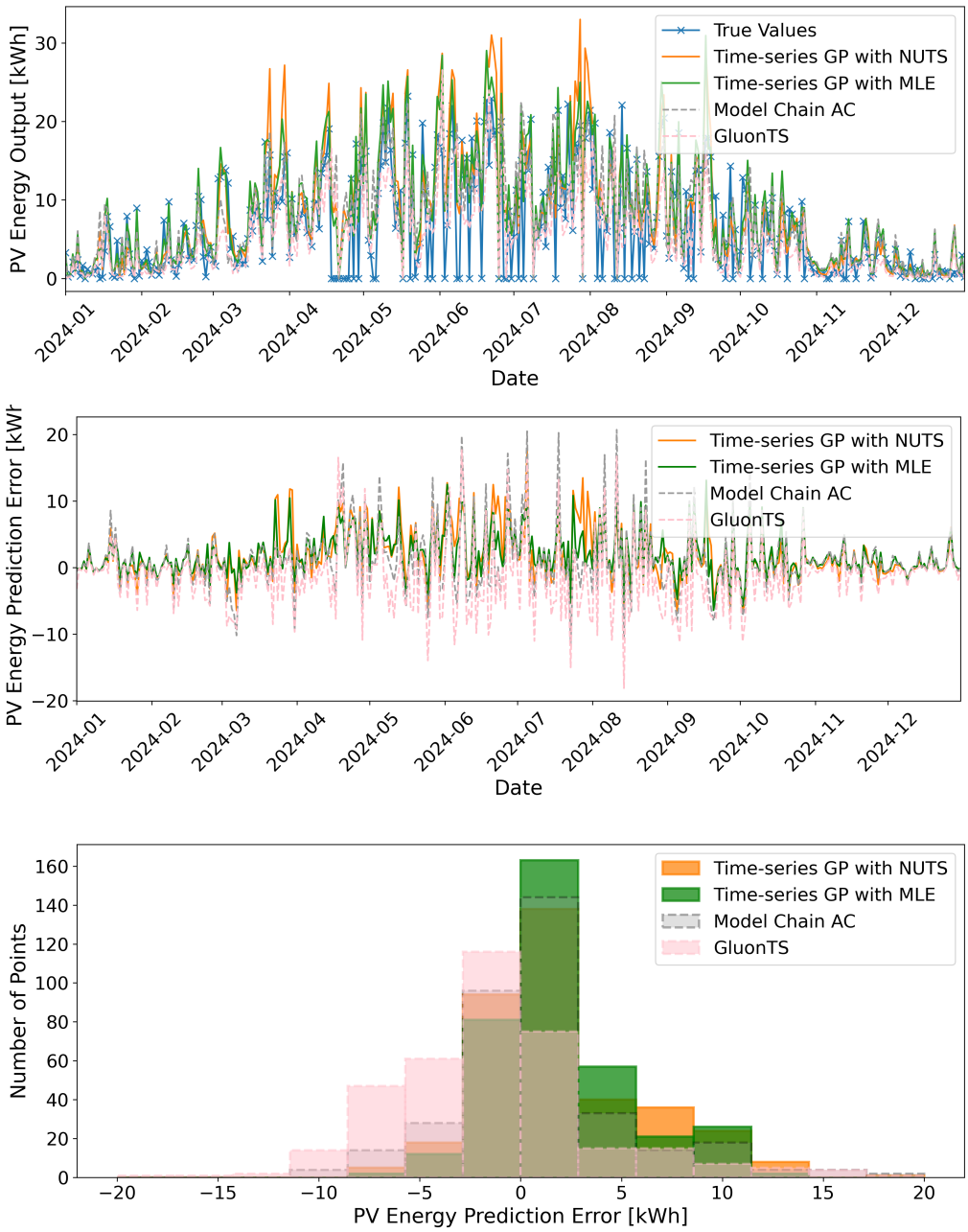

Figure A.15: UK Site 5: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 3.51 kW while the peak capacity is approximately 2.97 kW.

A.1.4 HK sites benchmarking results

Table A.6: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the HK site B. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.38	0.57	23.70	134.97	260.42	7.66	3.57	4.46	47.04
Time series GP (NUTS)	0.39	0.58	280.30	134.01	265.30	7.80	2.68	3.73	47.70
GluonTS	0.36	0.56	4.93	131.39	276.89	8.14	3.38	5.16	53.43
ARIMA	0.39	0.58	20.18	211.77	465.41	13.69	3.80	5.37	60.04
LSTM (concurrent)	0.43	0.70	148.96	538.19	742.05	21.83	3.85	5.48	60.03
LSTM (recurrent)	0.58	4.06	141.32	549.17	763.12	22.44	3.38	5.57	60.06
Transformer	0.49	0.69	2.38	465.53	700.53	20.60	4.45	5.86	66.42
Random Forest	0.41	0.60	219.09	949.21	1579.50	46.46	7.60	13.18	94.71
XGBoost	0.42	0.63	537.87	894.18	1468.89	43.20	6.60	11.50	83.26
Direct Predictors									
Deterministic model (AC)	–	–	–	138.42	273.58	8.05	4.85	5.73	53.51
GluonTS	–	–	–	–	–	–	5.01	6.70	73.22
LSTM (concurrent)	–	–	–	–	–	–	5.44	6.32	66.20
LSTM (recurrent)	–	–	–	–	–	–	5.27	6.19	62.18
Transformer	–	–	–	–	–	–	5.13	5.99	57.34
On-the-market Model									
Quartz (OCF)	–	–	–	642.32	846.58	24.89	3.38	3.97	74.59

