The University Of Sheffield.

Measuring the uncertainty associated with estimating national photovoltaic electricity generation: A Great Britain case study

Owen Thomas Huxley

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Science
Department of Physics and Astronomy

19th July 2021

Acknowledgements

I would like to thank everyone who has helped me directly or indirectly during the 4 years I have spent on this PhD. It has been a long road to completion and I could not have achieved it without the help of many other people.

Firstly, I would like to thank my PhD supervisors, Dr Alastair Buckley and Jamie Taylor. Alastair brought excellent direction and structure to the topics of my research. I have learnt a lot from him about breaking real world problems down into manageable chunks and focussing on the important parts. His door was always open and his guidance has been an enormous help throughout my PhD. In particular, the process of writing this thesis highlighted to me the importance of his overall direction of the research topics in my PhD.

Whilst Alastair provided the big picture, Jamie has always been there to help with the details. Over the last 5 years Jamie has taught me an enormous amount about both the electricity industry and the technical side of coding and data science. I am a far better researcher and data scientist for having worked with Jamie. There were many times when Jamie would stay late in the office working through a new idea for a research project or debugging some python code and for that teaching I am very grateful.

I would like to thank the rest of the Sheffield Solar team past and present. Without them, there would be no PV Live, no relationship with National Grid, no data or database system, no website and none of this research would have been possible. In particular, I would like to thank Aldous, Julian and Dan for putting up with me in the office and always providing help when I asked.

Without the funding of both EPSRC and National Grid ESO none of this research would have been possible. In particular, I am very thankful for the help of Kevin Tilley and Rob Nickerson from National Grid ESO. Kevin has always provided excellent direction on the Phase 3 NIA project and his guidance and help has greatly improved the research output of the whole Sheffield Solar group. Rob Nickerson was my manager and mentor for my 6-month placement inside National Grid ESO. He made me feel exceptionally welcome and his guidance was critical in the successful delivery of the project during my placement.

Finally, I would like to thank all my friends and family and in particular, Mum, Dad, Ciara, and India. Without their support I probably would not have made it to the end of this PhD. I am particularly thankful for the wonderful support India given me throughout through the highs and lows of the last four years. However, I would like to save the last words of thanks for my mum and dad. Without their endless help, guidance, and belief there is no chance that I would be here today. For their role in this work, which cannot accurately be summarised in a few short sentences, I am eternally grateful.

Abstract

Monitoring near real-time national solar PV output is an increasingly important part of operating an electricity system. Monitoring PV output requires knowledge of the output of many PV systems embedded in the distribution network whose generation are not directly visible through existing transmission system metering. In this thesis a review of 27 national solar PV monitoring services which provide national PV output estimates for 20 different countries was performed showing that every service follows the same general approach. First, the PV yield is modelled using a set of data from reference PV systems providing data in real-time. Then the modelled PV yield is scaled by an estimate of the national solar PV capacity to estimate the national PV output. National PV output is then used, along with similar measurements for other embedded technologies such as wind, to train and validate electricity forecasts which ensure efficient electricity market operation.

Using Great Britain as a case study, the total error and uncertainty associated with the estimates from a national PV monitoring service are analysed. There are three main sources of error which contribute to the overall error in the national PV output estimates; the sample bias error, the statistical error in the yield model, and the error in the national capacity estimate. For the GB PV monitoring service, the domestic sample was shown to be unbiased for estimating national PV output. However, at a regional level the domestic sample used in the GB service is biased for estimating commercial/utility PV systems. The statistical error in the yield model was shown to be $\pm 1\%$ providing that a sample size of at least 6000 was used. The error in the GB national capacity estimate was shown to be $\pm 5\%$. I can conclude that, the capacity error, at $\pm 5\%$, dominates the yield calculation error, at $< \pm 1\%$ and leads to an overall error in GB solar PV output estimates of $\pm 5.1\%$. I also conclude that solar PV measurements, and consequently national electricity demand forecasts, are currently limited by the state of national PV capacity registers.

Contents

List of Figures

List of Tables

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously been presented for an award at this, or any other, university.

Research Output

- GIS shape file definitions for the GB Grid Supply Points which are now used widely in the energy industry including in the National Grid ESO Future Energy Scenarios publications and in National Grid ESO's regional PV and wind forecasts. https://data.nationalgrideso.com/system/gis-boundaries-for-gb-grid-supply-points

- Reusable software for deriving GSP level shapefiiles. https://github.com/SheffieldSolar/GSP-Region-Mapping

- Pop-up University Public Outreach Talk, September 2019

- Digital Utilities Europe 2019 Presentation, May 2019

- The University of Sheffield, graduate school showcase poster presentation, March 2020

- Full Length Journal Article: "The uncertainties involved in measuring national solar photovoltaic electricity generation" Submitted to Renewable and Sustainable Energy Reviews on the 23rd of June

Chapter 1

Introduction

Climate change is the multi-year trend of rising global temperatures, driven by anthropogenic greenhouse gas emissions, and the consequent changes to global weather patterns and sea level [1]. There are no scientific bodies of international standing which dispute the fact that climate change is caused by humans [2].

Renewable technology will be at the forefront of any transition to a low carbon economy. Pathways which meet carbon budgets require significant CO2 removal [3, 4] or unrealistic reductions in energy demand [5]. However, one accord across successful pathways is a rapid switch from fossil fuels to renewable energy sources such as nuclear, onshore and offshore wind, solar photovoltaics, hydro, and biomass [3, 4, 5, 6, 7]. It is widely agreed that the energy sector should be the first to decarbonise [4, 8, 9, 10]. This will put renewables at the heart of the transition to a clean economy.

Variable renewable energy (VRE) sources, such as wind and solar, significantly increase the complexity of operating the electricity grid compared with conventional thermal generators (coal and gas). Historically, electricity supply and demand were decoupled from each other; with consumers being provided for by the central dispatch of electricity from large fossil fuel power stations. However, wind and solar are smaller than traditional generators and connect to the electricity grid at a lower voltage. Consequently, their generation is not managed centrally. In the UK for example, at the end of 2019 there were roughly 1 million solar PV systems and 10k wind farms contributing 37 GW of potential power production capacity and producing 67 TWh of electricity annually. However only 17 GW are actively controlled as part of the power system. The other 20 GW are passively contributing - generating weather dependent electricity. Consequently, there is a need to accurately monitor and forecast the generation from these variable renewable energy sources to support grid

operation.

The concern of this thesis is the estimation of uncertainties associated with PV monitoring. Following a brief introduction to the electricity system in Great Britain chapter 2 gives an overview of the background for modelling solar PV systems and a review of existing solar PV monitoring techniques in literature and practical applications of these techniques from grid operators and academic researchers across the world. Then in chapter 3, the uncertainty associated with modelling the solar PV yield is assessed. In chapter 4, the uncertainty in the national solar PV capacity information is modelled for the first time. Finally, in chapter 5, the total error and uncertainty in the national solar PV output estimates produced by the GB PV monitoring solution, PV Live, is calculated by assuming that the uncertainties identified in chapters 3 and 4 are independent and therefore combine in quadrature.

1.1 The GB Electricity System

1.1.1 Electricity system structure

The electricity system comprises generators and networks to transport electricity from where it is produced to where it is consumed. Electricity is a bought and sold as a commodity but unlike most commodities there is little room for supply surplus or shortfall. Energy storage can shift electricity at surplus for use at shortfall, but this process must be managed such that supply, and demand are always balanced. Contrastingly, the gas networks can store surplus gas by operating the network pipes at higher pressure. Therefore, the supply and demand of gas on the network does not always need to be balanced.

National electricity systems are typically managed centrally and owned by the state, or under a regulated private system. In Great Britain the electricity system has been regulated and privately owned since the late 1990's when the 12 regional electricity distribution companies were sold. In the current, privately owned GB electricity system there are a range of organisations that are licensed to operate different components of the overall system under different regulations. The functions involved in the GB electricity system include:

- *Generators*: electricity generators add electricity to the public electricity network for transport to consumers. Large scale generators must have a generation license from Ofgem [11].
- *Networks*: electricity networks are responsible for the transport of electricity from

generator to consumer. There are two types of network; the transmission network which enables bulk transport of power at a very high voltage around the country, the distribution network which steps this power down to connect individual consumers to a power supply. Network companies must either sign a transmission or distribution network licence [11].

- *Suppliers*: electricity suppliers purchase electricity in electricity wholesale markets on behalf of consumers. Suppliers pass through network and policy costs onto consumers which is how the networks make profit. Suppliers must sign a supplier licence from Ofgem to handle consumer contracts [11].
- *The system operator*: the electricity system operator is responsible for ensuring that demand and supply are always balanced. This involves accounting for mismatches in supply and demand from market contracts [11].

Table 1.1 details the different license codes that each type of market participant must sign to be allowed to operate their given role(s) in the GB electricity market. There are four different codes that market participants must sign up to and each code governs different roles and responsibilities:

- *Balancing and settlement codes (BSC)*: The Balancing and Settlement Code (BSC) is managed by Elexon and defines the rules and governance for participation in the balancing mechanism and the imbalance settlement processes [12]. It is a multi-party contract ensuring that imbalance payments for wholesale electricity supply and demand are settled accurately.

- *Connection use of system codes (CUSC)*: The connection and use of system codes are managed by National Grid ESO and are the contractual framework for connecting to and using the National Electricity Transmission System (NETS) [13].

- *Distribution connection and use of system agreement (DCUSA)*: The Distribution Connection and Use of System Agreement (DCUSA) is managed by ElectraLink and is a multi-party contract between licensed electricity distributors, suppliers and generators in Great Britain concerned with the use of the electricity distribution system [14].

- *Grid codes (GC)*: The Grid Code details the technical requirements for connecting to and using the National Electricity Transmission System (NETS). Grid Code compliance is a requirement of the Connection and Use of System Code (CUSC) [15].

- *Distribution code*: The Distribution Code is managed by the Energy Networks Association and details technical specifications governing the role of the 14 Distribution Network Operators (and 12 Independent Distribution Network Operators) in operating and developing the distribution networks in Great Britain, and for the connection of equipment to them [16].

- *Master registration agreement*: The Master Registration Agreement is managed by GemServ and details an agreement which all electricity Suppliers must sign and includes Green Deal obligations. The Green Deal is a Government initiative allowing consumers to make energy efficiency improvements without upfront cost, instead the cost is spread over their electricity bills. The Master Registration Agreement also sets out rules relating to metering and procedures relating to the change of supplier associated with any metering point [17].

- *Smart energy code*: The Smart Energy Code (SEC) is managed by GemServ and is a multi-Party agreement defining the rights and obligations of energy suppliers, network operators and other relevant parties involved in the smart metering management in Great Britain [18].

The electricity system is privatised; the networks, power stations, suppliers, and the system operator are all owned by privately traded companies. These companies operate monopolies and therefore the electricity system is heavily regulated to ensure a fair deal for consumers. The GB electricity system is regulated by the Office of Gas and Electricity Markets (Ofgem). Ofgem's role is to protect consumers interests, facilitate a competitive market, and monitor social (fuel-poverty) and environmental (emissions) issues within the industry. Market participants must acquire a license which sets out how each party must act. Table 1 details the different licences and the codes which each license must comply with. Failure to comply with the codes in any given license area will result in fines and removal of licences. Ofgem monitor compliance with audits and spot checks.

There are some market activities which are exempt from requiring a license:

- *Small scale electricity generators*: electricity generators smaller than $100\ MW$ in England and Wales, $30\ MW$ in South Scotland, and $10\ MW$ in North Scotland can connect to the distribution network without a license [19]. This currently includes all 13GW of installed solar PV capacity in Great Britain, since there are no solar systems larger than 100 MW. However, there are plans for future solar farms which will be larger than 100 MW and require a generator license.

Code	Interconnector License	Transmission License	Distribution License	Generation License	Supply License
Balancing and settlement codes (BSC)	X	-	X	X	X
Connection use of settlement codes	-	-	-	X	X
Distribution connection use of system agreement (DCUSA)	-	-	-	-	X
Grid codes	-	X	X	X	X
Distribution code	-	X	X	X	X
Master registration agreement	-	-	X	-	X
Smart energy code	-	-	-	-	X

Table 1.1: Electricity licensing and code compliance.

- *Aggregators*: electricity aggregators, aggregate smaller generation assets and therefore do not require a license [19].

1.1.2 The role of the system operator

In Great Britain, the System Operator is National Grid Electricity System Operator (NGESO).[1] National Grid ESO is a legally separate business and operates a distinct function to the National Grid electricity and gas networks business. They maintain security and safety of electricity supply in Great Britain which involves managing demand and supply to keep the grid within engineering tolerances. Namely, they must regulate voltage at 230V and

[1] At the time of submission, the UK Government has just announced proposals for a new Future System Operator (FSO) role [20]. To be run by an independent and impartial organisation and operate both the gas and electricity markets.

keep frequency to 50 ± 0.1 Hz. To do this the National Grid ESO calls on generators and ancillary service providers to help balance supply and demand. They do does this by issuing balancing service contracts via the Balancing Mechanism. The Balancing Mechanism is a service in which parties can submit offers for balancing services to the SO for a pre-agreed price [21].

1.1.3 The balancing and settlement code company

Elexon is the Balancing and Settlement Code Company (BSCCo) in Great Britain [12]. Elexon are responsible for managing the balancing and settlement arrangements for the GB sector. Elexon use the Balancing Settlement Code to calculate imbalance settlement charges that arise when suppliers or generators delivery was in imbalance with their stated position.

1.1.4 Market design

The electricity market is a bilateral contract market in which generators and suppliers and agree on the sale and purchase of energy prior to its delivery [22]. Parties agree to purchase and sell electricity via wholesale electricity trading contracts. These bilateral contracts are notified to the System Operator prior to delivery so that the System Operator has knowledge on how parties have agreed to deliver energy. The System Operator can then correct for any imbalance between the stated positions of electricity. This process is called "balancing". When the System Operator corrects for imbalance, the settlement process identifies each user's imbalance volume and levies imbalance charges. In Great Britain, Elexon are responsible for the settlement process. Imbalance charges are prohibitively expensive and difficult to predict. Therefore, imbalance charges incentivise market participants to accurately forecast their electricity consumption and production.

Figure 1.1 shows the structure of the electricity market in Great Britain. Contracts for the final delivery are traded many times before gate closure by the System Operator. Market participants will use the future/forwards markets to hedge some of their risk profile before gate closure and delivery. For example, a supplier might sign a contract with customers for a 1-year fixed deal to lock-in the price for the upcoming year and reduce the risk of price change. Approaching gate closure, market participants will look to align their contracts with their actual expected delivery for each settlement period. Market participants then use the day ahead and intraday markets to correct their stated positions to account for real-time

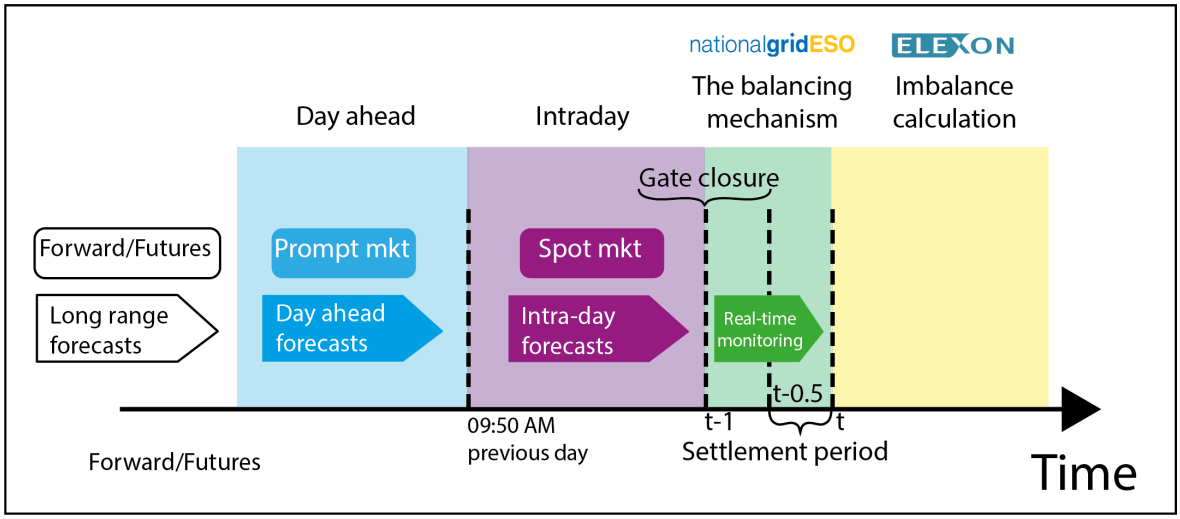


Figure 1.1: *The structure of the electricity market.*

conditions such as the weather and plant availability. At *11:00am* at the day-ahead stage all market participants must submit an Initial Physical Notification to the System Operator of their intended physical output position.

The System Operator announces gate closure *1 hour* before the end of each settlement period. At this point the System Operator takes on the role as system balancer and market participants must submit their Final Physical Notifications (FPNs). Market participants must also submit the maximum capacity of their plant for the settlement period in question. The System Operator uses the FPNs and their own demand forecasts to inform likely activity which they will have to take in the Balancing Market to balance the system. Participation in the Balancing Market by market participants is optional and involves BSC participants submitting offers (to increase generation or decrease demand) and bids (to increase demand or decrease generation). The System Operator must accept these offers and bids in the most economic manner. As well as using the Balancing Market, the System Operator can also procure balancing services in advance. These contracts include the provision of extra capacity to protect against the largest possible trip i.e. the loss of load if the largest generator were to trip off the network. Some of the balancing services being provided are called ancillary services. Ancillary services are contracts provided by market participants which can provide specific types of balancing services such as firm frequency response, short time operating reserve, reactive power, and black start services [23].

At the time of writing, there are no solar PV systems in Great Britain which exceed the $100MW$ limit for the small-scale generator license exemption. This means that solar PV owners are exempt from imbalance charges. However, the need to forecast solar PV

generation exists. Different solar PV monitoring and forecasting services are required relating to the different wholesale electricity markets: long range forecasts for new capacity installation and solar PV output are needed for next summer to help with operation of the forward/futures market, forecasts with a *24 – 72 hour* time-horizon are required for the day-ahead market, solar PV nowcasting services with a time-horizon of *0-6 hours* are needed for the intraday market, and solar PV monitoring services are needed to aid the System Operator with balancing the system in the Balancing Market and for training the aforementioned solar PV forecasting tools. For the use of training solar PV forecasting tools, the solar PV monitoring service only needs to operate at half-hourly resolution to match up with settlement periods. However, for system balancing and managing ancillary services the system operator requires a solar PV monitoring service with a smaller temporal resolution e.g., *5-minutely*, or better.

1.1.5 Networks

The transmission network in England and Wales is operated by National Grid Electricity Transmission and operates at $275\ kV$ and $400\ kV$ [24]. The transmission network in Scotland is operated by Scottish Power Transmission in Southern Scotland and Scottish Hydro Electricity Transmission in Northern Scotland. The transmission network also operates at $132\ kV$, $275\ kV$, and $400\ kV$ [24]. Additionally, there are Offshore Transmission Owners (OFTOs) across Great Britain who operate networks connecting offshore electricity assets (e.g. wind farms) to the mainland.

The distribution network in Great Britain is responsible for delivering power to consumers. It is broken up into 14 distribution license areas which are operated by 14 distribution network operators whose licenses are contracted to 6 companies [25]. Additionally, there are several independent distribution network operators (IDNOs) who operate small sections of distribution network. Normally, to supply a new housing development with a bespoke energy solution [26].

The distribution network is connected to the transmission network through Grid Supply Points (GSPs). The first part of the distribution network is called “extra high voltage distribution network” and operates at $132\ kV$. Afterwards electricity supply passes through bulk supply points which step the voltage down to $33\ kV$. Supply then moves through primary supply points to the “high voltage distribution network” which operates at $11\ kV$. Numerous secondary substations will then lower voltage to the $230\ V$ for use by the end consumer at their property’s connection points.

Traditionally, distribution networks have been concerned only with delivering power from the transmission network to the end consumer. However, renewable generation is being embedded inside the distribution network and this complicates the operation of both the distribution and transmission network. For example, domestic solar PV systems are connected to the distribution network at the lowest level. Therefore, it is very difficult to track power from these systems back up the network to the transmission network because the network is highly meshed through the secondary substations. Furthermore, whilst the topology of the distribution network will conform to a base case assuming standard operating conditions there are many layers of substations and operational constraints will cause power to divert from these standard conditions. Additionally, larger solar PV systems, such as solar farms, are often connected higher in the network at the primary, bulk, or grid supply point. At this level, if there is a network issue with the substation that the solar farm is connected to then it is likely that the solar PV generation will be lost altogether for the duration of the fault.

1.1.6 Demand forecasting

The Electricity System Operator for Great Britain, National Grid ESO (ESO), is required to provide national demand forecasting services to the electricity industry as stated in the Transmission license code. Accurate national demand forecasts enable generators and suppliers to accurately match their wholesale trading contracts with their metered positions. Thus, minimising use of the expensive Balancing Market and reducing reliance on CO_2 producing reserve power. This keeps costs-to-consumers low, minimises the production of CO_2, and ensures a secure and stable operation of the electricity grid. Under the RIIO incentives scheme, the System Operator is regulated to provide four forecasts to the electricity industry: the transmission-connected wind day-ahead, published by 5am; the day-ahead demand, published by *9 am*; the two-day-ahead demand, published by *5 pm*; and the seven-day-ahead demand, published by *5 pm*.