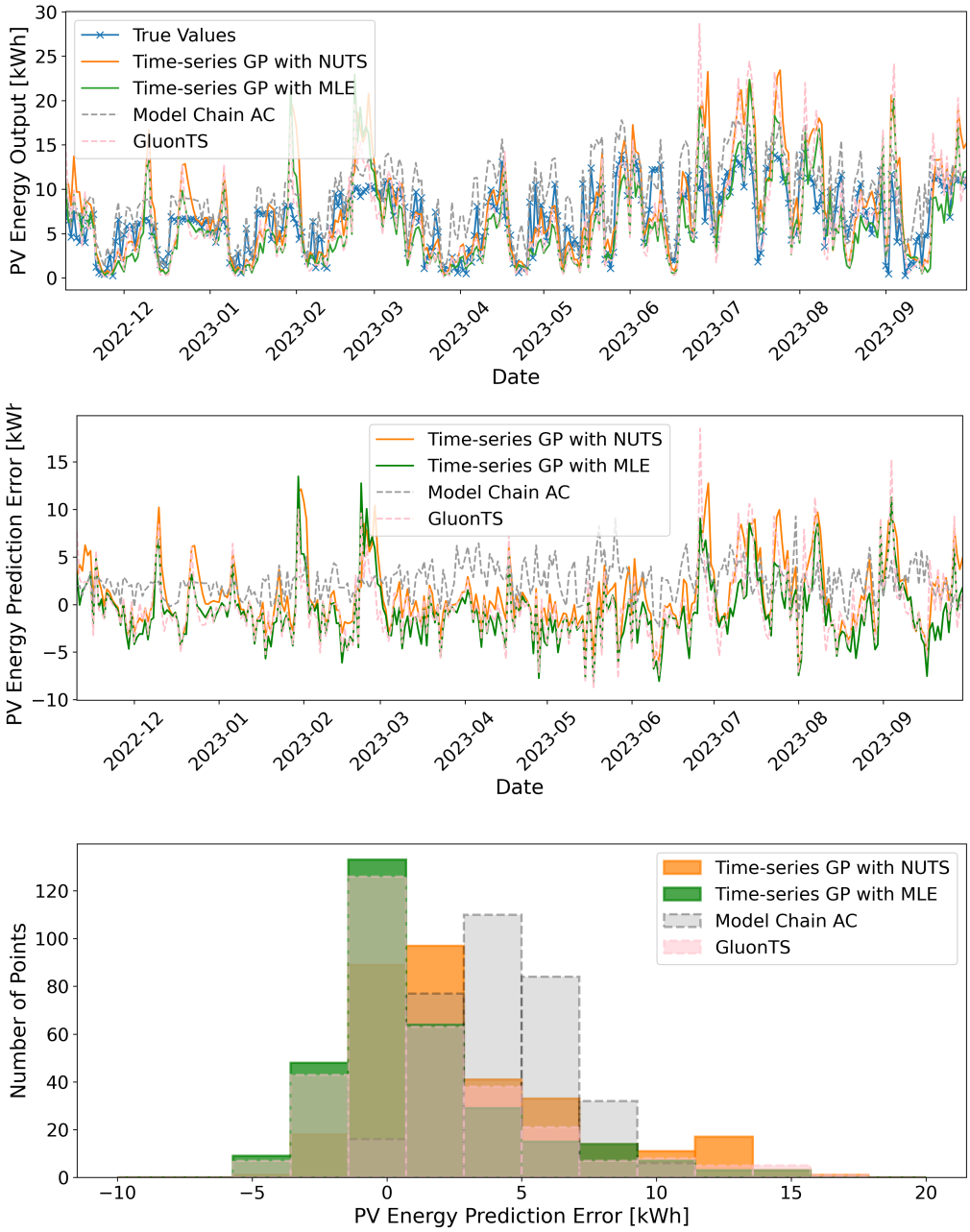

Figure A.16: HK Site B: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 34.10 kW while the peak capacity is approximately 28.99 kW.

Table A.7: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the HK site C. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.61	0.99	39.21	309.48	617.80	10.51	5.28	6.21	27.21
Time series GP (NUTS)	0.61	0.99	390.04	289.15	569.99	9.69	4.56	5.38	24.91
GluonTS	0.65	1.06	4.27	311.46	626.60	10.66	5.24	5.87	26.29
ARIMA	0.666	1.03	6.40	359.95	923.68	15.71	6.73	8.83	52.64
LSTM (concurrent)	1.17	0.82	209.96	1898.96	2559.75	43.53	5.69	7.43	62.87
LSTM (recurrent)	1.37	0.86	128.30	1686.65	2287.58	38.90	7.66	9.01	63.92
Transformer	0.77	1.08	2.49	2260.90	2401.51	40.84	7.51	10.48	66.92
Random forest	0.74	1.11	426.31	2470.65	3893.24	66.21	18.09	32.08	78.32
XGBoost	0.88	1.37	207.67	2331.74	3617.41	61.52	15.62	27.94	68.27
Direct Predictors									
Deterministic model (AC)	–	–	–	286.48	551.80	9.38	4.99	5.88	28.72
GluonTS	–	–	–	–	–	–	6.72	9.25	59.15
LSTM (concurrent)	–	–	–	–	–	–	18.04	21.00	63.87
LSTM (recurrent)	–	–	–	–	–	–	17.65	20.79	61.95
Transformer	–	–	–	–	–	–	17.38	20.33	58.90
On-the-market Model									
Quartz (OCF)	–	–	–	1398.24	2018.05	34.32	16.97	19.42	56.50

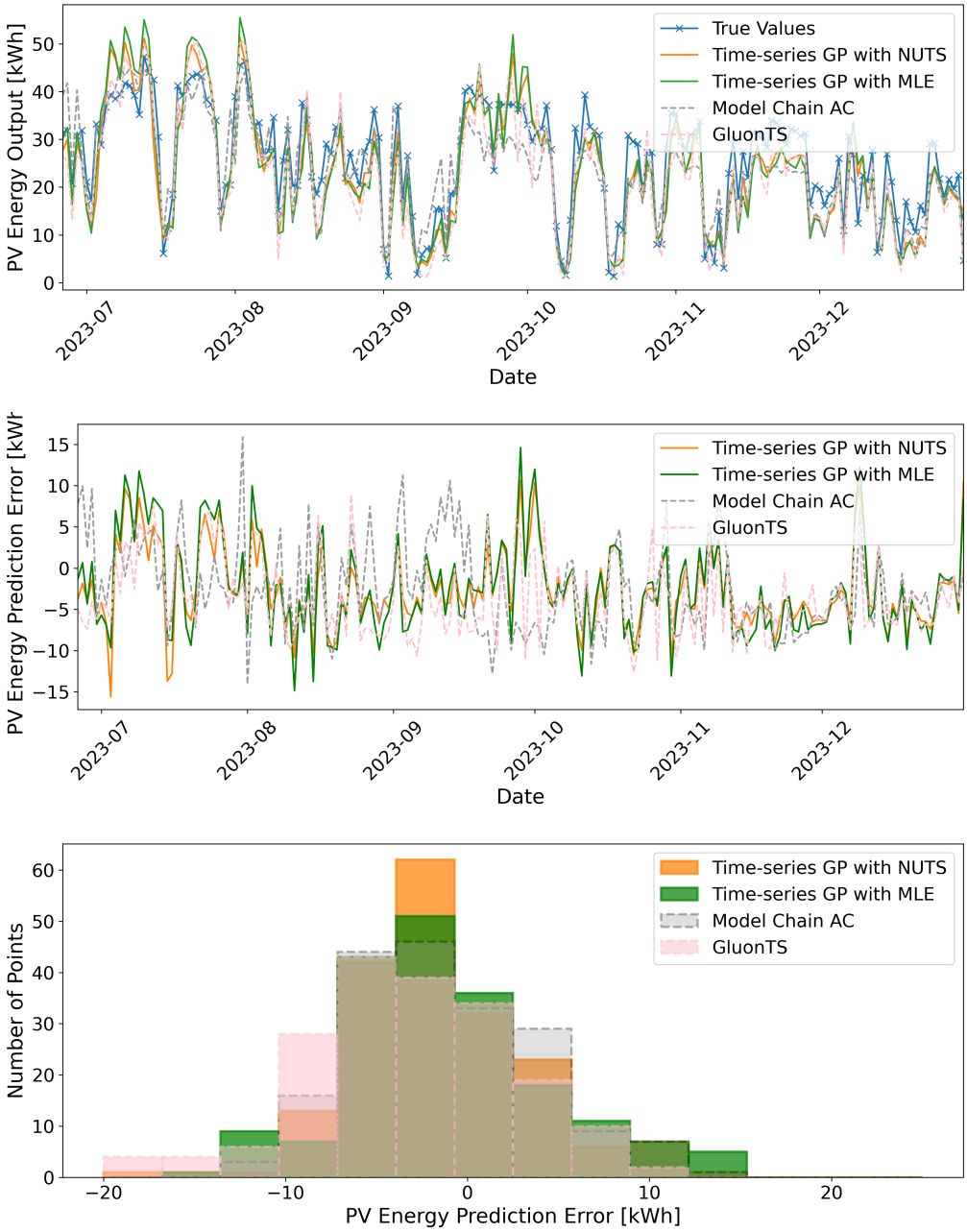

Figure A.17: HK Site C: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 5.88 kW while the peak capacity is approximately 5.00 kW.

Table A.8: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the HK site D. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.67	0.87	36.20	1421.64	3125.98	9.16	16.31	28.58	310.59
Time series GP (NUTS)	0.67	0.89	356.90	1366.20	2748.38	8.06	15.86	20.11	317.08
GluonTS	0.65	0.87	4.54	1429.27	3204.12	9.40	35.04	43.15	335.52
ARIMA	0.68	0.90	14.95	2014.80	4728.73	13.87	43.40	49.79	383.09
LSTM (concurrent)	0.88	1.19	275.53	5847.09	9425.73	27.64	42.64	48.81	377.64
LSTM (recurrent)	1.49	0.95	128.62	5990.61	9495.84	27.85	40.57	46.51	376.06
Transformer	0.72	0.93	2.56	9887.68	8791.87	25.78	32.54	44.95	347.42
Random forest	0.83	1.12	818.45	10581.12	16323.85	47.87	78.25	121.34	414.61
XGBoost	0.84	1.10	422.28	10014.27	15291.54	44.84	70.39	106.76	445.09
Direct Predictors									
Deterministic model (AC)	–	–	–	1504.40	3268.28	10.64	24.66	34.25	348.63
GluonTS	–	–	–	–	–	–	39.99	50.04	352.82
LSTM (concurrent)	–	–	–	–	–	–	116.69	132.03	425.04
LSTM (recurrent)	–	–	–	–	–	–	114.24	129.35	415.32
Transformer	–	–	–	–	–	–	112.06	127.77	400.83
On-the-market Model									
Quartz (OCF)	–	–	–	5177.68	9285.13	27.23	107.99	122.79	387.34