Figure 1.2 shows the daily electricity demand profile with the cardinal points which ESO are regulated to forecast superimposed. The $3C$, $4A$, and $4B$ cardinal points occur only in summer, and the DP point occurs only in winter, which highlights dependence on the time of year.

Each cardinal point can loosely be explained by societal behaviour: the $1A$ peak occurs between *01:00* and *02:00* and is caused by consumers with an *Economy 7* electricity tariff turning on their immersion heater; the $1B$ minimum occurs early morning at around *05:00*

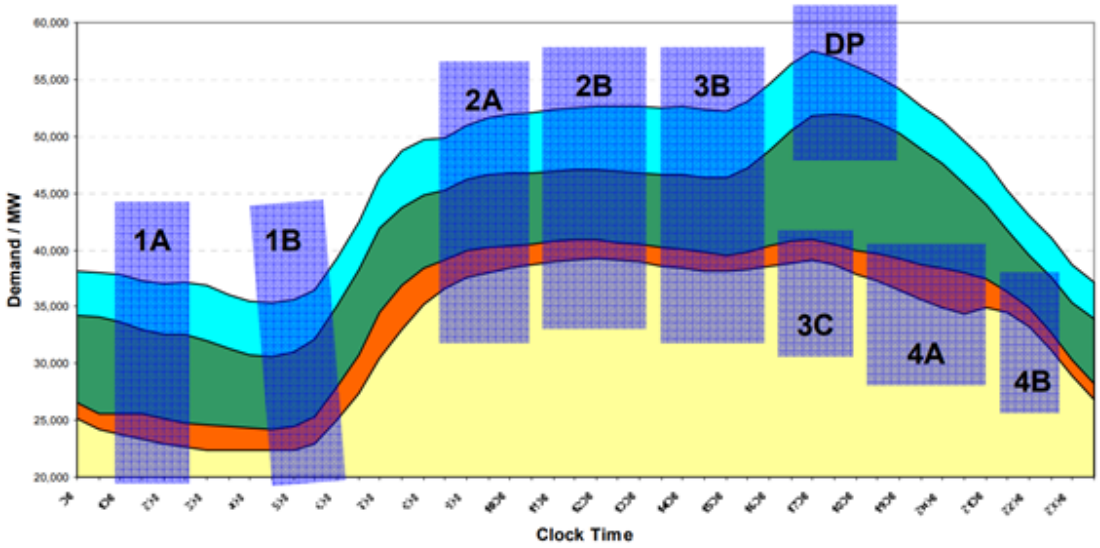


Figure 1.2: *The daily electricity demand profile for Great Britain, the blue and green curves represent winter days and the red and yellow represent summer days [27][24].*

when most people are in bed; the $2A$ peak occurs mid-morning at around *10:30 am* when most people arrive into school/offices; the $2B$ peak occurs just before midday and is caused by lunch time food preparation; the $3B$ point is the afternoon minimum; the $3C$ afternoon peak occurs only in BST and is caused by people getting home from work to make dinner and use in-home appliances; the $4A$ point is the evening minimum in summer; the $4B$ peak occurs only in summer and is caused by people turning the lights on as the sun sets; the DP (darkness peak) occurs only in winter and is the same as $3C$ peak in summer with added lighting demand in winter due to shorter daylight hours.

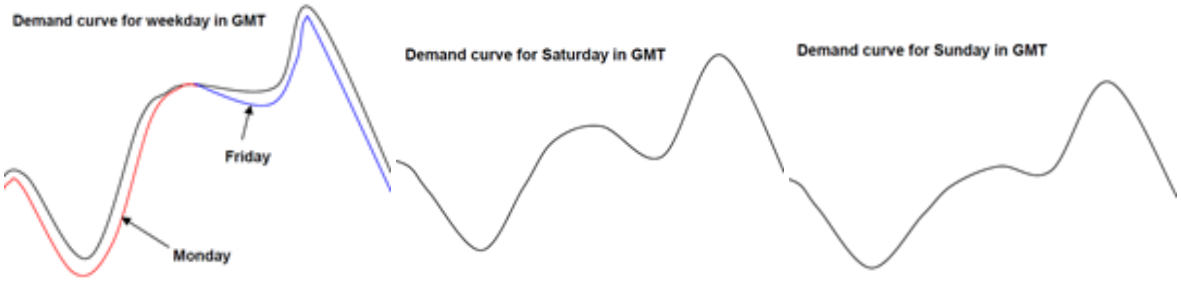


Figure 1.3: *The daily demand profile for weekdays (a), Saturdays (b), and Sundays (c) [27].*

There are many factors which affect each cardinal point in the national demand profile, and they can be summarised by four groups: the weather, the day of the week, the basic underlying demand trend, and the embedded generation. The weather affects each cardinal point roughly equally. Whereas, figure 3 shows that the effect of the day of the week is not universal for all cardinal points; the morning and evening cardinal points on Mondays and

Fridays respectively are suppressed due to reduced industrial load compared with other weekdays. The weekend demand profile is lower for all cardinal points, and the evening peak is flatter, compared with weekdays.

The basic demand profile accounts for long term trends in energy efficiency and levels of electrification in society. Therefore, it also effects all cardinal points equally. However, the embedded generation factor accounts for the reduction in national demand on the transmission network because of embedded generation on the distribution network. The embedded generation effects each cardinal point differently depending on the real-time spatial and temporal variability in the weather and in the spatial and temporal availability of embedded capacity.

ESO forecast demand for every cardinal point and to do this they forecast four separate components [27]:

- *Weather*: a weather component accounts for temperature, wind speed, and irradiance.
- *Day of week*: a day of week component accounts for intra-week variation in the daily demand profile caused by changing industrial loads and demand suppression caused by school and bank holidays.
- *Basic demand*: a basic demand component captures long-term trends in electricity demand relating to energy efficiency and the level of electrification of society.
- *Embedded generation*: an embedded generation component accounts for the reduction in national demand caused by the spatial and temporal variation in embedded generation.

ESO forecast each component individually and combine to estimate the national demand. To forecast each component ESO uses statistical techniques to fit linear regression and machine learning models using historic settlement electricity demand data.

There are three benefits towards ESO's compartmentalised forecasting approach. Firstly, in real-time ESO can compare operational data with forecast data for each component to identify errors in their total national demand forecast. For example, they can compare the input data used to train each component with actual conditions, to estimate each components error. Secondly, when ESO calculate how much reserve to procure for each settlement period they can account for the uncertainty in each forecasted component with respect to its relative volume. For example, on a calm and cloudy winter's day when solar and wind

generation will be low, a large uncertainty in the embedded generation is insignificant with respect to estimating national demand. Finally, when their national demand forecast breaks down this compartmentalised approach makes it easier to identify the cause of the forecast error. This is important for explaining their forecasting performance to the regulator, OFGEM, to avoid prohibitively expensive retrospective fines.

In recent years, the most significant source of error in ESO's national demand forecast has been the uncertainty caused by embedded solar PV generation. To reduce this error ESO have funded three separate solar PV monitoring projects aimed at delivering operational estimates of the solar PV generation in Great Britain. These projects enabled ESO to build a solar PV forecasting service which uses the monitoring data for training [28]. Together, the monitoring and forecasting services have enabled ESO to reduce the uncertainty introduced in their national demand model from embedded solar PV generation.

ESO have two requirements from their solar PV monitoring service: *real-time* solar PV monitoring to facilitate System Balancing and *historic* half-hourly estimates of the solar PV generation so that they can train their solar PV forecast model.

Real-time solar PV data is necessary so that the ESO control room can compare predicted solar PV output with the actual volume of electricity provided by solar PV in real-time. If the solar PV output prediction is high or low, the control room can correct by accepting offers and bids in the balancing market. Additionally, historic solar PV output estimates are needed to enable the training of forecast models which predict future solar PV output.

1.2 PV Live: the GB PV monitoring service

Since 2010, solar PV has become an increasingly important energy source for the GB electricity sector. On a sunny summer day solar PV can provice nearly a third of the GB's electricity supply. For example on the 28th of March 2021 solar PV provided 33.1% [29] of the electricity required by the GB electricity grid. Between 2010 and 2014, as the installed base of solar PV power increased from almost zero to 2.5 GW of solar PV capacity, National Grid ESO (ESO) started to experience errors in their forecast for the afternoon 2A, 2B, and 3B cardinal points. During this period PV generation was almost completely invisible to ESO. PV systems were connected to the distribution grid and even the largest systems were below the threshold for official transmission system metering. To minimise the error in the demand forecast relating to solar PV generation, ESO identified a need to improve real-time information on national solar PV generation.

In April 2014, ESO launched a Network Innovation Allowance (NIA) project titled "PV Monitoring: Phase 1". Prior to the project, the mean absolute forecast error for GB transmission system demand had increased from 322 MW to 422 MW between 2011 and 2014. This increase was mostly caused by embedded solar PV generation. The aim of the Phase 1 project was to set up PV and irradiance measuring systems at three substations in Great Britain. The three stations would provide a proof of concept for a larger measuring system with sites located at all the National Grid owned substations on the GB electricity network. The data from these sites would allow for estimation of the midday demand suppression caused by the distributed solar PV generation.

The project successfully set up two measuring stations but abandoned the third station because of difficulties accessing the substation site. The project concluded that rolling out PV measuring stations across the National Grid substations was not a practical approach towards monitoring solar PV generation because it was difficult to coordinate access to the substations and because connecting new data feeds to the secure operational transmission network was slow. The project recommended that alternative approaches towards solar PV monitoring be investigated.

In August 2015, following on from the first NIA project, ESO launched a second NIA project titled "PV Monitoring Phase 2" in collaboration with The University of Sheffield. The project had two aims: to develop methodologies for estimating the half-hourly national and regional solar PV output and to establish a live data feed to enable real-time national and regional solar PV output estimates for integration into ESO's forecasting operations.

Figure 1.4 gives an overview of the modelling approach for estimating the national and regional solar PV output in Great Britain which was developed in the Phase 2 NIA project between The University of Sheffield and ESO. The modelling approach can be broken down into three components as identified in figure 1.4: 1) creating a model for the PV yield for all PV systems in Great Britain using only data from a small subset of reference PV systems which provide data historically and in real-time; 2) upscaling the modelled PV yield by the known installed solar PV capacity of PV systems to estimate the solar PV generation of every PV system in Great Britain; 3) assigning the solar PV output from individual PV systems to nodes on the electricity network to enable a spatially resolved estimate of PV output which relates to the physical structure of the electricity grid.

The Phase 2 project focused on the first two parts of this process, creating statistical models for the national and regional PV yield[2] and deriving a national solar PV site list. At

[2] see the close down reports from the Phase 2 NIA project for details on the modelling methodology [30]

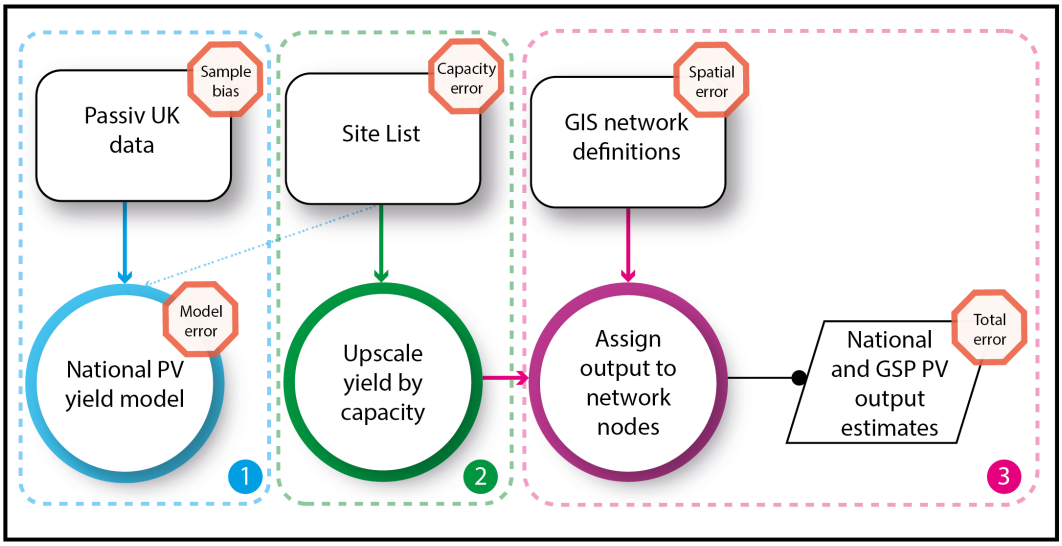


Figure 1.4: *The modelling approach of the GB solar PV monitoring service, PV Live. The diagram separates the model inputs (top row), modelling processes (bottom row circles), and model outputs (bottom row rhombus). Additionally, the modelling approach is broken down into three different parts enclosed by dashed: 1) modelling the PV yield; 2) upscaling the modelled yield using national capacity data to estimate the output for each PV system in GB; 3) assigning the output for each system in GB to a node on the electricity network.*

the start of the project, ESO already produced a national list of all PV systems capacity in Great Britain by synthesising multiple capacity data sources into a single list. However, a lack of validation meant that this approach suffered from double counting, under-counting and incorrectly located systems and consequently there were substantial data quality issues in the site list. Therefore, as part of the Phase 2 project new software for the derivation of a national site list was developed.

The PV Monitoring Phase 2 project was successful, delivering on all its aims and objectives. During the project, Sheffield Solar made a significant contribution towards the acquisition of a new real-time solar PV data feed from Passiv Systems. The Passiv System data feed delivers data from 22k solar PV systems distributed across Great Britain. Using this data, Sheffield Solar developed methodologies for estimating both national and regional solar PV output. They also developed a website [31], named "PV Live", and an API which deliver real-time estimates of the national and regional solar PV output to both the ESO control room and other parties in the energy industry.

At the end of the Phase 2 project, the data feeds provided by The University of Sheffield, Sheffield Solar research group were being used operationally inside the ESO control room. To inform the control engineers on the real-time PV power output and to train solar PV forecast models for use in the day-ahead market planning. In April 2016, when the solar

PV data was first implemented within the ESO control operations a reduction in GB demand forecast error of 100 MW was reported [30]. However, whilst the solar PV output estimates were providing value to the control room engineers, there remained a need for better understanding on the uncertainty associated with the national and regional solar PV output estimates.

In May 2018, a third and final NIA project titled "PV Monitoring Phase 3" was launched. The project was created to further develop the PV monitoring service created in Phase 2. In particular, it was focused on improving the national capacity site list (process 2 in figure 1.4) and on developing a robust methodology to define a lookup between individual solar PV systems and nodes on the transmission network (process 3 in figure 1.4), and on validating the accuracy of the overall model estimates for solar PV output. Four work packages were identified to improve the usefulness and the resilience of the service: developing 5-minutely estimates of solar PV output to better aid the role of ESO as system balancer; improving knowledge of the total installed solar PV capacity; improving the resilience of the data feeds in the service; and validating PV output estimates to better understand the uncertainty associated with the service.

The third project delivered successfully on all four work packages. The PV monitoring service now delivers 5-minutely estimates of solar PV output. Knowledge on the installed base of solar PV capacity has improved significantly with the development of a new approach for calculating a national capacity site list which removes errors in the old approach with double counting. New Geographic Information System files have been produced for defining the lookup between geographic location and topological location on the electricity grid [32]. Additionally, as detailed in this thesis and an accompanying publication, research has been carried out on the uncertainty in the GB national solar PV capacity data relating to both data collection errors and operational errors affecting the installed fleet, such as offline systems and network outages. A new data feed provider, SMA Solar, has been sourced to improve the operational resilience of the service and the representativeness of the sample data used in the model. Finally, as described in this thesis, a complete analysis of the uncertainties affecting solar PV output estimates has been undertaken to validate the accuracy of the solar PV output estimates.

1.3 My role in the Sheffield Solar team

I joined The University of Sheffield, Sheffield Solar research group as a PhD student in September 2017. When I joined Sheffield Solar in 2017, the Phase 2 project was ending and the Phase 3 project was in the planning stages. The PV Live service was important for electricity grid and trading operations both in the National Grid ESO control room and in the wider electricity industry. The PV Live website was achieving roughly 1000 unique monthly users and the API was at least as heavily used.

PV Live data was the de-facto standard for solar PV output data in Great Britain. However, despite this there remained a significant knowledge gap about the uncertainty associated with the solar PV output estimates. My role in the Phase 3 project would be to quantitatively assign uncertainty bounds on the solar PV output estimates produced by PV Live.

During my PhD I have worked on improving all three processes in figure 1.4; I helped develop a process for creating a lookup between multiple disparate capacity datasets for commercial/utility systems which was later used to create the latest national site list compiler software [33]; I developed python software which creates one Geographic Information Systems file containing geographically defined boundaries for the area served by each Grid Supply Point on the transmission network; and as covered in this thesis, I have performed a sensitivity analysis on the national solar PV output estimates which quantifies the total error and uncertainty associated with the national solar PV output estimates from PV Live.

Chapter 2

Background

I'd put my money on the sun and solar energy. What a source of power! I hope we don't have to wait until oil and coal run out before we tackle that.

– Thomas Edison

This thesis is concerned with investigating and understanding the sources of error and uncertainty in the data pipeline for the National Grid ESO PV monitoring service. This chapter provides background on the underlying datasets involved in National Grid ESO's PV monitoring modelling framework and a review of the techniques involved in monitoring and forecasting solar PV yield, power, and energy output. First in *section 2.1*, the characteristics of PV output data are presented. Then in *section 2.4*, PV metadata is presented with respect to the GB system. *Section 2.5* then gives an overview of the existing state of the PV monitoring literature and an international review of current services which monitor national solar PV output. Finally, different metrics for evaluating he accuracy of a PV monitoring service are presented in *section 2.6*.

2.1 Introduction to measuring PV output

To understand PV output data, it is prudent to first explain how electricity is metered. Before Watt-Hour meters, electricity was sold on a per lamp basis [34]. Competition in the electricity market led to the development of an electromagnetic electricity meter to track the electricity consumption of consumers and offer lower prices. Today, there are two general types of electricity meter: electromechanical, and electronic. Furthermore, the latest generation of electronic meters are said to be "smart" meters because they automate the meter reading process.

Electromechanical meters, sometimes called accumulation meters, use electromagnetic induction to turn a wheel such that the rotation of the wheel is proportional to the power flowing through the circuit. They measure the cumulative volume of power which has flowed through the circuit. The wheel will have some resistance which typically requires a few watts to overcome. Therefore, there is a minimum power flow which an electromagnetic meter can measure. However, the rotating disk offers a quasi-continuous measurement of the electrical power flowing through a circuit. In practise, an electromechanical meter has to be read by a human in order for it's output to be useful and so the power readings are quantised with finite precision when measurement is taken. Additionally, these meters cannot detect which time of the day the electricity was consumed.

Electronic meters take quantised measurements of the current and voltage to measure the instantaneous power in the electrical circuit. They do this at regular time intervals and estimate the electrical energy flow through the circuit by calculating the average power in any period and dividing by the length of the period. In Great Britain, electronic meters

record the total energy flow through the circuit every half-hour to align with settlement periods in the electricity market. Unlike electromechanical meters, electronic meters can record the time of consumption for the electrical energy flowing in the circuit.

Historically, in Great Britain, electricity meters were electromechanical and were non-half-hourly metered. Consumers were required to read the analogue display on their meter and submit readings to their electricity supplier at irregular intervals, usually of the order of a few months. In the last 20 years, electromechanical meters have slowly been being swapped for electronic meters. However, historically consumers were still required to read the electronic meters manually and so metering remained non-half-hourly. Consequently, domestic electricity demand data was only available at a high level from metering energy flow at nodes on the electricity grid such as grid supply points.

The UK Government is currently amid a national smart meter rollout in which every consumer is to be offered a smart meter. Smart meters are new electronic electricity and gas meters which connect to a mobile network and automate the process of taking meter readings. Thus facilitating more accurate and regular meter readings on behalf of the customer for the electricity supplier. In 2012, only 0.05% of electricity meters were operating in a smart mode. Whereas at the end of 2020, 27% of electricity meters in Great Britain are operating *"smart"*. Smart meters offer grid operators and electricity suppliers a much more detailed view of domestic electricity demand and will enable new technologies such as demand response to facilitate the energy transition to a net-zero electricity grid.

In Great Britain, electricity meters are assigned a Meter Point Administration Number (MPAN) which is unique and identifies each electricity supply point. An MPAN is defined as personal data and as such any data associated with it is closely protected by GDPR. Meaning that industry and research find it difficult to access the data collected by smart meters. In Great Britain, most MPANs are import MPANs which measure the electrical energy imported into a property. However, if there is onsite generation, then an export MPAN must be installed as well as the import MPAN to measure the energy which is exported from the property to the grid.