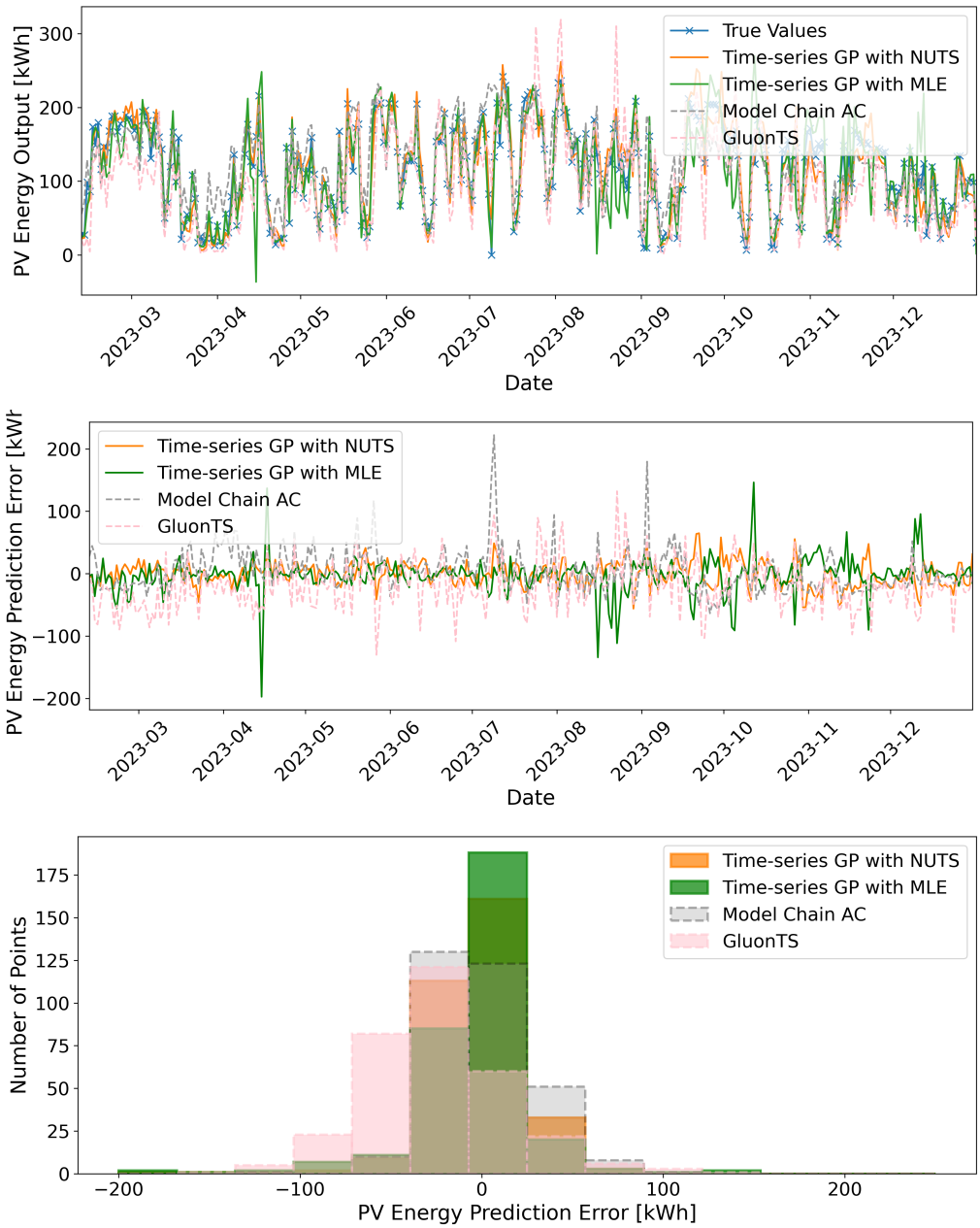

Figure A.18: HK Site D: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 34.10 kW while the peak capacity is approximately 28.99 kW.

Table A.9: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the HK site E. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.58	0.79	28.89	851.48	1899.69	5.57	5.09	7.15	48.36
Time series GP (NUTS)	0.58	0.79	304.23	911.03	1973.52	5.79	7.72	9.20	57.15
GluonTS	0.58	0.81	4.93	1317.91	941.36	2.76	12.43	17.34	74.05
ARIMA	0.60	0.81	29.52	3097.77	5546.16	16.28	16.42	23.13	105.36
LSTM (concurrent)	0.86	0.84	212.15	6592.64	9342.52	27.43	15.41	19.33	87.48
LSTM (recurrent)	0.77	0.85	228.78	7782.83	10904.06	32.01	19.71	23.46	122.15
Transformer	0.67	0.90	3.01	6914.64	10100.37	29.65	17.35	24.09	89.29
Random forest	0.72	0.93	3375.88	7293.64	10866.47	31.90	45.59	64.16	152.60
XGBoost	0.70	0.92	774.71	6912.51	10231.44	30.04	40.40	55.70	136.99
Direct Predictors									
Deterministic model (AC)	–	–	–	1043.96	2128.01	6.25	15.51	21.45	70.95
GluonTS	–	–	–	–	–	–	25.43	33.95	135.25
LSTM (concurrent)	–	–	–	–	–	–	78.02	90.33	96.54
LSTM (recurrent)	–	–	–	–	–	–	76.56	86.59	89.54
Transformer	–	–	–	–	–	–	75.00	86.53	89.92
On-the-market Model									
Quartz (OCF)	–	–	–	3721.76	6628.36	19.46	72.71	83.08	84.00

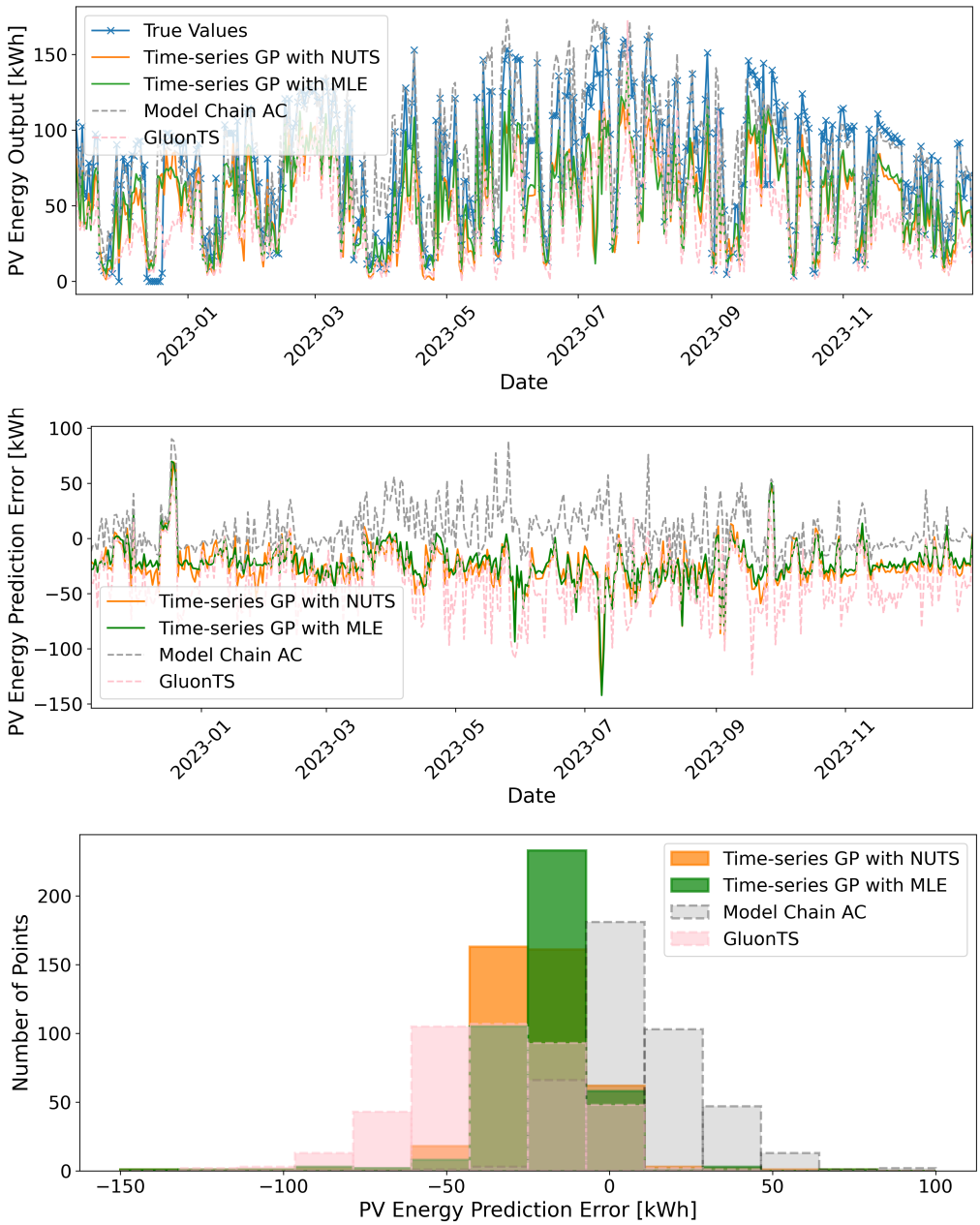

Figure A.19: (HK Site E: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 34.06 kW while the peak capacity is approximately 28.95 kW.