When domestic-scale generators, such as solar PV, were introduced the smart meter roll-out was still being planned and most electricity meters were non-half-hourly metered. Therefore, when domestic export MPANs were introduce they were also introduced as non-half-hourly metered. This legacy regulation still exists today and consequently, if a generator is smaller than $30\ kW$, the export MPAN is not required to be half-hourly metered or smart. Meaning that the export meter readings are generally supplied manually and irreg-

ularly by the generator owner to their electricity supplier. Smart meters can measure the half-hourly energy exported to the grid and automate the process of sending readings to the electricity supplier. However, this is not required by the BSC and many suppliers have not configured their smart meters to automatically read exports from onsite generation. Meaning that most export MPANs for generators smaller than 30kW operate in a *"dumb"* manner.

If the onsite generation is larger than $30\ kW$, then the BSC states that a half-hourly export meter must be installed. Systems larger than $30\ kW$ are commercial systems either installed on the ground or on large factory roofs. Half-hourly export MPANs record the net flow of electricity through the MPAN which is exported from the property to the electricity grid. When the onsite generator is producing more electricity than is being consumed, the reading will be positive and vice versa.

MPAN energy readings are uploaded to the Data Transfer Service for use in the retail energy market. Figure 2.1 illustrates some example setups for a domestic and commercial system. The domestic system has one export MPAN and its readings are a combination of the onsite generation and consumption. The PV output signal from this export MPAN will contain a significant volume of noise because of the onsite consumption. Contrastingly, the commercial system contains multiple MPANs and only one MPAN contains onsite consumption. The PV signal in the export data from the MPAN with onsite generation will contain noise relating to the onsite consumption. An additional complication is that the solar PV capacity connected to each MPAN will likely be different. However, for commercial PV systems with multiple MPANs only the whole system capacity is known so the MPANs must be grouped together. This matching can only be done through cross-referencing based on the address associated with each MPAN. Since there are often multiple MPANs in proximity from multiple PV systems and the address for each MPAN can often vary across MPANs associated with the same system, this metering process is very manual and imprecise.

In summary, there are several reasons why export MPAN data is unsuitable for building national scale PV monitoring services: most domestic export MPANs are metered too irregularly for modelling half-hourly PV power output; for all but the largest commercial PV systems, onsite consumption obfuscates the PV power output signal in the data from the export MPAN; where there are multiple MPANs per commercial PV system, grouping them together can be difficult because there often many PV systems in proximity and the capacity per MPAN is unknown. Consequently, electricity market settlement data is unsuitable for modelling national PV power output and alternative PV output data sources

are needed for modelling national PV output.

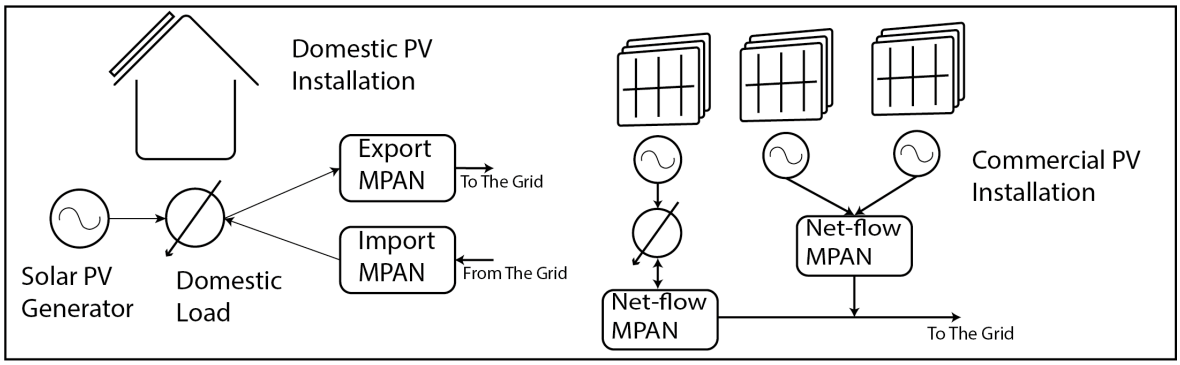


Figure 2.1: *A diagram contrasting the different generation and load profiles which can exist at a domestic solar PV installation and a commercial PV installation. The circle with an arrow through it represents a demand profile and a circle with a sinusoidal profile inside it represents AC supply from a solar PV installation. The domestic PV installation has one export and one import MPAN. The export MPAN measures the electrical energy exported from the property to the grid and is the net of the energy consumed and generated at the property. Whereas, the import MPAN only measures the total electricity imported from the grid. Contrastingly, the commercial PV system is metered by net-flow MPANs which measure the bidirectional flow of electricity into and out of the property (import is negative, and export is positive). In the diagram, the commercial installation illustrates a system setup with multiple MPANs and some onsite consumption. The onsite consumption is only connected through one of the MPANS, the other MPAN contains only generation behind the meter.*

2.2 The reference PV data source

Since 2015, National Grid ESO have been purchasing PV power output data from a company called Passiv Systems to facilitate estimating the total energy output of Great Britain's PV fleet. Passiv Systems supply metered PV output data from a distributed fleet of domestic PV systems. Passiv Systems can offer this service because they offer a commercial service to the solar PV industry in the UK and in Europe in which they automate PV meter readings. The Passiv Systems technology directly monitors the electrical power and energy output from the AC connection of the inverter.

The Passiv Systems data is made available to National Grid ESO for the purpose of modelling the real-time and future (forecast) national and regional PV output. Table 1 details the two data feeds provided by Passiv Systems: a real-time data feed and a historic day-plus-one data feed. The real-time feed provides 2-minutely readings of the instantaneous PV power for $\sim 1k$ PV systems. Whereas the historic feed provides readings of the half-hourly energy for $\sim 22k$ PV systems. The intraday dataset facilitates an estimate of the national

PV power output at the end of every settlement period for use by the National Grid ESO in their role as System Balancer. The day-plus-one dataset facilitates a more accurate historic estimate of the national and regional PV power output. These more accurate estimates are used to train machine learning models for a variety of uses in the electricity retail markets. Namely, for forecasting the day-ahead national and regional solar PV power forecast.

	Intraday	Day plus one
Number of systems	$\sim 1k$	$\sim 22k$
Metering method	Instantaneous power readings	Electrical energy readings
Temporal resolution	2-minutely	Half-hourly
Available from	$\sim$ real-time	9am the next day

Table 2.1: Descriptions of the intraday and day-plus-one Passiv Systems PV energy and power output data feeds.

The Passiv Systems historic day-plus-one PV output data contains measurements of the half-hourly energy output which aligns with the electricity market settlement periods. Contrastingly, the intraday Passiv PV output data contains 2-minutely measurements of the instantaneous PV power. These PV power output readings are used to estimate the half-hourly energy output using equation 2.1. Where $E(t)$ is the electrical energy delivered over time t, P is the power recorded by the metering device and Δt is the interval between measurements.

$$E(t) = \sum P \Delta t \tag{2.1}$$

When the Passiv data is ingested into the Sheffield Solar database, each half-hour period is required to have a full set (15) of readings. If any PV system does not provide the full set of 15 readings for any half-hour period, then no data will entered into the database for that system for the given half-hour. An exception is made at sunrise and sunset because these periods are expected to have fewer readings.

Missed readings arise in the Passiv data for many reasons: the PV system may be offline, an individual meter might go offline, the communications network which the meters use to upload readings might go down, the data from the meter could be corrupted, the Passiv UK

Ltd. server which stores the readings could go down, the API between the Passiv Systems central database and National Grid ESOs database might go down, National Grid ESOs server which ingests the Passiv Systems data could go down.

Data feed errors have different consequences for the intraday and day-plus-one data feeds. If a reading is missed in the intraday data feed, then the derived half-hourly energy data for that PV system will also be missing for the half-hour period in which the missed reading occurred. Consequently, the real-time estimate of national PV power output will be calculated with a smaller sample. The intraday data feed retrospectively downloads missed readings if they become available later and the PV Live model recalculates each estimate every 5 minutes. However, the effectiveness of these updates relies on the end-user of the PV estimates to continually re-poll the PV Live API for updates which may or may not occur. It is unreasonable to assume that this happens all the time. It is likely that users poll the PV Live API once, even if a later estimate is made which makes use of a larger sample of PV system data.

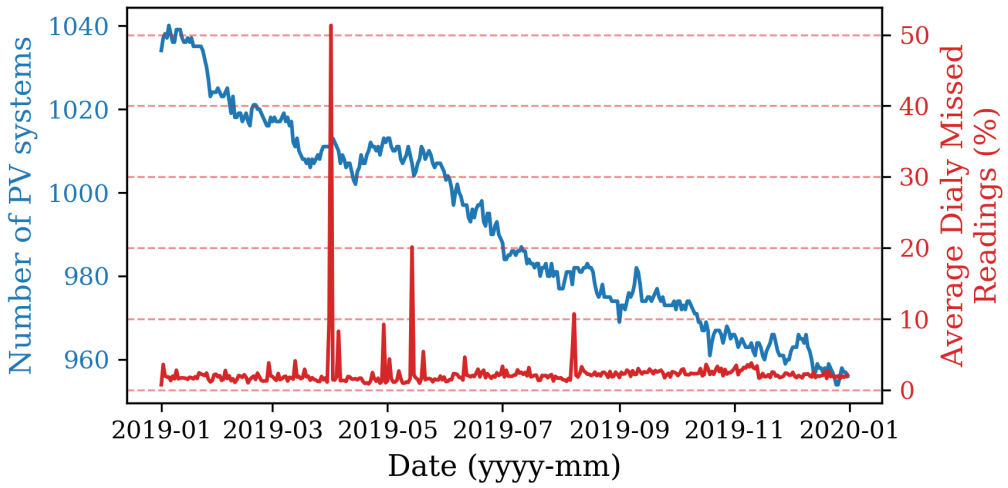


Figure 2.2: *The number of PV systems providing data during 2019 in the intraday data feed from Passiv Systems: the left axis shows the number of PV systems providing data each day (blue) and the right axis shows the mean percentage of missed readings (red) across all PV systems which provided readings for each day.*

In figure 2.2, the number of systems in the intraday data from Passiv Systems is shown along with the average percentage of missed readings per day across all PV systems which provided data. This graph shows that the intraday data feed reduced in size by nearly 10% over 2019. Furthermore, there were three days for which the average number of missed readings was more than 10%. In statistical learning, sample size is often strongly correlated with model accuracy with smaller sample sizes yielding less accurate model estimates [35].

It is important that the relationship between sample size and National Grid ESOs PV yield model error is understood so that the effects of changes to sample size, as seen in figure 2, can be evaluated robustly with respect to the effect on the accuracy and reliability of National Grid ESO's PV monitoring service. This research question is answered in chapter 3.

Figure 2.3 illustrates the difference between the 2-minutely instantaneous PV power and the half-hourly average of the 2-minutely power. The 2-minutely data contains a larger maximum energy output than the half-hourly dataset. This is because the averaging involved in the derivation of the half-hourly value smooths out the extreme measurements of power output.

One feature of PV power timeseries is that the largest instantaneous PV power occurs under broken cloud conditions as opposed to clear sky days [36]. This is because the temperature response of PV panels occurs over 7-minutes. Hence, on cloudy days, panel temperature does not increase significantly due to direct beam radiation because the panels are not irradiated by direct beam radiation for long enough to induce a temperature response in the panels. However, when illuminated by direct beam radiation the photoelectric effect responsible for power production is instantaneous.

The fact that the instantaneous PV power is larger than the half-hourly average power is only important for national PV monitoring if the instantaneous peaks occur concurrently across many PV systems across the country and the concurrency of a distributed fleet of PV systems is discussed in the next section.

2.3 Spatial and temporal variability in PV power

The concurrency of changes in PV power across many distributed systems depends on their spatial and temporal correlation. If all PV systems in Great Britain, which have a total capacity of $13\,GW$, are perfectly correlated. Then the highly variable nature of the 2-minutely data in figure 2.3 will cause large voltage and frequency variations on the GB electricity grid. Under such circumstances, National Grid ESO would have to procure enormous volumes of backup capacity. Incurring large financial and carbon costs.

In practise, PV systems across GB are largely uncorrelated at the 2-minute timescale. However, the spatial and temporal correlation between PV systems determines how many samples are needed to build an accurate PV power output model. If all PV systems in in a

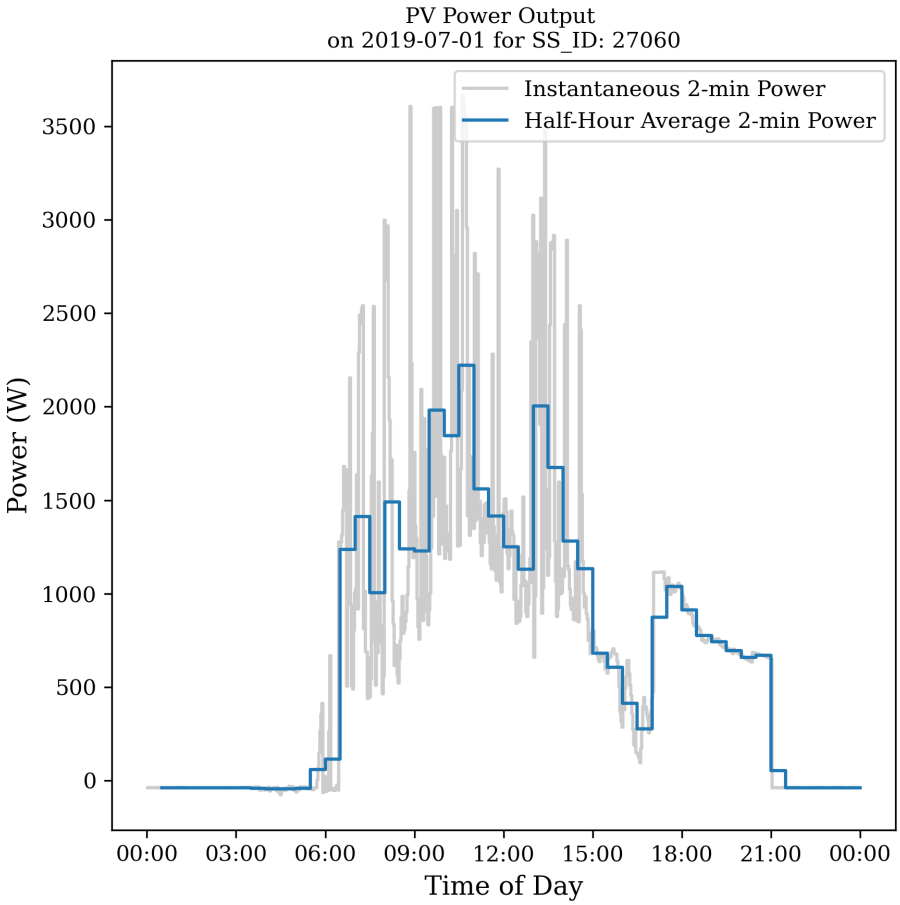


Figure 2.3: *Power output timeseries from one Passiv PV system. Grey is the 2-minutely instantaneous power and blue is the half-hourly average of the 2-minutely data.*

region are perfectly correlated, then only one system is needed to estimate the mean yield for the region. Hence, understanding the spatial and temporal correlation of PV yield across the GB PV fleet is important for estimating the necessary sample size for modelling national PV yield.

2.3.1 The solar resource

Solar PV power strongly depends on the local solar resource and so it is prudent to first discuss its variability. Solar radiation, often called the solar resource, is the electromagnetic radiation emitted by the Sun. The availability of the solar resource is defined using two terms which are often used interchangeably: *"irradiance"* which is the power density of radiation incident on a surface and has units of Wm^{-2} and *"irradiation"* which is the quantity of solar energy arriving at a surface per unit time and has the units of Whm^{-2}

[37].

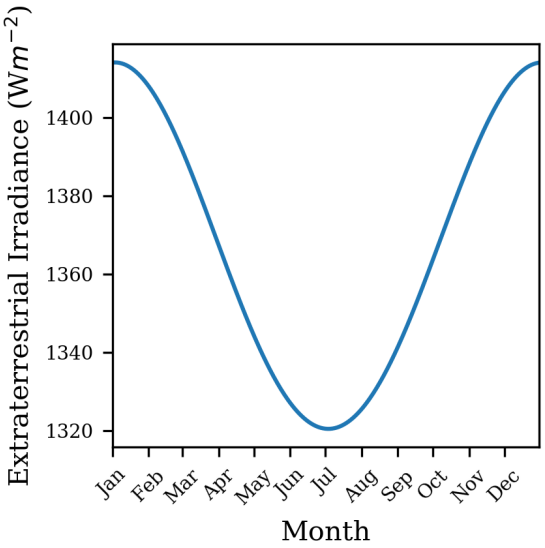


Figure 2.4: *The variation in the extraterrestrial solar irradiance across a year.*

The solar resource incident at the outer atmosphere is called the extra-terrestrial solar irradiance (I_{ext}) and its annual variation is shown in figure 2.4. The energy emitted by the sun is virtually invariant although the earth's orbit around the sun causes a 6.7% inter-annual variation in the extra-terrestrial radiation [37]. The extra-terrestrial irradiance can be calculated using equation 2.2 [38] which makes use of the solar constant and by the eccentricity of the earth's orbit.

$$I_{ext}^{norm} = SC\left[1 + 0.033\cos\left(\frac{360N}{365}\right)\right] \tag{2.2}$$

where SC is the solar constant and N is the day of year. The latest value of the solar constant is defined to be $1366.1Wm^{-2}$ [38].

Figure 2.5 [39] shows the different components of solar radiation at the earth's surface: direct, diffuse, and reflected. The radiation splits into the direct and diffuse components because it scatters as it passes through the earth's atmosphere. There are many particles in the atmosphere which could cause incoming radiation to scatter. Namely, cloud droplets and aerosols (dust, pollutants, smoke e.g. from volcanoes or forest fires).

The total solar radiation incident on a horizontal surface can be defined in terms of both the direct beam and the diffuse radiation. To do this the diffuse horizontal irradiance (DHI) and the direct normal irradiance (DNI) must be defined. The DHI is the diffuse radiation incident at a flat surface horizontal to the earth's surface [37]. The DNI is the direct

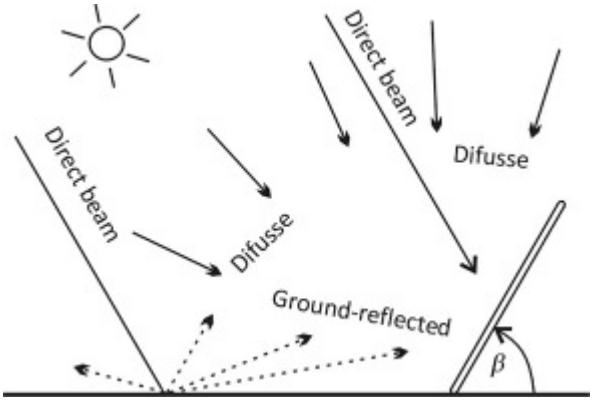


Figure 2.5: *The different components of solar radiation at the earth's surface [39]. β is the tilt of the panel measured from the horizontal plane at the earths surface. The direct beam is the solar radiation incoming parallel to the line drawn at the solar zenith angle (the angle between the sun's rays and the vertical [38]). The diffuse radiation is the solar radiation which has been scattered as it passed through the earth's atmosphere [37]. The ground reflected radiation has generally been reflected off of surfaces such as water, glass, and snow.*

beam irradiance incident on a horizontal plane perpendicular to the line drawn by the solar zenith angle [37]. The global horizontal irradiance (GHI) can then be defined, as the total radiation incident at a surface oriented horizontally at the earth's surface [37]. A mathematic derivation of the GHI is detailed in equation 2.4. The amount of GHI reaching the earth's surface and the ratio between the DNI and the DHI depends on many factors such as: location, time of day, season, local shading, and the weather.

$$GHI = DNI \times \cos(Z) + DHI \tag{2.3}$$

where DNI is the direct normal irradiance, Z is the solar azimuth angle - the angle describing the suns position in the horizontal plane taken by projecting the centre of the sun onto the horizontal plane and is generally measured as the angle such that South=0, East<0, and West >0, and DHI is the diffuse horizontal irradiance.

The clear sky irradiance, the clearness index, and the clear-sky index are three important variables in solar PV modelling and forecasting. The clear sky irradiance is the practical upper limit on the global horizontal irradiance when observed under perfectly clear skies with no albedo, clouds, or aerosols [40]. The clearness index is a unitless variable defined as the ratio between the global horizontal irradiance and the extraterrestrial irradiance [41]. Similarly, the clear sky index is defined as the ratio between the global horizontal irradiance and the clear sky irradiance [42].

Clear sky irradiance models can broadly be categorised as either empirical or physical

[43]. Physical models are based on radiative transfer models (RTM). They simulate the radiation attenuation for each layer of the atmosphere [43]. Example physical models are the SOLIS [44] and the MAGIC [45] models which are used in the CM-SAF and Metosat Second Generation satellite-based irradiance observation models, respectively. Empirical models are parameterised simplifications of the full physical attenuation process [43] which estimate the clear sky irradiance using only limited atmospheric inputs. Some example of empirical models are the simplified SOLIS model (sSOLIS) [46], REST2 [47] and ESRA [48]. All three of these models have been shown to be among the best performing models by today's standards and have a normalised RMSE of less than 5% [43]. Physical models are prohibitively computationally expensive for all but the most sophisticated use cases (e.g., satellite-based observation models) and so in most PV research empirical approaches are used instead.