Table A.10: Performance comparison across post-processing models, direct predictors, and the Quartz on-the-market model for the HK site F. Statistical models are highlighted in green, NN models in red, and ensemble-based tree methods in blue.

Model	Adjustment Factor			Power			Day-ahead energy		
	MAE	RMSE	Time [s]	MAE [W]	RMSE [W]	nRMSE [%]	MAE [kWh]	RMSE [kWh]	MAPE [%]
Post-processing Models									
Time series GP (MLE)	0.69	0.98	32.10	849.15	1734.37	5.98	9.20	11.57	183.70
Time series GP (NUTS)	0.65	0.93	385.34	880.22	1835.75	6.32	12.34	17.37	173.65
GluonTS	0.16	0.23	5.21	965.29	2006.77	6.92	13.73	17.91	185.30
ARIMA	0.16	0.22	40.44	5263.37	13734.83	18.14	21.51	28.91	204.67
LSTM (concurrent)	0.06	0.44	147.82	5509.48	7655.76	26.39	14.01	16.95	248.68
LSTM (recurrent)	0.06	0.45	171.66	5444.05	7586.94	21.15	15.07	18.12	250.38
Transformer	0.18	0.24	2.75	5852.60	5562.96	19.18	10.58	15.78	249.30
Random forest	0.19	0.25	26548.81	6133.90	8508.08	29.33	22.66	30.97	263.09
XGBoost	0.18	0.24	792.95	5854.67	8144.47	28.07	23.60	31.21	258.88
Direct Predictors									
Deterministic model (AC)	–	–	–	1101.69	2258.51	7.79	14.35	20.11	223.30
GluonTS	–	–	–	–	–	–	37.11	50.99	385.44
LSTM (concurrent)	–	–	–	–	–	–	45.02	62.10	430.98
LSTM (recurrent)	–	–	–	–	–	–	–	58.04	410.33
Transformer	–	–	–	–	–	–	39.74	54.01	396.67
On-the-market Model									
Quartz (OCF)	–	–	–	3602.73	6316.08	21.77	18.02	21.95	282.23

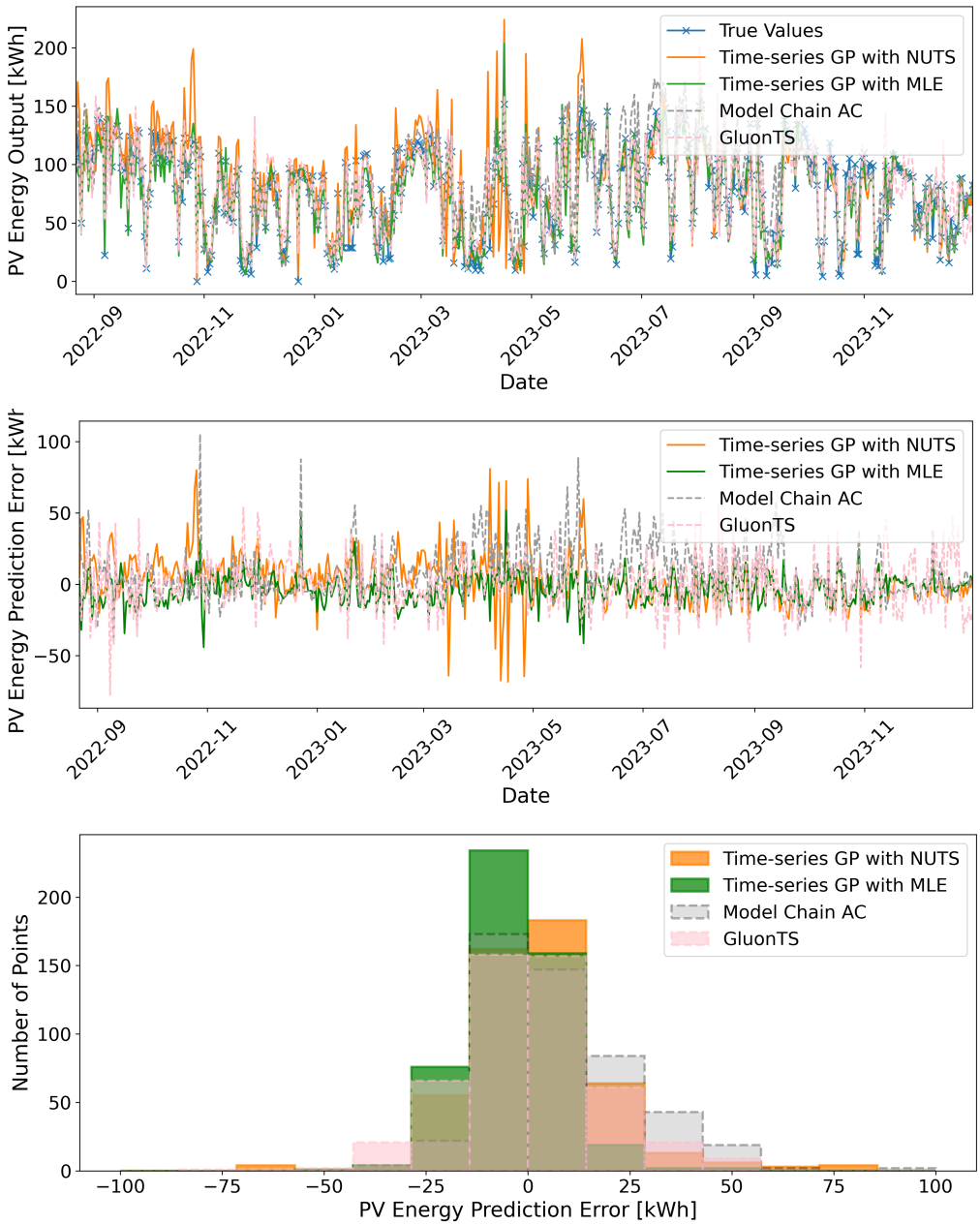

Figure A.20: (HK Site F: (Top) Observed PV day-ahead energy values vs. predictions from GluonTS, GP MLE and GP NUTS models, as well as the AC model. (Middle) Day-ahead PV energy prediction errors for GluonTS, GP MLE, GP NUTS, and the AC model. (Bottom) Distribution of the day-ahead PV energy prediction errors for GluonTS, GP MLE, and GP NUTS models, as well as the AC model. The nominal capacity is 29.01 kW while the peak capacity is approximately 24.66 kW.