2.3.2 Variability in solar resource and PV power

The availability of solar irradiance is strongly correlated with solar PV output, as seen in an example timeseries for a single PV system in figure 6 [49]. For this reason, solar irradiance has often been used as a proxy for solar PV power in research [50, 51, 52, 53].

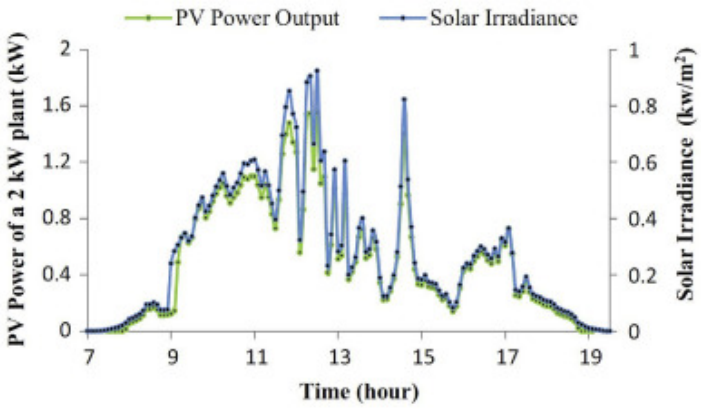


Figure 2.6: *In-plane solar irradiance and PV power timeseries for a single PV system for one day [49].*

Early studies investigating the variability of the solar resource at the earth's surface related to understanding variability in solar irradiance at a single location. In the 1990's the frequency distribution of short-term (1-5 min) irradiance measurements was shown to be more bi-modal than that of hourly data [54, 55]. Highlighting the on/off nature of the solar resource which occurs because of cloud movements. However, literature has also shown that the sub-hourly Global Horizontal Irradiance and the Direct Normal Irradiance

can be modelled as a function of the hourly clearness index and it's inter-hour variability [56, 57, 58]. Over time, extensive literature has developed which uses hourly irradiance data to model sub-hourly solar PV power and energy [59, 60, 61, 62, 63, 64].

In 1997, the first research investigating the variability in the solar resource at multiple locations was carried out by Otani et al. [65]. In their study they investigated the correlation of solar irradiance at 9 sites in Japan spread over a $4\ km$ by $4\ km$ region and showed that under broken-cloud the variability of the 9-sites average was $20 - 50\%$ lower than each independent site. Four years later, Wiemken et al. [66], using solar PV output as a proxy for the availability of the solar resource, showed that the power output changes for the average of 100 PV systems distributed across Germany were significantly smaller than the independent system changes. In 2006, Kawasaki et al. [67] informally named this the "Smoothing Effect" in a follow on experiment from the 1997 Otani study.

The next development was research into the correlation between pairs of PV systems distributed in space and time. Murata et al. [68] were the first to do this when they analysed the short-term correlation between the output of pairs of PV systems from a sample of 52 PV systems in Japan. They found that the PV systems correlation decreased as the distance of separation increased. A year later in 2010, Hoff and Perez [52] formalised this empirical result by defining a mathematical relationship for calculating the output variability of a fleet of identical PV systems (same capacity, orientation, and spacing). They showed that for a national PV fleet the output variability of the fleet is proportional to the inverse square root of the number of systems in the fleet.

In 2012, Hoff and Perez [51] expanded on the results from their 2010 paper by deriving an analytical approach for calculating the maximum short term output variability from an arbitrary fleet of PV systems. They found that the maximum possible variability of a fleet of distributed arbitrarily sized PV systems is equal to the capacity divided by $\sqrt{2N}$, where N is the number of systems in the sample.

Figure 2.7 [51] shows results from the 2012 Hoff and Perez study in which they investigated the correlation in changes in the clear-sky irradiance between many station-pair locations. The results shown are consistent with previous studies in that the correlation coefficients decrease with increasing distance and decrease more slowly with increasing periods of observation [68]. They show that for a period of observation of 1 hour, pairs of PV systems are correlated for a separation of up to roughly $50\ km$. Indicating that to model the hourly PV power for PV systems distributed over a large geographical area, the model must be able to resolve PV power with a spatial resolution of $50\ km$ squared. Over time

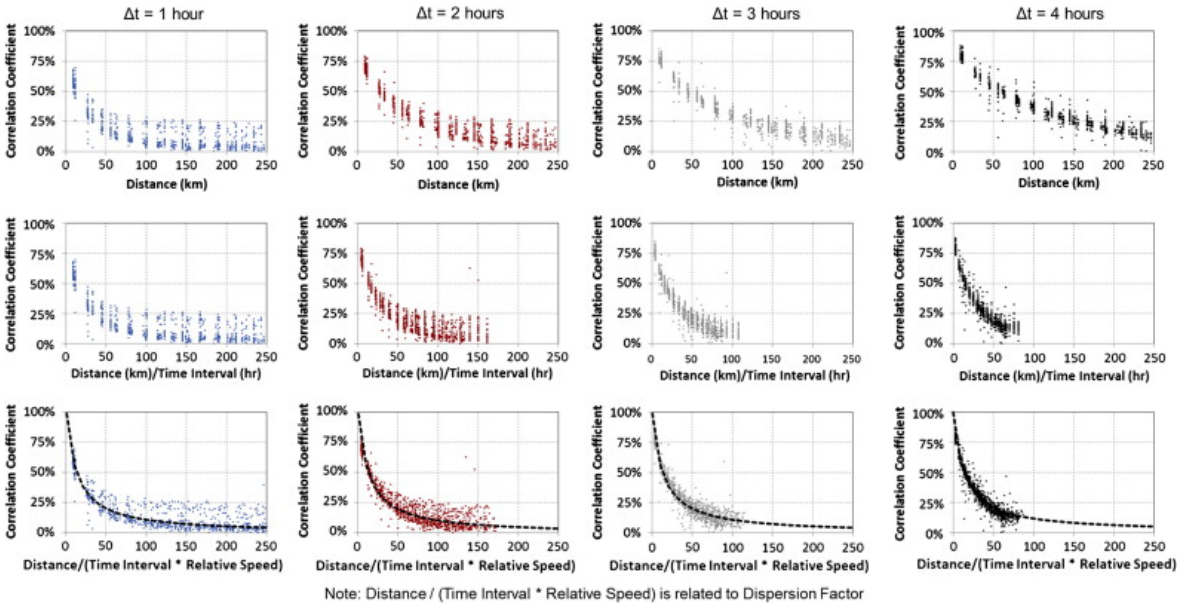


Figure 2.7: *Station-pair correlation: correlation coefficients for changes in the clear-sky index, calculated using satellite-derived irradiance data, between pairs of locations in the Southern Great Plains in the United States [51]. The columns show results for different periods of observation and the rows represent the correlation coefficients plotted against three related variables: separation distance, separation distance divided by time interval, and separation distance divided by time interval multiplied by the an indirectly measured relative cloud speed. The cloud speed relates to the dispersion factor which they introduced in their 2010 paper [52]. In the bottom row, the dashed line represents their model for station-pair correlation which is shown to be a very good fit to the data in this analysis.*

an extensive literature has developed on the temporal, spatial, and site-specific short term variability of fleets of distributed PV systems [36, 57, 69, 70, 71].

2.3.3 Implications for PV power monitoring

The "Smoothing Effect" implies that the peaks in the 2-minutely power data in figure 2.3 will be smoothed out in the national PV power profile. If the variability is low then this should mean that the half-hourly average of the 2-minutely power readings should provide an accurate unbiased estimate of the true half-hourly measured energy.

In figure 2.8, the Smoothing Effect has been visualised in the 2-minutely instantaneous PV power data from the intraday Passiv Systems data. Four random subsamples of sizes 1, 10, 100, and 1000 were selected using sampling with replacement. Aggregate yield timeseries for all PV systems in each sample were computed and timeseries for the 1st of July 2019 have been plotted in figure 2.8. The variability can be seen to significantly decrease

with each larger size of aggregation. The timeseries is significantly smoother for the aggregate of 1000 systems than for the single system. Thus, illustrating the smoothing effect in the Passiv Systems intraday dataset.

The highly variable nature of the single system timeseries means that the average of the 15 2-minutely PV power samples could result a significant over or underestimation of the PV power for a single system if all the measurements happen to be taken at peaks in the timeseries. However, the "Smoothing Effect" as demonstrated in figure 2.8, means that for large numbers of PV systems the aggregate 2-minutely PV power timeseries is smooth. This conclusion can be stated formally using Hoff and Perez's relationship [51] for the maximum output variability of a PV fleet. There are 1 thousand PV systems in the intraday 2-minutely data feed with a capacity of roughly $3\ MW$. Therefore, as per the Hoff and Perez relationship the maximum variability of the sample is 0.1%.

The significance of the smoothing effect for modelling national PV power depends on the exact model configuration and its inputs and outputs. National Grid ESO's national PV monitoring solution estimates a representative yield by taking the mean yield of a geographically optimised sample of PV systems. A full breakdown of the modelling methodology is given in a later chapter of this thesis. The size of this representative sample is of the order of 1000 systems. Therefore, the smoothing effect means that the 2-minutely data is sufficient for modelling the half-hourly energy of the individual systems since any errors in the half-hourly energy estimates of individual systems will be smoothed out when aggregating in the calculation of the representative mean yield.

Figure 2.9 illustrates the correlation between pairs of PV systems in the intraday Passiv data as a function of their haversine distance-of-separation. In this analysis, the Passiv sample was resampled for four different periods of observation (2 minutes, 10 minutes, 30 minutes, and 1 hour). The variability of each PV system was then evaluated by calculating the change in yield from one period to the next. Finally, the correlation coefficients between all possible combinations of system-pairs were calculated. The correlation coefficients have been plotted in figure 9 as a function of the haversine distance of separation between the PV systems for each period of observation.

The relationships shown in figure 2.9 closely follow the relationships shown in figure 2.7 from the 2012 Hoff and Perez paper, however, they appear to decrease faster in 2.9 than in 2.7. In the Passiv correlation analysis, metred PV power readings have been used to calculate the correlation coefficients. Whereas, in the Hoff and Perez paper [51], Hoff and Perez used the clearness index as a proxy for PV system data. Clearness index is a good

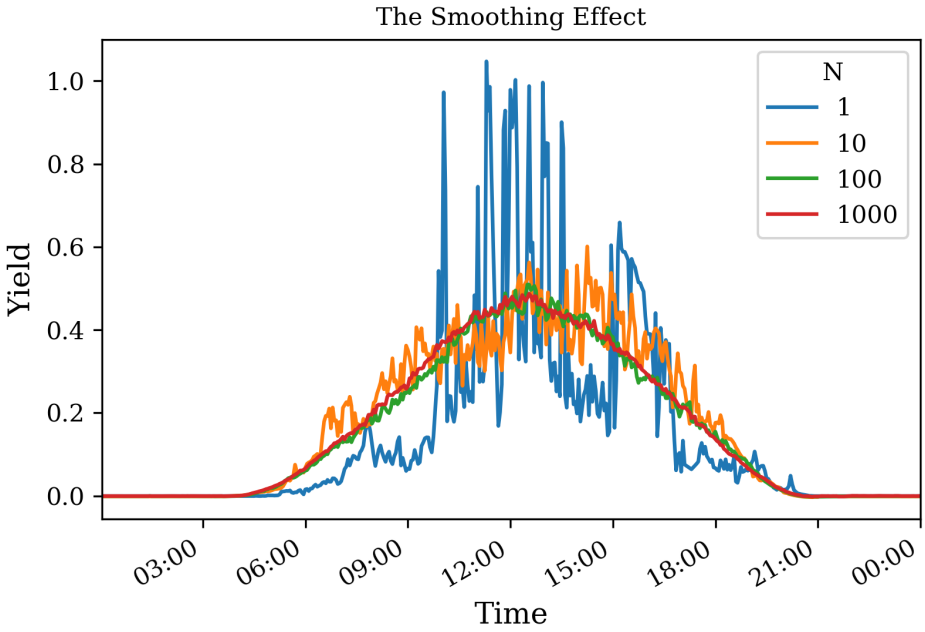


Figure 2.8: *The Smoothing Effect. The 2-minutely instantaneous PV power readings from the intraday Passiv Systems data has been aggregated for four different sample sizes. Blue is the single system, orange is* 10 *systems, green is* 100 *systems, and red is* 1000 *systems.*

descriptor for the spatial and temporal variability of PV, however, it will not correlate perfectly with PV power due system configuration factors which both vary randomly between systems and affect the power output of PV systems, such as shading and system orientation and tilt. Therefore, it is expected that pairs of real-world of PV systems will have lower correlation than pairs of clearness indices and therefore the correlation of PV systems should decrease faster with increasing distance and time. The relationships in figure 9 add to existing results in literature [18–20,25,36] which state that the correlation of solar PV power decreases with increasing distance of separation.

National Grid ESOs PV monitoring system is half-hourly to align with settlement periods in the electricity market. The bottom left graph in figure 2.9 shows that at a half-hourly period of observation, PV power output is strongly (> 60%) correlated up to distances of $10\ km$ or larger . Indicating that National Grid ESOs national PV power model be able to resolve PV power at a resolution of roughly $10\ km^2$ squared.

This estimate of the distance-of-correlation can be used to estimate the sample size which would be needed to fully capture the variability in PV power at a half-hourly temporal resolution. Great Britain is roughly $200\,000\ km^2$. Therefore, assuming PV capacity is distributed evenly across Great Britain, it is estimated that $20k$ PV systems will be required

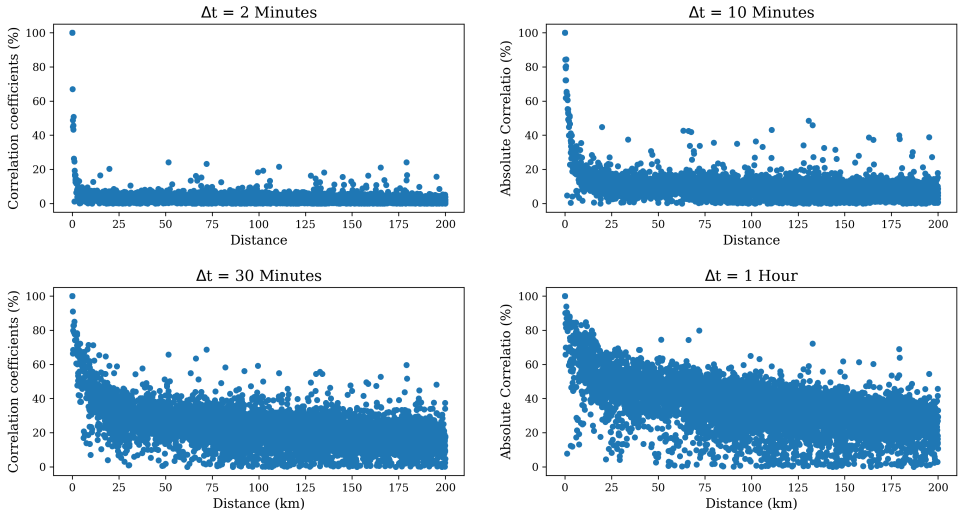


Figure 2.9: *The station-pair correlation coefficients calculated for the intraday Passiv sample for four different observation periods (2 minutes, 10 minutes, 30 minutes, and 1 hour).*

to capture all the variation in PV power across Great Britain. However, PV capacity is not evenly distributed across Great Britain, so the true sample size is likely to be smaller than this estimate. The exact number of PV systems necessary to model GB PV power will depend on the geographical distribution of both the GB PV fleet and of the sample and will be explored in detail in this thesis.

2.4 Characteristics of PV metadata

Solar PV metadata is needed for monitoring and forecasting the current and future solar PV output for systems which are unmetered. National solar PV monitoring services work by modelling the solar PV yield using data from reference systems and upscaling using solar PV metadata for the entire fleet of PV systems. Therefore, there is clearly a need for a consolidated list of the metadata associated with all solar PV systems.

For monitoring solar PV output, solar PV metadata must characterise each solar PV system using a minimum of three parameters: the date on which the system started generating electricity, the direct current (DC) capacity of the panels of the solar PV system, and the location of the solar PV system.

The start date is needed for two reasons. To inform when the system started contribut-

ing electricity to the grid and to estimate the performance induced degradation which occurs to the system as it ages due to damage from the solar irradiance. Performance induced degradation depends on the local climate [72] and for the UK, systems degrades by $0.8 \pm 0.1\%$ on average annually [73].

The location of the system is needed because the yield of the PV system is largely weather dependent and for assigning solar PV output to nodes on the electricity network. Therefore, the yield of the system will vary significantly across the country in relation the specific weather conditions in each location. Knowing the location of each system allows the calculation of a spatially resolved yield which closely matches the spatial distribution of the installed fleet of solar PV systems.

In the United Kingdom and in the EU, the exact location of a domestic solar PV system is considered personally identifiable data and as such is covered by GDPR. However, as per figure 2.9, half-hourly solar PV output in GB is correlated up to distances of $\sim 10\ km$. Therefore, for use in solar PV monitoring services domestic solar PV systems can be grouped using spatial regions of aggregation which are smaller than the distance of correlation for the specific period of observation required. For example, PV systems could be aggregated across a $\sim 1\ km$ grid in the UK for half-hourly use.

Often utility systems are in fields with hard-to-define locations and currently, there is no standard for defining the location of a utility scale system. Sometimes the location of a utility-scale system is defined as the adress of the farmhouse associated with the owner of the land but often this can be km's away from the location of the actual system. Sometimes utility-scale systems are identified using the postcode of a nearby road. However, often rural postcodes cover large areas, and the geocoded locations can often lie km's away from the actual location of the system. In some cases, utility scale systems are identified by their companies head office which can be in a completely different area of the country to the system. Traditional postal addresses are clearly unsuitable for defining the locations of a solar PV systems as a system may be installed in a field or on the roof of a house, factory, or outbuilding.

The DC capacity of the Panels is needed so that the performance of PV systems of different sizes can be compared. It is important that the direct capacity of the panels is used instead of the maximum power output from the alternating current connection to the inverter (AC capacity). This is because the size of the inverter with respect to the size of the solar PV system varies between installations. To calculate the system yield, the capacity is used as a scaling factor between different systems of different sizes. Therefore, if the

AC capacity of the inverter were used then the variability of inverter size to system size between different systems would contribute towards uncertainty in the modelled yield for each PV system. Hence, the DC capacity of the panels is used instead since this value is comparable across all solar PV systems regardless of the size of the inverter.

Additionally, the orientation and tilt of the PV systems is useful metadata but not essential. The orientation and tilt of the PV system strongly affects the solar PV yield of the system. These details are difficult to manually record and are seldom included with recorded solar PV metadata. However, the orientation and tilt of solar PV systems can be calculated from data provided by reference solar PV systems [74]. The distribution of the orientation and tilt from the reference PV systems can then be used to stochastically estimate the orientation and tilt of the entire fleet of PV systems. This approach is widely used in the solar PV monitoring tools in both academia and industry [75, 76].

2.4.1 GB solar PV metadata sources

In Great Britain there have been many different subsidy schemes; the Feed-in-Tariff (FIT), Renewable Obligations Certificates (ROC), Contracts for Difference (CfD). As such there are many different sources of solar PV metadata associated with tracking the installations of solar PV systems for each subsidy scheme. However, this system data was collected to facilitate the subsidy administration. The data was not collected with PV output modelling in mind and this has resulted in inadequate data quality for modelling purposes and poor interoperability between datasets.

Electralink

Electralink were founded as a company in 1998; their remit was to provide a secure independent and low-cost data transfer system for all parties in the electricity sector [77]. Electralink facilitate the data transfer service and as such they collect data for all electricity meters on the distribution system. By analysing the timeseries of the net electricity flow through each MPAN, Electralink have identified MPANs associated with solar PV systems and for these systems they have collated PV system metadata by cross-referencing the address of the MPAN with address information in other governmental capacity datasets, such as the REPD.

The Renewable Energy Planning Database

When a PV system is installed, the owner must apply for planning permission from their local planning authority. These planning applications are collated on a national scale by BEIS and documented in one database called the Renewable Planning Database (REPD). The REPD comes with a warning that the metadata recorded is accurate for the proposed PV system only and may not accurately reflect the installed PV system.

Subsidy datasets

The Office of Gas and Electricity Markets (OFGEM) are responsible for administering subsidies for PV generation. There are different types of subsidy depending on the size of PV system and when it was installed; the Feed-in-Tariff (FIT), Renewable Obligation Certificates (ROCs) or Contracts for Difference (CfDs). To administer the different subsidies OFGEM need to know the metadata for a given PV system in each scheme. Therefore, OFGEM has a list of PV metadata for any PV system in Great Britain which has been awarded government subsidy. The accuracy of OFGEM's data, relies on accurate data entry by the PV installers. Furthermore, the FIT was scrapped in April 2019 so there is currently no subsidy scheme for domestic PV and consequently no register tracking the new domestic capacity. OFGEM supply data in two separate databases; one database for all the FIT qualifying systems, one database for all other systems that qualify for subsidy e.g. ROCs or CfDs.

The Embedded Capacity Register

The District Network Operators are required under their license agreement to publish a list of all systems larger than 1 MW which are connected to their distribution system. Separate registers are published by each of the 12 district network operators.

Micro-generation Certification Scheme

Most domestic systems are accredited through a trade organisation scheme known as the Microgeneration Certification Scheme (MCS). For a system to be accredited, the installer must be certified with the MCS and a fee must be paid upon registration. The MCS has a register containing data on all accredited systems, but this data is not publicly available.