A.2 Observed-to-predicted Adjustment Ratio Results

A.2.1 UK site adjustment factor results

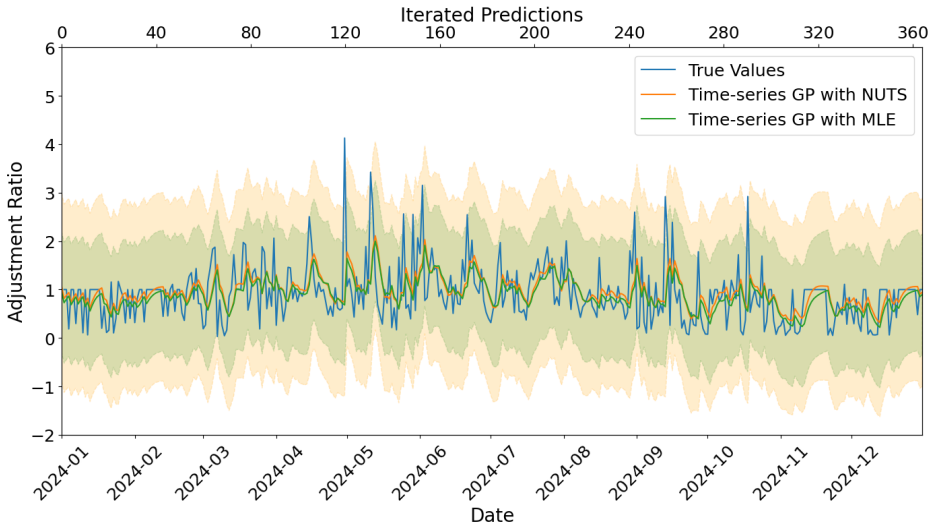

Figure A.21: Time-series GP at the UK Site 1: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

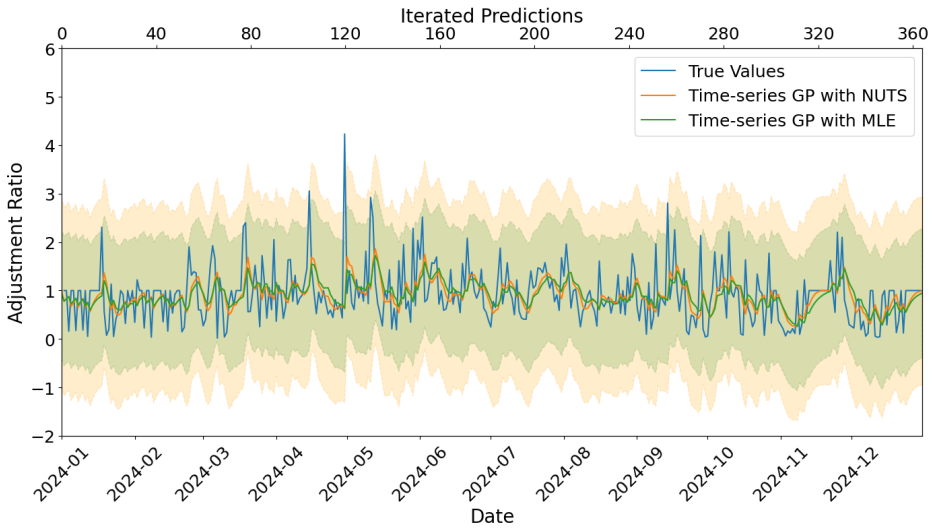

Figure A.22: Time-series GP at the UK Site 2: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

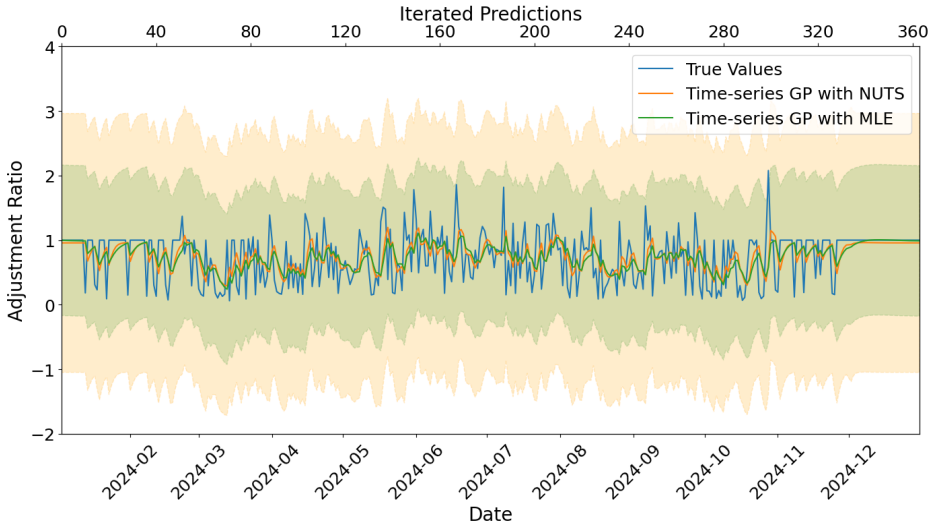

Figure A.23: Time-series GP at the UK Site 3: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

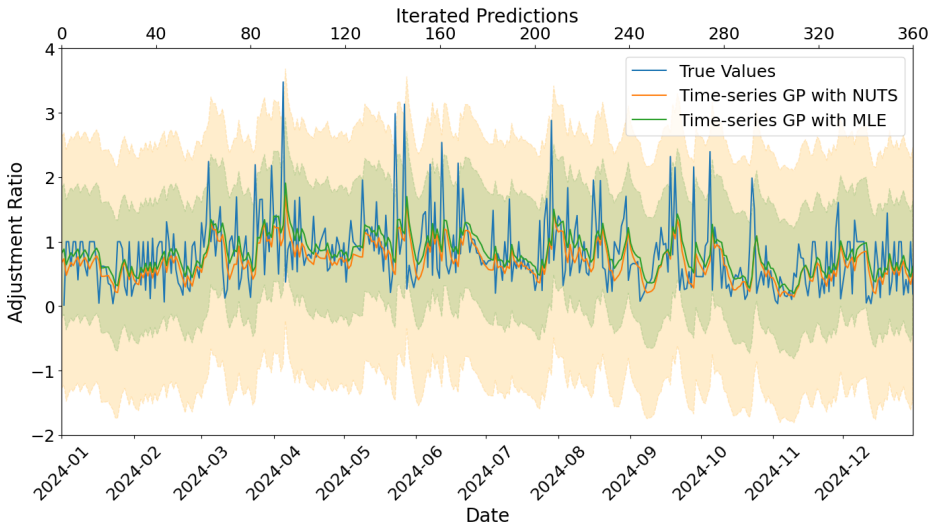

Figure A.24: Time-series GP at the UK Site 4: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

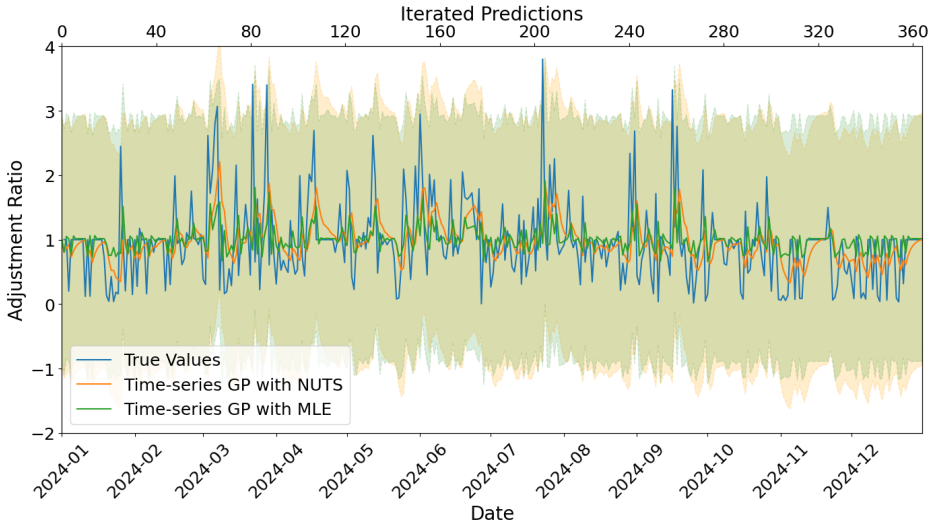

Figure A.25: Time-series GP at the UK Site 5: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

A.2.2 HK site adjustment factor results

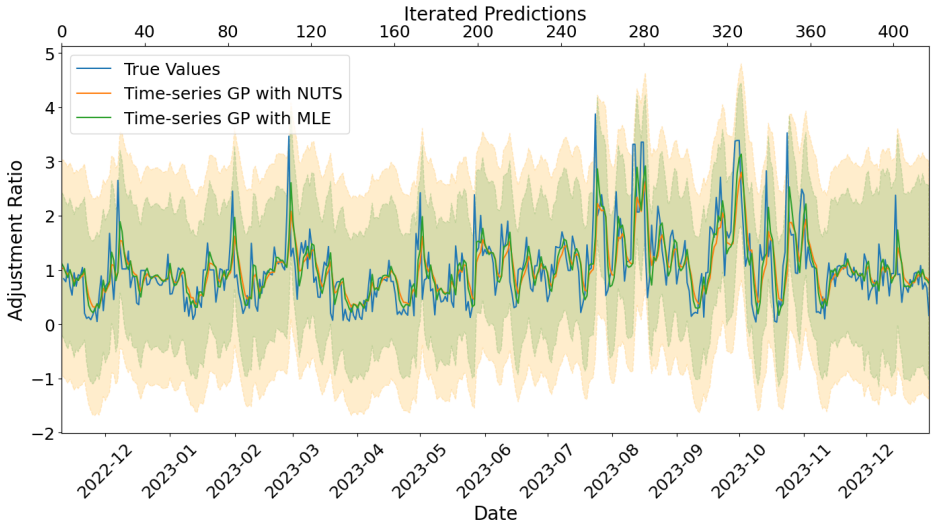

Figure A.26: Time-series GP at the HK Site A: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

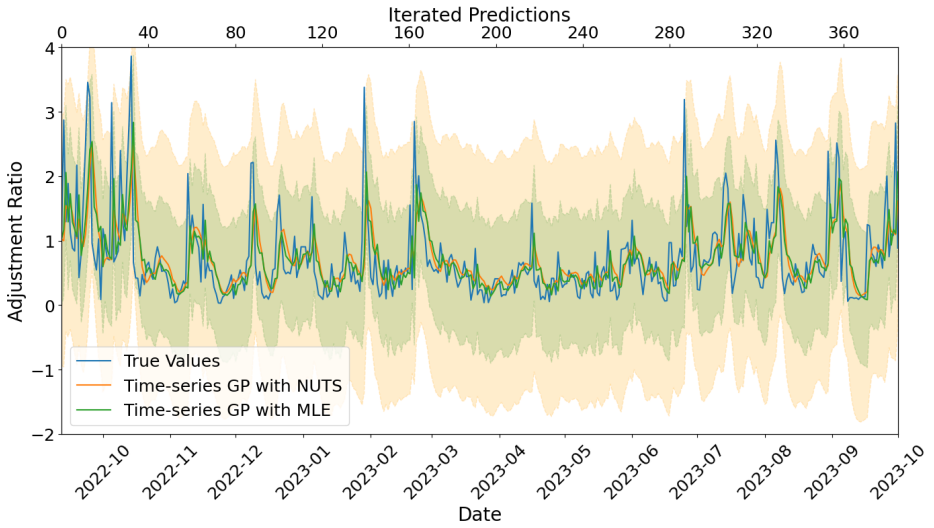

Figure A.27: Time-series GP at the HK Site B: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