Solar Media

Solar Media are a solar trade organisation. They collate PV system metadata from a variety of sources including news reports, local press, and the databases already named. They claim to keep a comprehensive list of all GB PV systems $\geq 300\ kW$.

In 2019, the UK Government commissioned the Energy Data Task Force (EDTF). The EDTF was a committee led by the Energy Systems Catapult and was tasked with identifying the changes needed to energy data collection, processing, storage, and dissemination. They surveyed the energy industry to assess the needs and requirements of energy systems data. In June of 2019, they released a report with the findings of their survey [78]. In the report they stated five recommendations titled: digitalisation of the energy system, maximising the value of data, visibility of data, coordination of asset registration, visibility of infrastructure and assets. These recommendations should result in one consolidated PV system data list and should eliminate cross referencing problems. However, as of March 2021 their recommendations have not been implemented.

2.4.2 Compiling national metadata lists

Different countries have approached the problem of collating national solar PV metadata lists in different ways. Germany [79] and Italy [80] have both legislated for one consolidated source of solar PV capacity information which is collected and maintained by government. The German register is the only example of a publicly available consolidated solar PV capacity dataset. Whereas, to access the Italian dataset users must register with the Italian transmission system operator and obtain a login for their GAUDI system. The German and Italian registers are the only registers which provide mechanisms for recording when a solar PV system has been decommissioned. However, despite the existence of the mechanism in the German register since 2017, no PV systems have been recorded as decommissioned as of March 2021. Obviously, this is not reflective of the true number of decommissioned systems in Germany as the PV sector is 20 years old.

Where consolidated national solar PV registers are not made available by a government or transmission system operator. National solar PV site lists have been derived by compiling multiple disparate and overlapping sources of solar capacity information.

Our method, as used by National Grid ESO produce the GB national solar PV site list synthesises four sources of capacity information: the Government Feed-in-tariff register

[43], the Renewable Energy Planning Database [81], Renewable Obligation Certificate register [82], and a commercial capacity dataset from Solar Media [83]. These datasets include location, system install date, and system size and are brought together, and cross checked to ensure that there is no double counting. A single comprehensive site list is compiled (GitHub [33]) which includes every known system in Great Britain. Systems which occur in one or more source registers but do not have location data are allocated pseudo-randomly according to the likelihood of location based on the capacity-weighted spatial kernel density estimate of known PV systems. The growth in the GB capacity as measured by the site list compilation [33] is shown in figure 2.10.

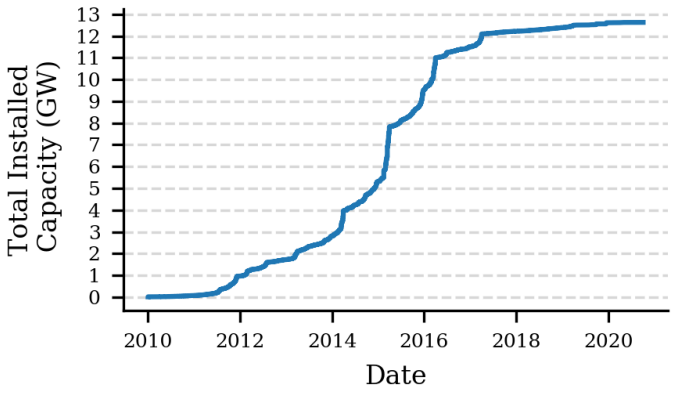


Figure 2.10: *The cumulative total installed (DC) capacity of the panels in the GB PV fleet.*

The Dutch Solar PV service [84] uses a similar approach to GB, compiling capacity information from governmental, commercial, and scientific reports. Public data is collated from the Nationaal Solar Trendrapport, data from the Dutch Central Bureau and Statistics, and the Klimaatmonitor. Additionally, some confidential and closed data sources are included from Solar Monkey, Eindhoven University of Technology, and the Dutch transmission service operators (Stedin and Alliander).

The Australian solar PV monitoring service [85] also collates data from a multiple sources. Data for small-scale systems with capacities less than $100\ kW$ are selected from the Clean energy regulator's Small-Scale Generation Unit database. This dataset includes most domestic systems installed since 2001 but it has no mechanism for handling decommissioned systems. Solar PV systems with a capacity greater than 100 KW are taken from the Large-scale Renewable Energy Target database and this database does have a mechanism for tracking decommissioned systems.

The Open Power System Data (OPSD) [86] project attempts to centralise and standardise power system data for many European nations. It documents capacity data for 1.7 million

renewable power plants across Germany, Denmark, France, and Poland. This capacity data is available to download in standardised csv files. However, it cannot publish commercially sensitive datasets such as are used in the Dutch and the GB solar PV site lists. Therefore, the GB site list on the OPSD is different from the site list used in the GB solar PV monitoring service because it does not include the commercial data from Solar Media.

In conclusion, collating metadata for renewable assets is important for their efficient grid integration. Germany leads the way in this regard and is the only country with one consolidated dataset of renewable asset metadata which is made publicly available. Where public datasets are not available, users of renewable metadata are left to produce their own lists using different disparate sources of metadata. As is the case with the GB, Australian, and Dutch solar PV monitoring services, different sources of metadata exist due to multiple subsidy schemes for different sizes and types of system and often these different datasets have significant overlap. This makes compiling one national solar PV site list complicated and time consuming. The Open Power System Data project is attempting to remove the need for this complicated and time-consuming work by publishing cleaned and standardised datasets of metadata for renewable assets for countries in Europe.

2.5 PV Monitoring Review

2.5.1 Correlations with PV power output

There are environmental and site-specific factors which affect PV system output. Environmental factors include temperature, wind speed, irradiance, and the solar position. Site-specific factors include: the orientation and tilt of the panels, the quality and performance of the specific panels in the installation, the size of the inverter in relation to the DC capacity of the panels, the quality of the inverter, the quality of the wiring and fitting, the amount of shading, and whether the panels are soiled.

Irradiance

Figure 2.6 demonstrates the relationship between solar irradiance and PV power output on a clear-sky day in Kuala Lumpur, Malayasia. Figure 2.11 shows a positive correlation between PV power output and solar irradiance and the data for this plot was taken from the same experiment in Kuala Lumpur as in figure 2.6. The correlation coefficient in figure 2.11

is 0.988 indicating a very strong correlation. Hence, solar PV power is largely governed by the solar irradiance incident at the panels.

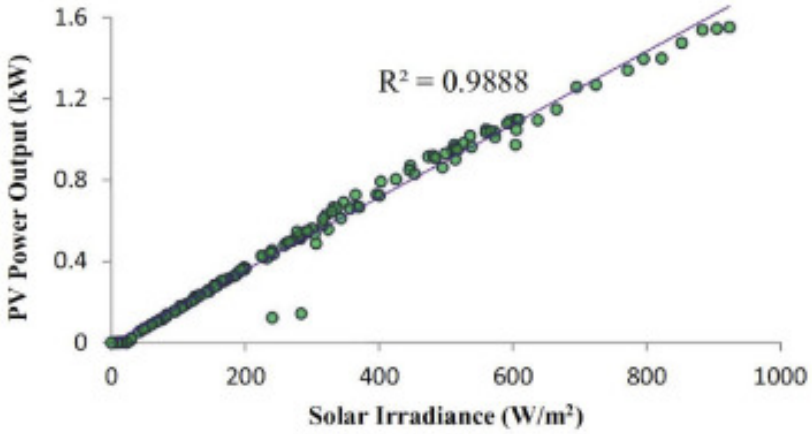


Figure 2.11: *Correlation of PV power output and solar irradiance for a solar PV system located on the roof of a building at the University of Malaya, Kuala Lumpur, Malaysia [17].*

Clouds

Given that solar irradiance is the most significant factor effecting solar PV power, it follows that factors which directly affect the solar irradiance incident at the surface of the earth will indirectly affect the solar PV power output of a PV system. One such factor is the cloudiness. The volume of solar irradiance which arrives at the earths surface is dependent the number and type of clouds [87]. Therefore, cloud motion, birth, and dissipation are important factors for solar PV power output due to their effect on solar irradiance.

Traditionally, clouds have been modelled using satellite images. However, historically satellite-based images only achieve a resolution of 16 - 50 km [88]. Recently, satellites with a resolution of 3km per pixel have been deployed [89] and researchers, such as OpenClimateFix [90] are trying to make use of these more accurate satellite images for nowcasting clouds and solar irradiance, for short term (0-6 hour horizon) solar PV forecasting applications. However, these techniques are still in the developmental stages.

Energy-markets mostly use half-hour metering and figure 2.9 indicates that for this period of observation solar PV models must be able to depict changes in solar PV power across distances of less than ~ 10 km. Therefore, traditional satellite-based cloud models lack sufficient resolution for modelling solar irradiance for solar PV monitoring and forecasting purposes. Because of this, researchers have started using terrestrial sky-imaging for monitoring the movement of clouds [87]. However, this approach is restricted almost ex-

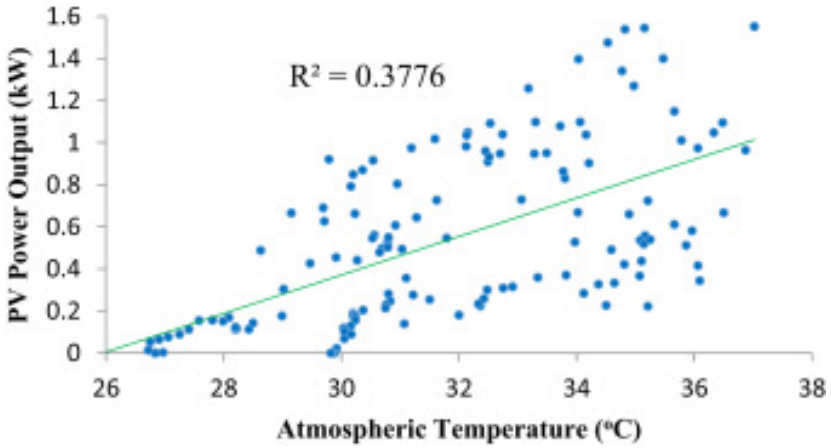


Figure 2.12: *Correlation between PV power output measured in kW and atmospheric temperature measures in degrees Celsius [17].*

clusively for single system forecasts because networks of distributed terrestrial sky imagers are not yet available.

Temperature

The relationship in figure 2.12 shows that the power of a PV system has a medium correlation with the atmospheric temperature. This follows since temperature and irradiance are strongly correlated but the efficiency of solar PV cells reduces by about -0.4% per Kelvin increase in cell-temperature [91]. The cell-temperature in solar PV panels is governed largely by the ambient [92] temperature. However, since there is delay between an increase in ambient temperature causing an increase in cell-temperature, a time-lagged ambient temperature parameter can also be a useful factor for modelling solar PV power.

The temperature response of the PV panels to incident solar irradiance occurs over a 7-minute timeframe [36]. Therefore, under broken cloud conditions on days with large peak irradiances the efficiency of PV panels can remain higher than on comparable clear-sky days and consequently large peak power measurements can be recorded.

Wind speed

Given that the performance of PV panels is linked to temperature, it also follows that solar PV power will be correlated with wind speed since wind has the effect of lowering surface temperature. Wind speed has been shown to be a significant factor for modelling cell-temperature [93, 94]. However, wind speed has often been omitted from PV power models

because it contributes little to model accuracy. This is because it only affects PV power output for high cell temperatures and therefore most of the time it is redundant [49].

2.5.2 Review of monitoring techniques

Solar PV is the first in a new age of distributed energy resources. In Great Britain, there are 1 million solar PV systems distributed across the country. Historic monitoring methods used for tracking energy production from large-scale thermal generators are not suitable for monitoring national solar PV production. Therefore, across the world, governments, researchers, and transmission and distribution network operators have developed new methods, tools, and services for monitoring solar PV power production.

Monitoring solar PV power at a national scale involves modelling solar PV yield and then upscaling the modelled yield using data on national solar PV capacity. Section 2.4.2 covered how different countries have tackled the problem of estimating national solar PV capacity and in this section the approaches for modelling solar PV yield will be discussed.

In general, there are four approaches which can be used to model national solar PV yield: statistical sampling methods, statistical learning timeseries methods, physical methods, and ensemble methods [95].

Statistical sampling methods involve collating data for a set of reference systems for which real-time data is available and providing an adequate data source they are easy to implement. Stratified sampling methods are then be employed to select a representative subsample from the reference systems which best matches the dynamics of the system being modelled. Finally, the sample yield in each region is estimated as the mean yield for all systems in the representative sample and the PV output can be estimated by upscaling using capacity data for known systems in each region. Statistical sampling methods are highly dependent on a representative and real-time reference dataset.

Statistical learning methods are also highly dependent on historical reference data and easy to implement. They use data from reference PV systems and atmospheric observations to train statistical models which capture the correlations between a set of input (e.g. irradiance, temperature, wind speed) and output parameters (e.g. solar PV yield). They have two advantages over statistical sampling methods, they can facilitate forecasting using easily available atmospheric data and often they can be trained on historical PV and meteorological data and used to make real-time predictions without a real-time metered PV data feed.

Therefore, they do not require a real-time reference PV data feed. Under this definition machine learning and artificial intelligence approaches are considered statistical learning approaches.

Physical methods involve breaking down the attenuation processes which govern the transmission of solar irradiance through each layer of the atmosphere and then the production of electricity given the flux of irradiance incident at the surface of the solar PV panel. Physical models used to model PV power can be broken down into two parts: modelling solar irradiance using Numerical Weather Predictions, terrestrial sky-imagery, and satellite-imaging; and modelling solar PV power production using the physical principals which govern the thermal properties of the PV cell, the conversion of irradiance into DC power inside the semiconductor, and the conversion of direct to alternating current inside the inverter.

Ensemble methods involve any combination of statistical or physical approaches [95]. For example, an NWP could be used for modelling the global horizontal solar irradiance. The solar PV power could then be modelled using a statistical relationship learned from historic data from reference PV systems. Using global horizontal irradiance from the NWP as an input to estimate solar PV power in either a statistical or physical model for converting irradiance to PV power. Ensemble methods are used so that the combination of two different models can overcome the weak features of each model used independently [95].

Statistical sampling techniques, also known as upscaling, have been extensively used for modelling solar PV yield for unmetered systems using data from reference systems. Upscaling is defined as estimating information for a population by extrapolating only information from a subset of the population [64] and has become the standard technique used by grid operator across the world for estimating the national solar PV power output [64, 92, 31, 85, 84, 96].

The use of upscaling for modelling solar PV output was seen in 2010 when Shcierenbeck et al. [97] published a methodology for modelling the solar PV output in the Transnet control area in Germany. The authors used an inverse weighting method to upscale the output from a sample of reference systems to estimate the PV output of every PV system in their control area.

Following on from the work of Schierenbeck et al. [97], Golnas et al. [98] analysed 55 PV systems between $30 - 500\ kWp$ in the state of New Jersey, USA for which they had 15-minutely PV output data measured in kW. They calculated the daily yield (total output for

each system normalised to the total potential output under standard test conditions) which they then scaled by the clear-sky plane of array irradiance as calculated by the Bird clear sky irradiance model. They called this performance metric the Bird Performance Index. To estimate the performance of some unknown system they then calculated the daily weighted mean of the Bird Performance Index using data from the 54 other PV systems. Finally, they estimated PV output by multiplying by the plane of array irradiance at the location of the PV system in question for each 15-minute period throughout the day. Their model performed poorly for periods of observation of less than one day.

Lonij et al. [99] also derived an upscaling methodology for calculating solar PV yield. First, they calculated the clear sky yield as defined in Lonij et al. [100]. Using this they defined a performance metric, K, which is the PV system yield normalised to the clear-sky yield. To simulate PV power for any PV system they calculate K for the four closest reference PV systems and assume that the median value for K is representative of the PV system under prediction. Thus, upscaling the output of the reference systems to enable estimates of the output of PV systems for which no measured data is available. Using their real-time estimates of K and an estimate of the wind speed the authors then defined a solar PV forecast formula. For example, the performance of a PV system at (x, y) was calculated as $K(x - v_x dt,\ y - v_y dt)$ where v_x and v_y are the wind speed in the x and y direction respectively and dt is the forecast horizon. The authors achieved a root mean square error on their forecasted PV output of $\sim\ 10\%$ for forecasts with a 1-hour horizon time. At a 15-minute time-horizon, the RMSE was reduced to 6%. However, the authors did not detail the error in their real-time upscaling methodology.

Engerer and Mills [42] further developed the work of Golnas et al. [98] and Lonij et al. [100, 99] by defining a new parameter, the clear-sky index for photovoltaics K_{pv}. The clear-sky index for photovoltaics is defined as the ratio of instantaneous PV power output to the instantaneous theoretical clear-sky power output. The novel development of this work is the fact that the clear-sky power output has been defined purely theoretically without the need for any historic power timeseries. Thus, enabling the estimation of PV power for PV systems for which there is no historic data. Engerer and Mills demonstrated that their new metric performed better than the previously defined clear-sky indices in Golnas et al.[98] and Lonij et al. [100] when both PV system characteristics are known, and the orientation of both PV systems was similar.

Beck et al. [101] were the first to derive a specific real-time nowcasting methodology which relied only on PV power data and did not use any other meteorological information

e.g. irradiance. They created a polynomial model for solar PV power derived solely from PV power timeseries. They tested their model within the control area of the German energy provider Stadtwerke Passau GmbH (SWP). The SWP control area is 7 square kilometres and in 2014 contained 800 PV systems with at total capacity of 23 MWp. Beck et al. used data from 8 PV systems in their analysis. They found that the PV power of systems can be scaled from nearby PV plants using their polynomial approach. However, their model did not account for the geographical location of each PV system and therefore would break down when used over a larger geographically distributed system.

Engerer and Hansard [75] described a PV monitoring service for Canberra, Australia. They used data from 200 reference PV systems to model the K_{pv} performance factor first defined in Eneger and Mills [42]. They then calculated K_{pv} for unmetered PV systems using a nearest neighbour approach in which the number of neighbours is calculated on the fly by minimising an RMSE error metric. In order to model system yield and output for unmetered PV systems, they collated system metadata for all PV systems in Canberra. However, the metadata did not include information on system orientation or tilt. Therefore, they simulated system orientation and tilt from a known distribution of orientation and tilts of 500 PV systems in Canberra. Once they had fully characterised the system metadata of all 15000+ PV systems in Canberra and modelled the K_{pv} performance metric, they upscaled the modelled performance using the system descriptions for each system. They demonstrated that their PV monitoring system was able to successfully detect transient clouds. However, they did not quantify the accuracy of their regional estimates of solar PV power output.

Reference PV system data and irradiance data has been combined with some success in literature. Saint-Drennan et al. [102] used reference PV data to correct global horizontal irradiance data. Inage et al. [103] successfully used data from 45 reference PV systems to correct the aggregated satellite derived power forecast for a large 1.32km squared solar farm. However, Bright et al. [64] were the first to combine reference PV system data with irradiance data for the purpose of improving national solar PV monitoring estimates. Bright et al. compared four approaches for modelling solar PV output: a satellite only approach, an upscaling only approach, a correction approach, and an ensemble approach. They found that the hybrid approach performed only marginally better than a simple upscaling approach using only data from reference PV systems.

All the approaches described so far have relied on the upscaling of data from reference PV systems and can all be described as statistical upscaling, or statistical learning timeseries methods, or as an ensemble approach. However, physical methods can also be used to

model solar PV power. Pfenninger and Staffell [63] developed a method which relied only on satellite derived reanalysis irradiance and weather data to calculate solar PV output. First, they interpolated the global horizontal irradiance values from the EMCWF Merra-2 dataset. Then they calculated the direct and diffuse irradiance at the solar PV system using the Liu and Jordan [41] irradiance decomposition models. Finally, they estimated the solar PV output of individual systems using their own model for the thermal performance of a PV system. They made their model available online at a site named renewablesninja [104]. Whilst renewablesninja only produces historical estimates for individual PV systems, in practise it could be re-purposed to work with live and forecast irradiance data from an NWP and alongside some national solar PV metadata information it could enable a national solar PV monitoring estimate. However, this approach is limited by the accuracy of the satellite derived irradiance data which is currently 5%.

Another approach which worked independently of reference PV data is seen in Schepel et al. [76]. The authors derived a PV monitoring service for The Netherlands with no need for real-time reference PV data. They used meteorological data measured at 46 weather stations across The Netherlands to estimate PV yield using a series of statistical univariate parametric models corresponding to the different physical processes which govern solar PV production: the amount of in-plane irradiance, shading, soiling, surface albedo, system efficiency, panel temperature, model efficiency, and inverter efficiency [76]. To estimate the PV yield of a PV system in real-time the yield is modelled for the nearest weather monitoring station. The authors state that all locations in The Netherlands are less than $35\ km$ away from their nearest weather monitoring station. However, figure 2.9 shows that for hourly periods of observation PV systems in Great Britain are correlated up to distances of $\sim\ 20\ km$. The Netherlands and the UK have similar climates and therefore it is likely that 46 weather stations are insufficient for capturing the spatial variability in solar PV for hourly timeseries data in The Netherlands.

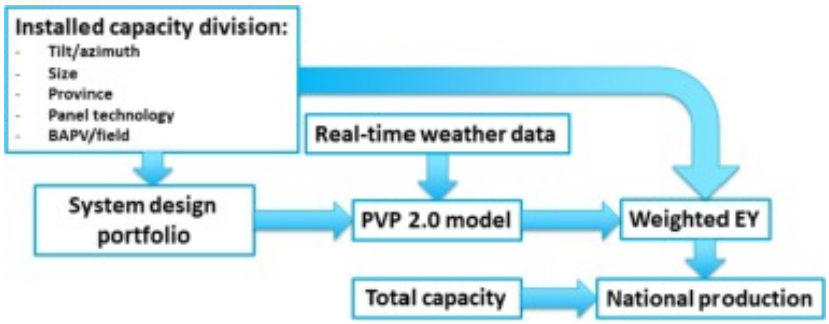


Figure 2.13: *The Dutch National Solar Energy Production model flow. EY is electricity yield (referred to in this thesis as yield) and has units* WWp^{-1}.

The authors compile one list of all PV systems in The Netherlands by combining multiple commercial, governmental, and research capacity datasets. This national capacity list is used to derive a system design portfolio. This is a portfolio of PV systems which are representative of the installed capacity of all PV systems in The Netherlands. To do this they define the six metadata parameters used in their PV yield model: installation region, PV module technology, tilt, azimuth, system type (ground mount or roof-top), and capacity in kWp. They then defined 475 PV system designs, each with unique characteristics, which represent the installed PV capacity. A detailed explanation of how each PV system design is calculated can be found in section 5 of their paper [76].

Figure 2.13 [76] shows the flow of data in the Dutch PV national solar PV output model. Once they have trained their PV yield model, sourced their real-time weather monitoring data, and created their system design portfolio for representative PV systems in The Netherlands. Next, Schepel et al. estimate the yield of each of the 475 designed PV systems. From this they calculate a representative yield for each of the distribution control areas in The Netherlands. The representative yield is then upscaled using the capacity data from their national site list to estimate the regional and national solar PV output in The Netherlands.

Schepel et al. [76] have validated their model against reference data from 26 PV systems located at different sites across The Netherlands. They found that their model has -11% to $+5.5\%$ error on the estimate of individual PV systems and that it underestimates generation 84.6% of the time. Indicating that their national PV output estimates are negatively biased. This information is available on their website and in an accompanying publication, but it is not shown on their PV output graphs or provided with their data download.

2.5.3 Review of existing PV monitoring services

Two general problems must be solved to build national scale systems to measure PV generation. The first problem is to estimate the installed PV capacity that is connected across the electricity grid in terms of the geographic location of the PV systems, their capacity, and their topological connection into the electricity network. The second problem is to calculate, for each time step of interest, the average normalised yield of the fleet systems (MW/MWp) so that it can be scaled to a regional or national power output by multiplying by capacity.

In general, engineers have used two general approaches to model national solar PV yield: using meteorological data to estimate PV yield using either a statistical or physical

modelling approach [63, 76] or using a statistical upscaling methodology to estimate the regional or national yield from a set of reference systems providing data in real-time [64, 31, 105, 101, 106].

Engineers in different countries have also assigned different licenses to their datasets. Some come with an explicit open license are therefore truly open datasets. However, some engineers have made their data publicly available without specifying an appropriate open license. Whilst these datasets are public, they are not open.

A international review of solar PV monitoring services has been performed to better understand the state of PV monitoring across multiple electricity grids. To find solar PV monitoring services multiple search techniques were used. First, the top 20 countries, ranked by total installed solar PV power were searched using a key-term search with the following terms: country name, solar, power, PV, electricity. Then the website for each countries transmission system operator was identified and searched for live electricity data.

In addition to searching for the top 20 countries by solar PV capacity, the European ENTSO-E transparency platform was used to identify European TSOs which provide solar PV monitoring data. If a TSO was identified as providing solar PV data to the ENTSO-E transparency platform then it was also included in the search and the PV monitoring service was located using the same methods detailed above.

Once a PV monitoring service had been found online, the webpage was searched for an open data license accompanying the data. If no license was found, then the broader website was searched for a copyright notice or a legal page. If no explicit open data license was provided but the data was publicly available for free download, then the services has been designated as a public non-open data source. Data being publicly available is not sufficient to make it open.

The results of the survey are presented in table 2.2. The survey identified 27 services providing data for 19 different countries. Multiple services were found for Germany (7) and the UK (3) which is an indication of the maturity of their solar PV sectors. However, there is duplication of data across many of these services. For example, all of the solar PV generation data in the UK originates from the Sheffield Solar research group and the primary data source for the German services is the four German TSOs.

There are some countries with developed solar PV sectors for which solar PV monitoring services could not be identified. For example, no service was identified for China which has 254 GW of installed solar PV capacity [107]. Other notable exceptions, which are all in-

side the top 20 countries as ranked by solar power, are: India, Vietnam, South Korea, Brazil, Turkey, South Africa, Taiwan, and Mexico. Together, these countries provide 100 GW of solar PV power. Additionally, although a PV monitoring service was identified for the US state of California which accounts for 31 GW of US solar capacity. No monitoring services were found for any other states which make up the rest of the 73 GW of solar PV capacity installed in the US. The process for finding solar PV monitoring services is definitely not without error and it is possible that these countries/states provide solar PV monitoring services but they have not been included in this survey.

Of the 29 services identified in table 1, only the British [30], Dutch [76], Australian [85], and TransnetBW [105] services provide detailed dissemination of their modelling methodology in an easy to find location on their website. Furthermore, only five [108, 109, 110, 111, 112] provide an explicit open license for their PV output data. The remaining services provide the data publicly but without the presence of an explicit open license and are therefore non-open data sources.

The GB [31] and Australian [75] services use reference data to calculate a spatially resolved estimate of the PV yield which they upscale using installed capacity data collated from government datasets to estimate national and regional PV output. The GB service does calculates the mean yield from a statistically representative sample and upscales using national capacity data provided by the National Grid ESO. Whereas, the Australian service calculates the k_{pv} performance factor for each PV system in Australia [42] from a sample of neighbouring systems and upscales to calculate the generation of each system using national capacity data derived from multiple commercial and subsidy based datasets.

The German TransnetBW [105] service uses a hybrid approach, using both solar irradiance data and data from reference PV systems to calculate a spatially resolved PV yield which they upscale using solar capacity data collected by the German government through the EEG register. The SMA Solar, Amprion, 50Hertz, and Belgian services use reference data to calculate a representative yield which they upscale using national capacity data. However, the methods used in their yield calculations are unexplained and so it is unclear whether these estimates resolve the spatial variability in PV output relating to the spatial distribution of the installed capacity of their fleet.

The Dutch PV portal [84] uses meteorological data to estimate PV yield. They have defined parametric models for the different physical processes which govern solar PV production: the amount of in-plane irradiance, shading, soiling, surface albedo, system efficiency, panel temperature, model efficiency, and inverter efficiency [76]. Using real-time meteo-

rological data and their yield model they estimate the performance of several PV systems which they have calculated to be representative of the installed capacity of the Dutch PV fleet. They validated their model against 26 sites and found that their model has between -11% to +5.5% error on the estimate of individual PV systems and that it underestimates generation 84.6% of the time. Indicating that their national PV output estimates are negatively biased. This information is available on their website and in an accompanying publication, but it is not shown on their PV output graphs or provided with their data download.

The GB and Dutch service are the only services to provide any quantitative error analysis with their model estimates. The GB service plots the 95% confidence interval for the mean of their representative sample. This is not a true representation for the error associated with their PV estimates because it does not account for the representativness of their reference data or the accuracy of their national capacity data. Whilst the dutch system performed a more robust error analysis by validating against measured PV systems, they do not provide this information on their website or in the accompanying data download. The Amprion, 50Hertz, SMA Solar, Belgian, Czechian, and Danish services provide a note acknowledging that the output of the national fleet has been calculated by upscaling data from a sample of reference systems. However, they do not specify how many systems have been used to produce this estimate or attempt to verify the accuracy of their capacity data.

The GB and Belgian services provide a note explaining that their capacity data is not error free. The GB service refers to lag between the installation of a solar PV system and it appearing on their capacity register. The Belgian service states that they can only estimate the solar PV generation for solar PV systems which they have accurate solar PV capacity data for. However, they do not provide any comment on how many solar PV systems this might be or what percentage of the total fleet they believe that they have accurate capacity information on.

Most of the services provide data with an adequate temporal resolution (< 1 hour). Most services provide data at either 15-minutely, half-hourly, or hourly resolution. A few services provide data at 1 - 5 minutely resolution. However, the Spanish service only presents daily aggregated output data.

The SMARD service and the ENTSO-E Transparency Platform collate estimates from multiple TSOs for Germany and Europe, respectively, without any explanation for how each TSO arrived at their estimate. The remaining 16 services, present solar PV generation data with no explanation of provenance.

Table 2.2: A survey, performed in April 2021, on state of the art solar PV monitoring services.

Nation	Name	Provider	Inputs	Period	License	Display	API	Method	Error analysis	Official Nat Capacity (GW)	Service Capacity (GW)	Service Max Gen (GW)
Japan [113]	Electrical Japan	Electric Power Company (Kitamoto Lab, NII)	Japanese utilities	Utility dependent	Public non-open	Timeseries	None	None	None	67 [107]	None	∼40
Germany [114]	Energy Charts	Fraunhofer ISE	TSOs	Hourly	Public non-open [115]	Timeseries	None	None	A comment on error	53.78 [107]	54.86	37.2
Germany (TenneT TSO) [116]	Actual and forecast solar energy feed-in	TenneT	Unspecified	15-min	Public non-open	Timeseries	Manual csv export	None	None	18.88 [117]	None	12.301
Germany (Amprion TSO) [118]	Photovoltaic Infeed	Amprion	Reference PV data and capacity data from EEG renew. reg.	15-min	Public non-open	Timeseries and tabulated view	Manual csv export	Comment on upscaling	Acknowledge that their estimates differ to year-end accounts in the EEG	11.89	None	7.638
Germany (TrannetBW TSO) [105]	Key Figures	TransnetBW	Reference PV data and capacity data from EEG renew. reg.	15-min	Public non-open	Timeseries	RSS feed and manual csv export	Full dissemination [105].	None	6.671 (13/04/21) [117]	None	4.4
Germany (50Hertz TSO) [119]	Photovoltaics	50Hertz	Reference PV data and capacity data from EEG renew. reg.	15-min	Public non-open	Tabulated view	Manual csv export	Comment on upscaling	None	12.973 (13/04/21) [117]	None	9.51
Germany [96]	PV Yield Produced in Germany	SMA Solar	Reference PV data and capacity data from EEG renew. reg.	15-min	Public non-open [120]	Timeseries and regional heatmap	None for free service	Comment on upscaling	They acknowledge that their reference data may not be representative.	53.783 [107]	51.080	37.4
Germany [121]	SMARD	German Federal Network Agency [122]	ENTSO-E Transparency Service [117]	15-min	CC-BY-4.0 [44]	Timeseries graph and tabulated data view.	Data can be manually exported to pdf; csv; xls; xml.	No method provided	None	53.783 [107]	50.41	33.2 (01/06/20)
Luxembourg [121]	SMARD	German Federal Network Agency [122]	ENTSO-E Transparency Service [117]	Hourly	CC-BY-4.0 [110]	Timeseries and tabulated view	Manual export to pdf/csv/xls/xml	None	None	0.195 [107]	1.85×10^{-3} [121]	0.874×10^{-3} (28/03/2020)
USA California [123]	Today's Outlook I& Supply	California ISO	Unspecified	5-min	Public non-open [46]	Timeseries	Full API and manual csv export	None	None	31.3 [124]	14.066 [125]	12.335
Italy [126]	Actual Generation	Terna	Unspecified	Hourly	Public non-open [50]	Timeseries	Manual export to xlsx/csv	None	Warning that data changes based on best available reference data	21.6 in [107]	21.08 (31/12/20)	11.8 (08/04/21)

Table 2.2: A survey, performed in April 2021, on state of the art solar PV monitoring services.

Nation	Name	Provider	Inputs	Period	License	Display	API	Method	Error analysis	Official Nat Capacity (GW)	Service Capacity (GW)	Service Max Gen (GW)
Australia [85]	APVI Solar Map	Australian PV Institute	Reference data from 6k systems and capacity data from RET database	30-min	Public non-open	Timeseries and regional heatmap	None	Full dissemination [64]	N/A	17.627 [107]	20.2	7.7
Great Britain [31]	PV Live	Sheffield Solar and National Grid ESO	Reference data from 22k systems and capacity data from TSO	30-min	Public non-open	Timeseries and regional heatmap.	Full API [127]	Full dissemination	95% confidence interval, and comment on lag in capacity data	13.653 [107]	13.08	9.68 (20/04/20)
Great Britain [128]	Electric insights	Drax	Re-publishing Sheffield Solar data	30-min	Public non-open	Timeseries	None	None	None	13.653 [107]	None	9.12 (20/04/20)
Great Britain [129]	BMRS	Elexon	Unspecified	30-min	BMRS Open Data License [108]	Tabulated view	Full API and manual export to xml/csv	None	None	13.653 [107]	13.378	9.120 (20/04/20)
Spain [130]	REData	Red Electrica Spain	Unspecified	Daily	Public non-open	Bar chart of total generation mix	Manual export to JSON/csv/xlxs/svg/png	None	None	14.089 [107]	None	N/A
France [131]	Eco2Mix	RTE France	Unspecified	15-min	CC-BY-4.0 [112]	Timeseries and regional heatmap	Full API	None	None	11.733 [107]	10.675	7.809 (29/03/21)
The Netherlands [84]	The Dutch PV Portal	TU Delft	Meteorological data from KMNI and capacity data from multiple datasets	10-min	Public non-open	Timeseries and regional heatmap	Manual csv export	Full dissemination [76]	None displayed online, [76] states -11% to +5.5% uncertainty in indiv. estimates	10.213 [107]	None	N/A
Belgium [132]	Solar Power Generation	Elia	Reference data and capacity data	15-min	Elia Open Data License [109]	Timeseries and regional heatmap	Manual xls export	Comment on upscaling	Acknowledge missing capacity data	5.646 [107]	5.64	3.86 (19/03/21)
Czech Republic [133]	All data	CEPS	Reference data and capacity data	1-min	Public non-open	Timeseries	Manual export to txt/xls/xml	Comment on upscaling	None.	2.073 [107]	None	N/A
Austria [134]	Where does the electricity come from?	Austrian Power Grid	Unspecified	Hourly	Public non-open [135].	Timeseries and animation	None	None	None	2.2 [107]	None	N/A
Poland [136]	Generation of Wind farms and Solar farms	Polskie Sieci Elektroenergetyczne	Unspecified	Hourly	Public non-open	Tabulated view	Manual export to pdf/xlsx/csv	None	None	3.936 [107]	None	N/A
Denmark [137]	Energi Data Service	Energinet	Unspecified.	5-min	Danish Public Sector Open License [111]	Tabulated view	Full API	Comment on upscaling	None.	1.3 [107]	None	N/A

Table 2.2: A survey, performed in April 2021, on state of the art solar PV monitoring services.

Nation	Name	Provider	Inputs	Period	License	Display	API	Method	Error analysis	Official Nat Capacity (GW)	Service Capacity (GW)	Service Max Gen (GW)
Portugal [138]	Actual Generation	REN	Unspecified	Hourly	Public non-open	Tabulated view	Manual xls export	None	None	1.025 [107]	None	N/A
Hungary [139]	Hungarian Power System Actual Data	Mavir	Unspecified.	15-min	Public non-open	Timeseries	Manual export to xls/xlsx	None	None	1.953 [107]	None	N/A
Slovakia [140]	Daily Operational Data	Slovakia Transmission System Operator	Unspecified	N/A	Closed	N/A	N/A	None	None	0.593 [107]	None	N/A
Estonia [141]	Elering Live	Elering	Unspecified	Hourly	Public non-open	Timeseries and tabulated view	Full API and manual csv export	None	None	0.13 [107]	None	0.158

In all cases, national output is calculated by taking the product of the modelled yield and an estimate of the installed capacity. However, only the GB, Dutch, German, and Italian services provide any explanation of the provenance of their national capacity data and Germany and Italy are the only countries to have one consolidated national capacity dataset. To estimate national capacity, the GB, Dutch, and Australian service providers must collate capacity data from multiple overlapping government and commercial capacity datasets. The Belgian service acknowledges that it can only calculate PV power for systems for which it has "detailed background information". However, they do not comment on what proportion of systems they have detailed background information on or where this data originates from.

The PV generation estimates from these 26-services feed directly into national demand forecasting tools and any uncertainty in the PV input data manifests directly as a requirement for additional backup electricity generation, having significant (multi-million Euro) financial and carbon costs. The uncertainty in PV yield can be calculated directly from the mathematical approach used in the yield modelling. However, there is far less sophistication (or research) in the approach used to estimate the uncertainty in measurements of installed capacity. The overall error in the PV generation reported by these national tools depends (in quadrature) on the uncertainty in capacity and the yield modelling error. Therefore, it is prudent to investigate, understand and minimise both error sources.

Exploring the output of these solar PV monitoring services further highlights significant differences in what each service is modelling. Some services (GB, The Netherlands, Fraunhofer ISE, SMA Solar) estimate the output for all grid connected systems including small scale domestic. Alternatively, some services estimate the output for transmission connected solar PV (Japan and California). Additionally, some services estimate the grid feed-in for solar PV systems e.g. the output minus any self-consumption. An example of this is highlighted by a 5 GW discrepancy between the maximum generation reported from the German services from SMARD, reporting grid feed-in, (32.3 GW) and Fraunhofer ISE/SMA Solar, reporting output for all systems. (37.3 GW / 37.4 GW).

Each country, region, TSO and DSO, has PV fleets with unique characteristics and access to different data sources of varying quality. All the approaches listed in table 1 are valid given the data available to the service provider and the requirements of the specific end users which their service has been designed for. However, all services lack robust estimates of the uncertainty associated with their estimates. This problem is highlighted by the ENTSO-E Transparency Platform [117] which presents solar PV output estimates from

all European TSOs (and from the UK TSO).[1] Each estimate of solar PV output is presented without uncertainty limits and without clear definition of the scope of each estimate. For example, data from the German TSOs and the GB TSO is presented side-by-side even though they are estimating different quantities with different uncertainties.

Solar PV is the first large scale example of a distributed energy resource and these monitoring services are the first in a new generation of grid services which will monitor the national and regional demand and generation from a variety of technologies. The uncertainty associated with each technology will depend on many factors such as the climate, data volume/quality and modelling method. For end users to make efficient use of each estimate robust uncertainty estimates for each monitoring service and a clear definition on the scope of the service are essential.

2.6 Evaluation of PV output estimates

For assessing the accuracy of PV output estimates numerous metrics have been introduced [49, 88]. The most basic error measurement is simply the error for any given period of observation given as:

$$\epsilon = x_{pred} - x_{meas} \tag{2.4}$$

with x_{pred} as the predicted variable and x_{meas} as the measured variable over the period of observation. This metric gives the error associated with a single model estimate, this could be the total national PV output, or it could be the output of an individual PV system. Given that PV output is mostly modelled at half-hourly resolution, presenting the errors for all half-hours at once does not make sense. Instead, there are some standard measures to assess the accuracy across many time-periods two of which are the $rmse$ and the $bias\ error$ which are given as:

$$rmse = \sqrt{\left(\frac{\sum \epsilon^2}{N}\right)} \tag{2.5}$$

[1] Following the post Brexit Trade and Cooperation Agreement (TCA), Great Britain is no longer obliged to publish data on the ENTSO-E Transparency Platform and as such all publication ceased on tht 15th of June 2021.

$$bias\ error = \frac{\sum \epsilon}{N} \tag{2.6}$$

with ϵ as the error for each period and N the total number of periods under consideration. The bias error is a good measure of systematic uncertainties in model estimates. A positive bias error indicates that the model estimates are too large and a negative bias error indicates that the model estimates are too small. The quadratic relationship in the rmse means that large forecast errors have a larger impact on the rmse error metric than small forecast errors. This property of the rmse metric can mean that it is useful for reflecting the importance of large errors for stable grid operation. However, it also amplifies model errors on days with larger volumes of PV output and can make comparison between different months/regions difficult. One way to overcome this issue and facilitate unbiased comparison between different periods with different volumes of PV output is to normalise the error metric in equation 2.4 prior to calculating the rmse or the bias error. For example:

$$normalised\ error = \frac{x_{pred} - x_{meas}}{x_{norm}} \tag{2.7}$$

with x_{pred} as the predicted variable, x_{meas} as the measured variable, and x_{norm} as the normalisation variable. Some example normalisation variables are x_{meas}, the range of x_{pred}, the max of x_{pred}, or some modelled variable which accounts for the system-to-system variation such as the clear-sky PV power K_{pv} as defined in Engerer and Mills [10].

Another method for assessing the accuracy of model estimates for solar PV output is to consider the correlation and goodness of fit between the estimated and measured solar PV output. For example, the coefficient of determination (R2) and the Pearsons Correlation Coefficient which are given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{2.8}$$

with ϵ_i as the error residual defined in equation 2.4, x_i as the predicted variable, and $\overline{x}$ as the mean of the predicted variables.

$$\rho_{x,y} = \frac{\sigma(x,y)}{\sigma_x \sigma_y} \tag{2.9}$$

with $\sigma(x,y)$ as the covariance between the measured and predicted variables, σ_x is

the standard deviation for the measured variable and σ_y is the standard deviation for the predicted variable.