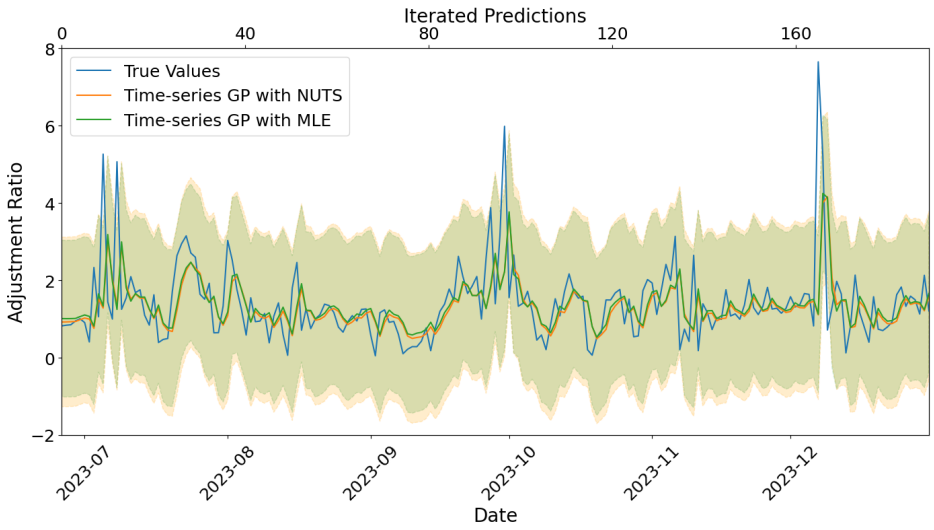

Figure A.28: Time-series GP at the HK Site C: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

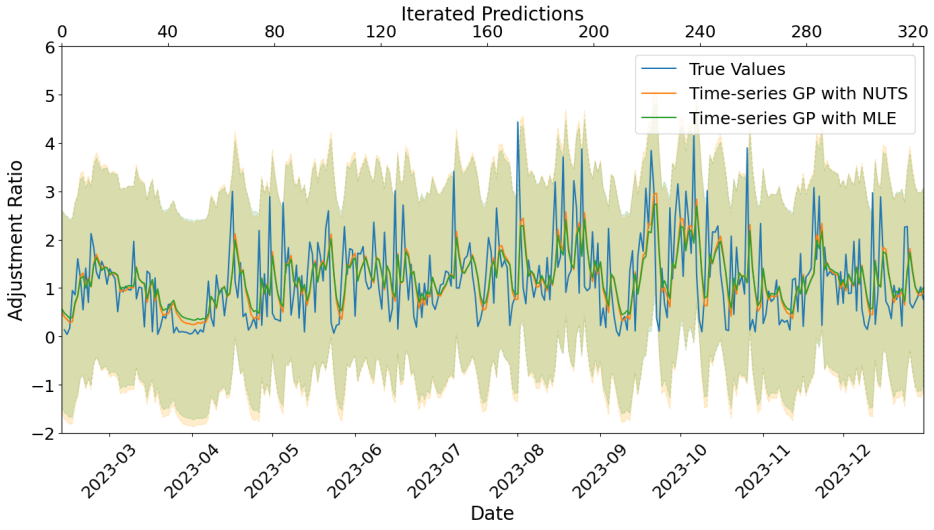

Figure A.29: Time-series GP at the HK Site D: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

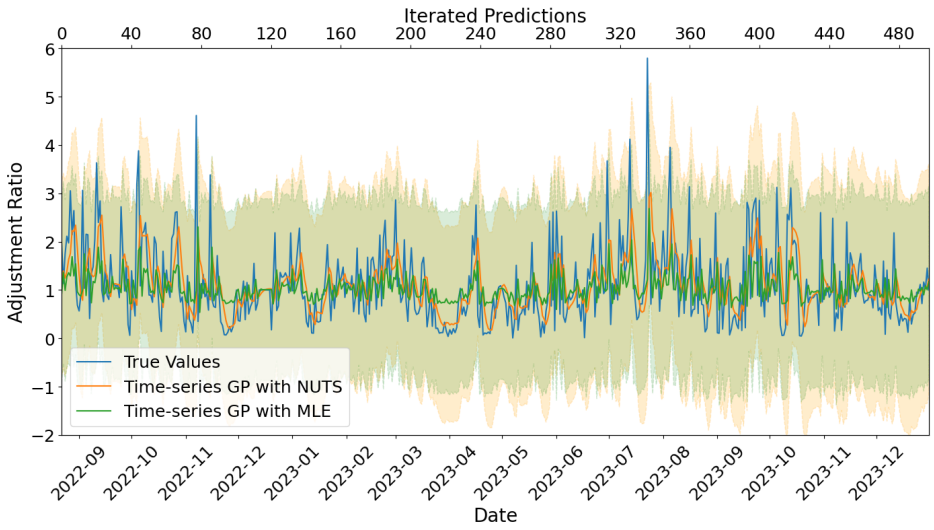

Figure A.30: Time-series GP at the HK Site E: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.

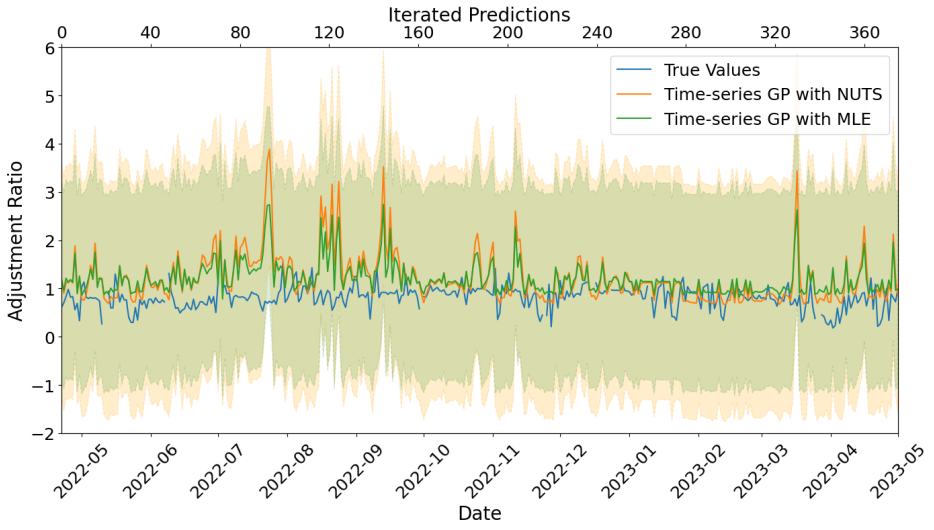

Figure A.31: Time-series GP at the HK Site F: predictive mean (solid) with 95% confidence intervals. MLE vs. NUTS are overlaid to illustrate the effect of hyperparameter marginalisation on uncertainty.