Chapter 3

Accuracy of modelling PV yield

What you do makes a difference, and you have to decide what kind of difference you want to make.

– Dr Jane Goodall

3.1 Introduction

As discussed in the introduction the GB solar PV monitoring service, PV Live, can be broken down into three parts as illustrated in figure 3.1: 1) modelling the PV yield, 2) scaling the modelled yield by capacity to estimate the output for each PV system in GB, 3) assigning the output of individual PV systems to nodes on the electricity network. This thesis is concerned with estimating the uncertainty in the *national* solar PV output estimates. Therefore, it is concerned with estimating the sample bias error, the statistical model error, the capacity error, and their total error when combined together to estimate the national solar PV output. Since there is no ambiguity about whether a system is located in Great Britain, this thesis is not concerned with estimating the spatial error which is affected by both the accuracy of the geographic boundaries for the real-time area served by each Grid Supply Point, and the accuracy of the location of each PV system.

Additionally, as discussed in section **??**, the national PV yield and output modelling methodology predates the start of this research. Therefore, it is important to note that the PV Live method, whilst detailed in this chapter for completeness, is not a contribution of this thesis. To this end, whenever "we" is used in the following chapter it is used to denote work done by the Sheffield Solar research group which does not relate to the contribution of this thesis.

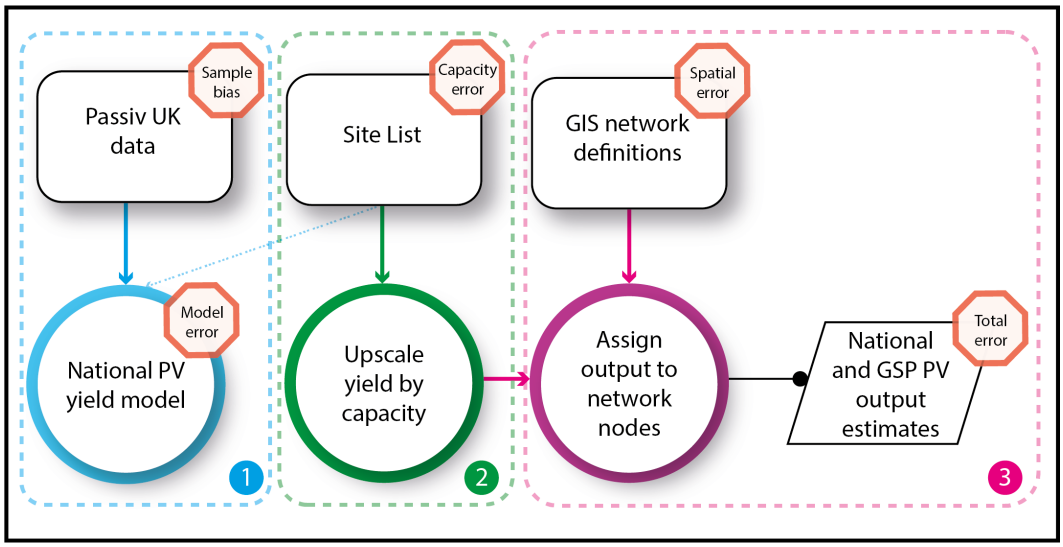


Figure 3.1: *The modelling approach of the national and regional GB solar PV monitoring service, PV Live.*

In this chapter, the PV yield model will be analysed to understand the total uncertainty in the estimates of the national PV yield. The PV yield model uses an upscaling methodol-

ogy which is the standard approach in the energy industry where where sufficient data is available from reference PV plants [106, 102, 97, 64, 75]. For example, in 2011 the German federal network agency recommended that the German TSOs use an upscaling approach for estimating the solar PV output of their fleets [106] and the GB solar PV monitoring service has been using an upscaling approach since 2015 [30]. There are examples in the literature of attempts to estimate the the error and uncertainty in national solar PV output estimates [64, 106, 102, 76]. Some researchers have focused on assessing the error for estimating the output of individual systems [76, 64]. However, as discussed by Saint-Drenan [106] the total error on the estimation of the average performance of a fleet of PV systems is governed by central limit theorem and therefore, it is more appropriate to assess the error on the total estimate of the national yield.

As discussed in the section 2.1, there is no ground-truth data on national solar PV output. Therefore, assessing the error on the national yield estimates is not straightforward. Figure 3.1, identifies the two main sources of error associated with estimating the yield in an upscaling approach: the *statistical error* resulting from the number of systems in the sample and their temporal and spatial distribution; and the *sample bias error* which is whether the sample is statistically representative of the wider population. These two sources of uncertainty will be investigated in this chapter.

This chapter is structured as follows. First, in section 3.2 the GB PV yield model methodology is presented. Then in section 3.3, a case study on the accuracy of the intraday PV estimates is presented. In this analysis estimates from the intraday PV Live model are compared with the more accurate historic estimates. This analysis highlights the error contribution from the real-time accuracy of the capacity data and the size of the sample of systems. The accuracy of the national capacity data is investigated in more detail in chapter 4. Section 3.4 investigates the accuracy of the PV Live national PV yield estimates with respect to the number of systems in the sample. Finally, in section 3.5 the accuracy of the PV yield estimates are tested against reference data from a set of solar farms to better understand the sample bias error.

3.2 PV Live Yield Model Method

Upscaling PV power from a sample of reference PV systems to estimate the power of some unmetered PV systems is common in literature [61, 102, 106, 97, 98]. Most of these methods involve approximating the output of an unmetered PV system with that of its nearest neigh-

bours. In this paper, an upscaling approach is employed which makes use of the location and capacity of a sample of PV systems along with the full site list of the GB fleet.

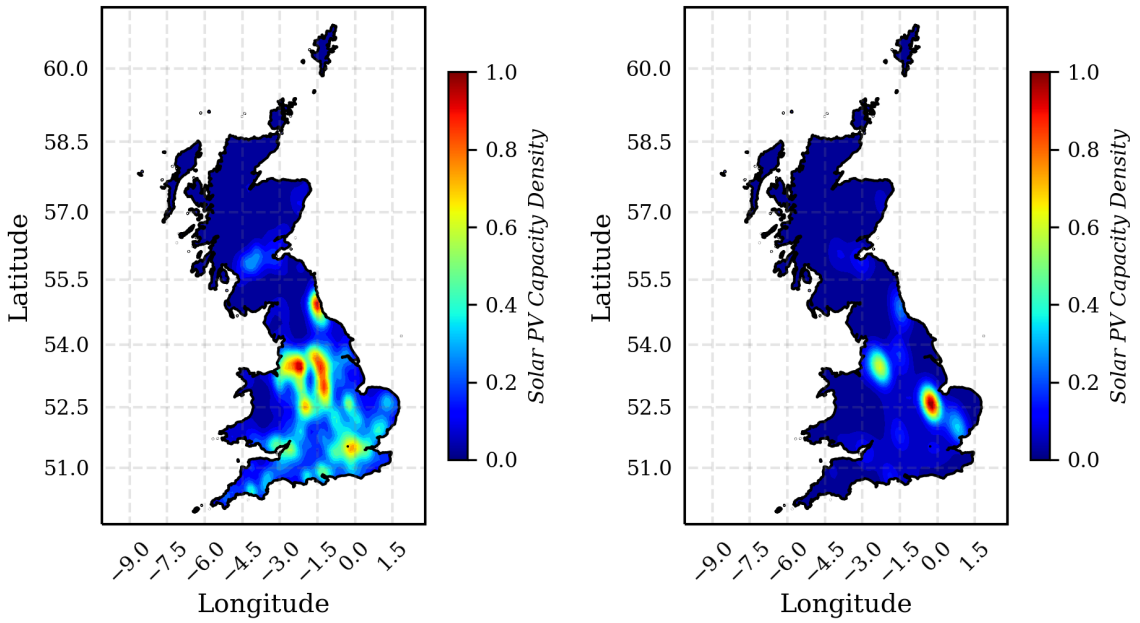
The available sample includes 20,000 PV systems with location, capacity and half hourly time series energy generation and has been made available for this study through National Grid ESO and Passiv UK Ltd. These systems are distributed unevenly across Great Britain as shown in the spatial density plot figure 3.2b.

In order to achieve a spatially representative subsample of PV systems from which to calculate the representative PV yield for GB, we have devised an optimisation algorithm. The algorithm seeks to select a subsample of PV systems from the available sample such that the subsample's spatial distribution matches as closely as possible the geographical distribution of all PV capacity. To do this we optimise the cost function given in equation 3.1. Where $X_{i,j}$ is the value of the cost function for each cell, $C_{i,j}$ is the capacity of the GB PV fleet in each cell, C_T is the total GB PV capacity, $S_{i,j}$ is the subsample capacity, and S_T is the total capacity of the subsample.

$$X_{i,j} = \left(\left(\frac{C_{i,j}}{C_T} \right) - \left(\frac{S_{i,j}}{S_T} \right) \right) \tag{3.1}$$

To select the spatially representative subsample from the 20,000 Passiv systems, we divide GB into a grid of N cells of 3° lon and 1.25° lat, with an origin at -9° lon and 49.9° lat. To start with there are no systems selected in the representative subsample and all of the systems in the Passiv sample are available to be selected. The aim of the algorithm is to select a representative sample such that the ratio of the sample capacity in each grid cell to the total sample capacity matches as close as possible the ratio of the total capacity of the GB PV fleet in each grid cell to the total GB PV fleet capacity.

The algorithm loops over every cell in GB and calculates, for each cell, the value of the cost function (equation 3.1) arising from adding the smallest system by capacity from the available Passiv sample located in the cell to the representative subsample. One system is then added to the representative subsample from the cell which minimised the value of the cost function. This system is also removed from the sample of Passiv systems which are available to be selected. The process is then repeated N_passiv times, and each time one more system is added to the representative sample and a new value of the cost function is calculated. Finally, the most representative sample is selected as the number of sample systems which minimised the cost function.



(a) *Site List Number Density*

(b) *Sample Number Density*

(c) *Site List Capacity Density*

(d) *Sample Capacity Density*

Figure 3.2: *Spatial density maps of the GB solar PV feet and the sample of $\sim 22k$ reference systems which are available for this analysis. a) The non-parametric number density of the GB fleet. 1b) and the sample 1c) The capacity-weighted non-parametric density of the GB fleet d) and the sample.*

$$Y = \frac{\sum_{i=1}^{N} \frac{G_i}{C^*_i}}{N} \tag{3.2}$$

$$C^* = C\left(1 - \frac{y \times D}{100}\right) \tag{3.3}$$

$$G = \sum_{j=1}^{N_{fleet}} G_j = Y \times \sum_{j=1}^{N_{fleet}} C^*_j \tag{3.4}$$

$$SE = \frac{\sigma_s}{\sqrt{N}} \tag{3.5}$$

The representative sample is then used to calculate the representative yield of the GB PV fleet using equation 3.2. Where Y is the representative yield (kWh / kWp), N is the number of systems in the representative sample, G_i is the generation of system i in kWh, C^*_i is the effective capacity of system i in kWp, as defined by equation 3.3. In equation 3.3, C is the installed capacity of a system in kWp, D is the percentage degradation of installed capacity per year, and y is the time in fractional years since installation. The effective capacity accounts for the degradation in performance of a PV system over time by capturing variation in system performance caused by system age. This essentially applies a correction for the difference in ages between the sample systems and the GB fleet. An average degradation rate of 0.8±0.1% per year, as previously measured by Taylor et al. [73], is used.

In equation 3.4, the representative yield is then scaled by the total capacity in the site list to give the national generation output G (kWh). For any individual system, the estimate G_i will not be an accurate estimate of its generation. However, due to the central limit theory [142] the average performance of the sample will converge on the average performance of the national PV fleet providing a large enough sample is achieved.

The modelled national PV yield is then scaled to national PV generation using the cumulative effective capacity from the site list. Finally, the standard error on the yield is calculated using equation 3.5, where σ_s is the standard deviation of the representative sample and N is the number of systems in the representative sample.

3.3 Case Study: Accuracy of the real-time PV output estimates

As described in section 2.2 there are two streams of PV power output which are made available through Natioanl Grid ESO to the Sheffield Solar team for modelling PV generation in Great Britain. The first is an intraday dataset providing 2-minutely instantaneous PV power output data for ~ 1000 systems in very-close-to real-time. Additionally, there is a historic dataset of half-hourly energy readings for $\sim 22k$ systems which is made available at 9am on the following day.

The national PV power model is simulated at least three times: once at the end of every settlement period using the intraday sample, at 11am the day afterwards using the larger day-plus-one sample, and again when there is a capacity update provided by National Grid ESO. In this section the accuracy of the intraday PV power estimates will be analysed.

There are three important factors which affect the accuracy of the operational intraday PV power estimates:

- *Lag in reported capacity*: The intraday capacity estimate is missing capacity from some PV systems due to the lag between installation and reporting in National Grid ESO's capacity register.
- *Sample size*: The day-plus-one data set contains $\sim 22k$ systems whereas the intraday contains $\sim 1k$.
- *Different metering methods*: The intraday input data contains instantaneous PV power output readings whereas the day-plus-one input data contains readings for the half-hourly energy generated.

In this section, the accuracy of the intraday PV power model will be investigated with respect to these three factors.

3.3.1 Method

A case study was performed in September 2019 to investigate sources of error in the intraday PV power model. The period of the study was March 2016 – August 2018. This period

was selected to ensure that all PV capacity installed within the study period had made it onto the most recent PV capacity update. The PV power model has been simulated four times to allow for a ceteris paribus comparison between the intraday and day-plus-one PV yield estimates. Each simulation of the PV power model will now be described.

The PV power model was simulated historically using the day-plus-one sample data and the retrospectively updated PV capacity data. These estimates provide the "actual" value for PV generation used in this analysis. This simulation used all available PV generation data from roughly 22k PV systems distributed geographically across Great Britain and the retrospectively updated PV capacity estimate. Therefore, this is the best available estimate for the half-hourly PV power output in Great Britain.

The PV power model was simulated in real-time using the intraday PV power sample data and the operational estimate of PV capacity. This estimate is the real-time view of PV generation which National Grid ESO use in their role as System Balancer. This estimate suffers from all three factors affecting the intraday PV power estimates: lag in reported capacity, smaller sample size, and the instantaneous-power metering method.

To determine the impact of each factor, the PV power model was simulated twice more to investigate in isolation the effect of the lag in reported capacity and then the metering method. The PV power model was simulated using the intraday PV power sample data and the retrospectively updated PV capacity data. This experiment controlled for the lag in PV capacity. Finally, the PV power model was simulated using a subset of the day-plus-one PV power such that systems were only included if they also provided data intraday and the retrospectively updated PV capacity data. This analysis controlled for both the lag in capacity and the different metering methods.

The results of these four experiments were subset between the hours of 9am and 3pm to remove low sunrise and sunset readings which are clustered around zero and would skew our results. The day-plus-one with retrospective capacity was used as our dependent variable and plotted on the x-axis and each of the three response variables (intraday, intraday with control for capacity lag, and intraday with control for capacity lag and metering method) were plotted on the y-axis. Ordinary Least Squared Regression has been used to plot a line of fit to each graph and for each plot the R Squared and the RMSE were calculated.

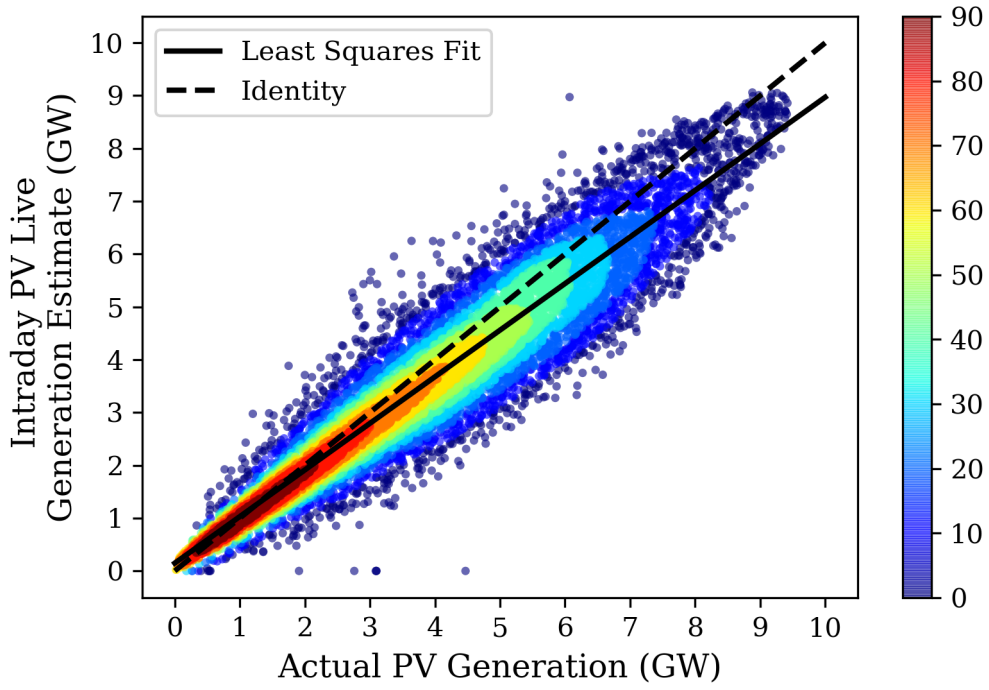
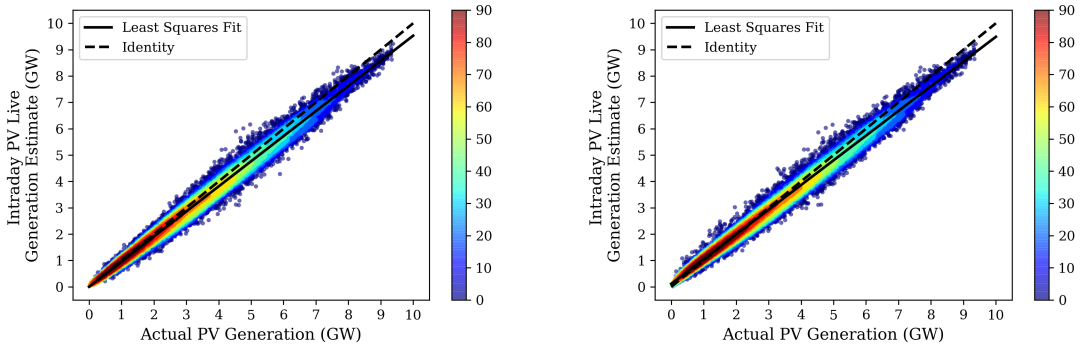


Figure 3.3: *Half-hourly estimates for the hours of 9am to 3pm for data between March 2016 and August 2018 of the operational intraday PV power model against the day-plus-one PV yield power model.*



(a) *Capacity controlled* ***(b)*** *Metering method controlled*

Figure 3.4: *Bivariate fits comparing the results from the GB PV output model estimates when caluclated using the intraday sample of* $\sim$ 1000 *systems with the full historic analysis calcuated using the most up-to-date capacity information and a sample of* $\sim$ 20000 *systems. In figure a) the lag in reported capacity has been controlled and in b) both the lag in capacity and the different metering methods between the intraday and the day-plus-one samples have been controlled for.*

3.3.2 Results

Figure 3.3 shows the relationship between real-time and historic PV output estimates. The line of fit in this plot has an R Squared of 0.9 and a Root Mean Square Error of $640\ MW$. Indicating that the real-time model can explain 90% of the variance in PV output and is accurate to within $640\ MW$. The identity line is plotted on figure 3.3 as a dashed black line and shows the expected line of fit of the model and the Ordinary Least Squares fit is plotted with a solid black line. The line of fit falls below the identity line when PV generation exceeds $\sim 1.75\ GW$. Suggesting that the intraday model estimates are negatively biased.

Figure 3.4 shows two bivariate plots for the results from the two simulations of the intraday model with control variables. In figure 3.4a, control for the lag in PV capacity has improved the R Squared from 0.9 to 0.98 and reduced the Root Mean Square Error from 640 MW to 295 MW. In figure 3.4b, additionally controlling for the metering method did not change the R Squared to 2 significant figures and insignificantly reduced the Root Mean Square to 293 MW. Both plots indicate that for periods with large PV output the intraday model underestimates the actual PV generation. This is the same result as seen in figure 3.3, although here the relationship is weaker.

Figure 3.5, shows the daily-range-normalised bias error for the yield estimates computed with both control variables. In this plot, the bias error has been normalised by the daily range of the historic PV outputs and then binned by the historic PV output in bins of width

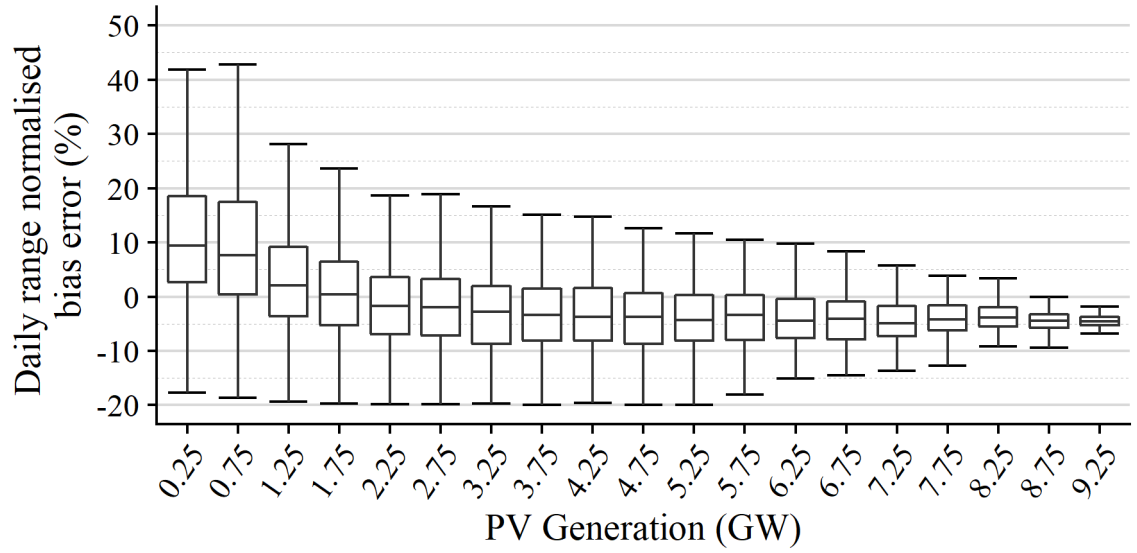


Figure 3.5: *The daily-range-normalised bias error.*

0.5. This figure shows that when the PV power output is less than 1.75 GW there is a positive bias in the real-time yield estimates and when PV power output exceeds 4 GW there is a negative bias error of $\sim -4\%$ in the real-time yield estimates.

3.3.3 Discussion

This study investigates the accuracy of the estimates from the real-time PV output model. There are several differences between the real-time estimates and the more accurate historic simulations: namely, the capacity data, the metering methodology, and the sample size. In this study I analysed the error in the real-time estimates by controlling for the capacity data and the metering method. The real-time estimates were shown to have an RMSE of $640\ MW$ and controlling for the lag in capacity reporting reduced the RMSE to $293\ MW$. Additionally controlling for the metering method had a negligible effect on the model error. Therefore, the source of this remaining error must be the smaller sample size used in the real-time model compared with the historic model.

Lag in capacity reporting is responsible for half of the error associated with the intraday PV model. This capacity error is caused by a lag between a PV system being installed and its capacity being included in the PV capacity register. Lag in PV capacity data exists to some extent in all PV monitoring services in table 2.2 since it is impossible to know exactly when a PV system first starts generating. Both the GB and the Belgian national PV monitoring services provide some comment about an error introduced in their data due to a lag in

reporting PV capacity. However, this is the first time that the error due to lag in reporting has been quantified in terms of effect on the accuracy for the national PV output data. The method for compiling the national site list for Great Britain, as described in section 2.4.2, involved simulating an uplift to the site list to correct for the lag in capacity reporting and then spatially allocating any unreported capacity according to the geographic density of the existing PV fleet. However, forecasting PV installations is difficult because it depends on many factors which are hard to predict such as the price of solar PV panels and subsidy. This result highlights the need for improved PV capacity registers which facilitate faster data collection.

The need for an accurate capacity register was first identified in Germany in 2006 by Felten et al. [143], when they performed a survey of the German PV industry and showed that ad-hoc capacity accounting resulted a 30% under reporting of installed PV capacity. More recently, in 2020, Stowell et al. [144] used a more sophisticated computer vision technique coupled with the OpenStreetMap data and community to map European PV capacity. Their results suggest that UK Governments estimate of installed PV capacity could be under reported by as much as 16% (2.6 GW). These studies have focussed on the total capacity error and have ignored any temporal effects on the capacity error. This work goes beyond previous results to quantify the error in a national PV power model caused by the lag between the installation and reporting of solar PV capacity. Highlighting the need for timely reporting of PV capacity; best practise would be a system register which tracks PV capacity of planned PV systems along with a prediction of the first-generation date and has a mechanism for updating individual system data to keep track of changes to system capacity as they age. For example, in Italy, the GAUDI system [80] is one centralised database which tracks energy systems from planning through to decommissioning.

The small sample size used to produce the real-time estimates is responsible for the other half of the error in the real-time PV output estimates. The sample size error causes the model underestimate the PV output by $\sim -4\%$ when the total PV generation is greater than $4\,GW$. The dependence of this error on the sample size is well documented in the literature [106, 102, 64]. Central limit theorem dictates that larger sample sizes will have lower standard errors for estimating the population mean [142] and in literature larger sample sizes have been shown to lead to smaller errors [106]. In particular, Hoff and Perez [52] showed that the variability in the PV power output of a sample of PV systems is inversely proportional the number of systems in the sample. This result follows on from the results in figure 2.9 which show that in Great Britain PV systems are correlated over distances of 10km. Therefore, for Great Britain, $\sim 20,000$ sample PV systems are needed to fully

capture the spatial variability in PV output.

The sample size error implies that for the smaller sample size the PV yield model is not sampling enough high-yield PV systems. I believe that this is caused by the chi-squared optimisation algorithm getting stuck in a local minimum because the solver runs out of systems to sample in one of the southern grid squares which has a large volume of installed solar PV capacity. If the solver runs out of systems to sample in one grid square it might be able to marginally improve the global sum of the chi-squared metrics for all grid squares by adding sample systems to grid squares in the north where there are unselected sample systems. However, eventually this will result in too many northern systems and continuing to add systems will only result in a less representative sample. The critical research question for solar PV monitoring is: "How many PV systems are needed to model a given national PV fleet?". This research question is investigated in section 3.4.

The operational intraday PV power estimates have been shown to be accurate to within $640\ MW$. The cost of this error in terms of System Balancing is given in [145] which states that a $10\ MW$ error in the national demand forecast costs £20 million in system balancing costs. The error in the intraday PV model will not directly correlate with an error in the national demand forecast because the more accurate historic PV power estimates are used for forecast training and because there are many other components which make up the national demand forecast. However, it is likely that errors in the intraday solar PV monitoring service create system balancing costs somewhere between hundreds of thousands and millions of pounds per year.

3.3.4 Conclusion

In this section, the real-time PV output estimates have been investigated as a case study to understand the different drivers of error in solar PV monitoring. The real-time PV power estimates are accurate to within $640\ MW$ and when the PV output is larger than $4\ GW$ there is a -4% bias error which is caused by under-sampling high yield southern PV systems. Improvements to reduce the lag between installation and reporting of solar PV capacity can reduce the RMSE to $295\ MW$ and a larger sample could reduce this error by a further $293\ MW$.

3.4 Statistical model error

Figure 3.6 shows the modelling approach for the national GB PV output model. In the remainder of this chapter, the error associated with the first part of this process, as highlighted in figure 3.6, will be analysed. Frist, in this section, the statistical model error associated with the sample size used in the model will be investigated. Preliminary analysis in the previous chapter identified a 293 MW RMSE in the real-time PV output estimates caused by the sample size. In this chapter the full relationship between the error of the yield model and the sample size will be explored. Then in section 3.5, the sample bias error will be explored so that the total error in the yield estimates is understood.

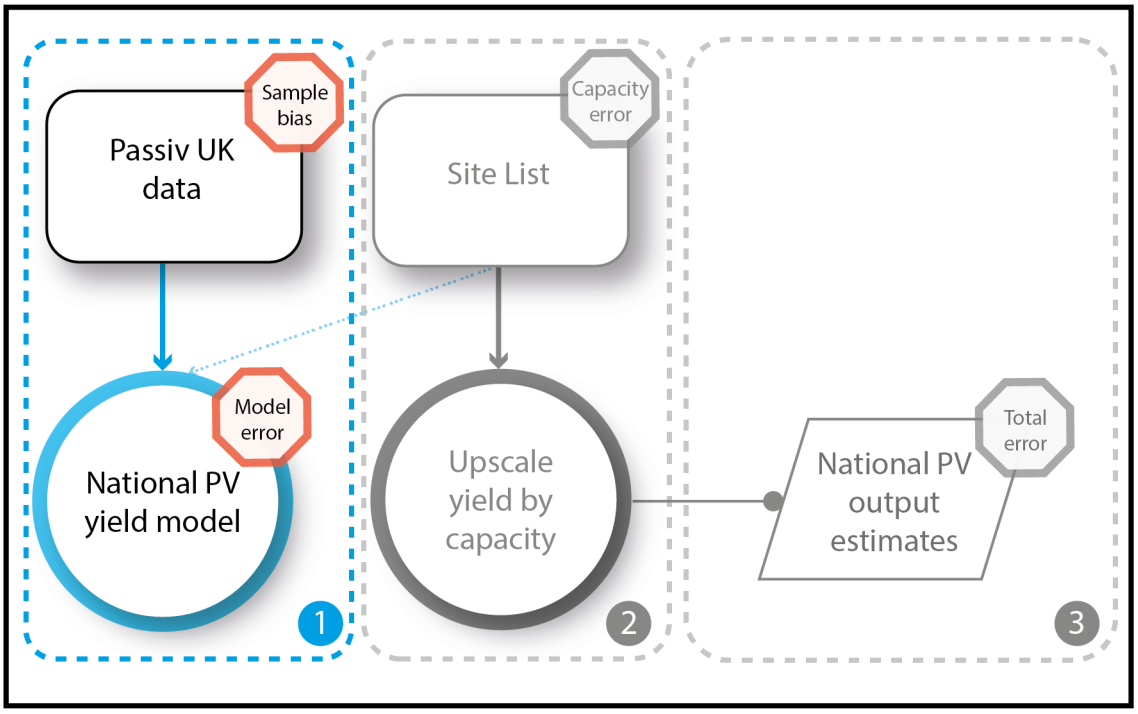


Figure 3.6: *The modelling approach of the national GB solar PV monitoring service, PV Live.*

No country facilitates comprehensive solar PV metering. Therefore, no precise data exists on the solar PV output of a national PV fleet. This means that perfectly quantifying the error on an estimate of national solar PV yield or output is impossible. Nevertheless, there are examples in the literature of attempts to quantify the error in a national yield estimate. Engerer and Hansard [75] used a transient cloud to demonstrate that their regional PV yield model was able to capture the varaibility in PV power output caused by the passing cloud. Schepel et al. [76] investigated the error in the estimation of individual systems and found that their model was able to estimate the output of 26 individual PV systems to between -11% to +5.5%. Schierenbeck [97] demonstrated that the average RMSE for single plants

ranges between 6% and 7%. However, they also showed that the error for a fleet of 140 plants is around 0.9-1.1%. A similar result was reported by Saint-Drenan [106] where he showed that the reduction in error for a fleet of PV systems follows central limit theorem and he demonstrated that for one TSO area in Germany the error in modelling solar PV yield using a sample of 311 PV systems was $\sim 1.5\%$.

One approach for quantifying the national yield error is leave one out cross-validation. Leave one out cross validation involves computing the model using N-1 times, where N is the number of reference systems used in the model, and for each computation the modelled yield is compared with the left-out system. This approach gives the most complete assessment of the model error for any individual period and minimises measurement error introduced by removing part of the sample for testing. However, numerous sample sizes (N) up to 20000 need to be tested in order to understand the full relationship between sample size and error. Additionally, a full year of data is needed to capture the inter-annual variation in solar PV output. Therefore, this approach is prohibitively computationally expensive.

The approach used to estimate the yield error in this analysis is similar to the approach used by Bright et al. [64]. In their analysis, Bright et al. randomly selected a set of reference PV systems and used an upscaling methodology to estimate the output of a randomly selected set of target PV systems. However, the PV systems in Bright et al. were all located in a small region and therefore, their analysis ignores the spatial distribution of the PV systems. Whereas, this analysis is considering the error in the GB PV fleet and therefore it is not reasonable to assume that the PV systems are evenly distributed. Hence, in this analysis, a randomly selected set of reference PV systems will be used to compute the GB PV yield and this yield will be compared with the GB PV yield as computed with the full available sample.

3.4.1 Method

A sensitivity analysis has been performed in order to understand the influence of the sample size on the statistical error in the yield model. From central limit theorem, it is expected that as the sample size increases, the the standard error in the measurement of the population mean will fall proportionally to the inverse square root of the number of systems in the sample. In this analysis, the PV yield model was computed for the year 2017 with half hourly resolution using the full available sample of 22,000 systems. These estimates are then

compared with further simulations of the PV yield model made with smaller sample sizes between 50 and 15,000. Each sample size is selected using random sampling with replacement. For each sample size and for each half hour, the yield was computed N-times where N is 10,000; 1000; 500; 100 for samples sized between 50 and 900; 1000–6000; 7000–12,000; and 13,000–15,000. The number of repeats for each sample size had to change because computing many repeats for the larger sample sizes was computationally prohibitive because the run time of the algorithm for selecting the representative sample increases as $\mathcal{O}(n^2)$. However, 10,000 repeats is unnecessary for the larger sample sizes because the standard error on the mean reduces with the square root of N. The results of the Monte Carlo analysis are compared against the PV yield estimates produced by the model when it is simulated using the full sample of ~ 20000 systems.

Test sample size (000's)	Number of repeats
0.05 - 0.9	10,000
1 - 6	1,000
7 - 12	500
13 - 15	100

Table 3.1: The number of simulations performed for each sample size.

3.4.2 Results

The results in this study have been confined to 10am – 2pm because the low yield readings around sunrise and sunset will skew the analysis and the hours around midday are most important since this is when PV yield is maximised. The mean mean yield across all random samples for each period has been computed for all sample sizes as per equation 3.6. Where, n is the sample size, t is the half-hour period of observation, N is the number of model repeats for each sample size as given by table 3.1, $\overline{Y}_{n,t}$ is the average yield across all N repeats, for each sample size n at time t, and $Y_{n,i,t}$ is the yield for each of the N computations of the model with n randomly selected reference sites at time t.

The mean yield as calculated by equation 3.6 gives an estimate of the average performance for each sample size n at time t. This estimator accounts for the spatial distribution of each random sample, the spatial distribution of the national PV fleet, and the spatial variability of the weather across the GB PV fleet and each random sample. In this analysis, this estimator is used as a measure of the likely yield computed using n reference systems at time t.

For each sample size n and for each period t, the bias error is calculated as per equation 3.7, where $Y_{n_{tot},t}$ is the yield at time t computed using the full available sample. The absolute percentage error is then calculated for sample size n and for each time t as per equation 3.8, where $\epsilon_{n,t}$ is the bias error as calculated in equation 3.7 and $Y_{n_{tot},t}$ is the yield computed using the full available sample at time t.

$$\overline{Y}_{n,t} = \frac{\sum_{i=1}^{N} Y_{n,i,t}}{N} \tag{3.6}$$

$$\epsilon_{n,t} = \overline{Y}_{n,t} - Y_{n_{tot},t} \tag{3.7}$$

$$APE_{n,t} = \frac{1}{N} \sum_{t=1}^{N} \left[\frac{\epsilon_t}{Y_{n_{tot},t}} \right] \times 100 \tag{3.8}$$

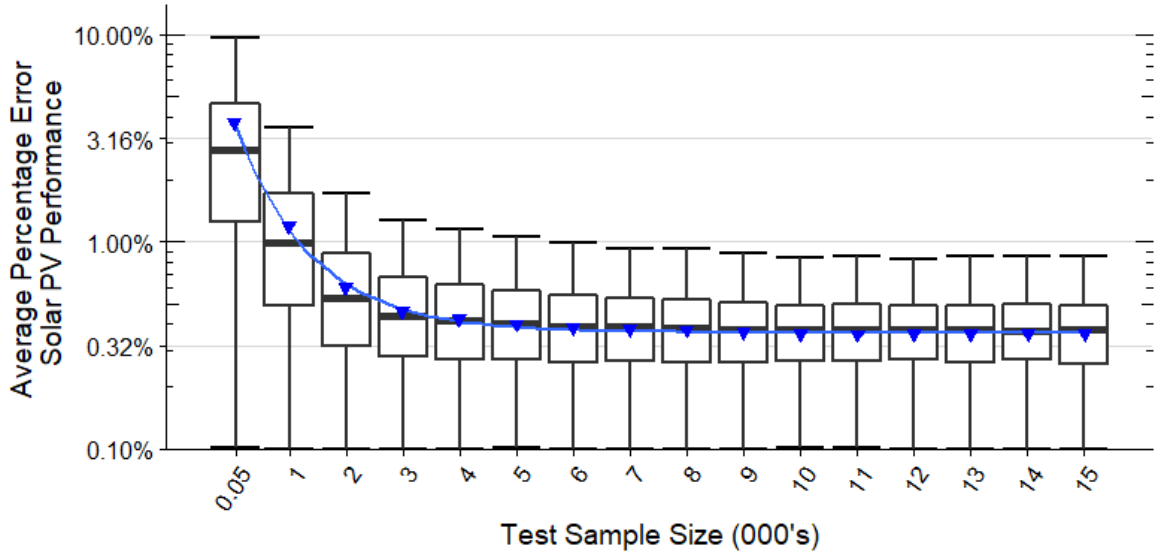


Figure 3.7: *The relationship between sample size and model error for the national PV yield model for the year 2017.*

In figure 3.7, boxplots of the absolute percentage error, calculated as per equation 3.8, for 2017 have been plotted for each sample size. A log scale has been used on the y-axis so that the trend for larger sample sizes is easier to see. As the sample size increases the absolute value of the relative bias error and its variance decreases. For sample sizes ≥ 6000 the absolute model error plateaus. For these sample sizes the upper quartile is around 0.6%. Implying that 75% of the time the PV yield model operates with an absolute model error of $\leq 0.6\%$. Furthermore, the max whiskers of the boxplots are $\leq 1\%$, suggesting that if a

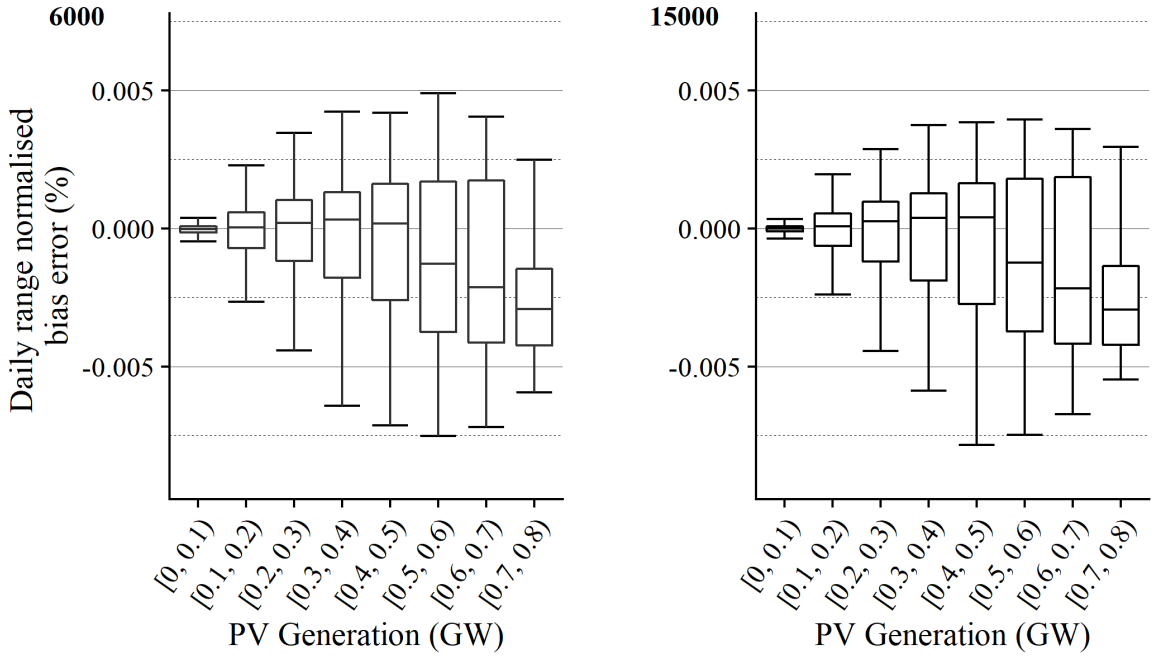


Figure 3.8: *National PV yield model bias error as a function of actual PV yield for a sample size of 6000 and 15000.*

sample size of 6000 systems is achieved then the absolute model error is $\leq 1\%$.

The bias error results have been binned according to $Y_{n_{tot},t}$ in bins of width 0.1. In figure 3.8, box plots have been plotted for each bin for sample sizes of 6000 and 15000. These sample sizes have been chosen because they highlight the change in bias error as the model plateaus. Figure 7 shows that the PV yield model is unbiased when the PV yield is less than 0.5 and negatively biased for PV yields greater than 0.5. Indicating that although the overall model error plateaus for sample sizes greater than 6000 (as shown in figure 6) the model error for high yield periods is still present.

Figure 3.9 is a heatmap of the yield bias error containing periods for which the actual yield fell between 0.5 and 0.6. The smaller sample sizes contain the largest variation in the yield bias error. When the sample size ≤ 1000, there are red squares in January, February, June, and December. Indicating that the PV yield model is overestimating the PV yield for these months. For the rest of the months the model generally underestimates PV yield as indicated by the blue squares in the heatmap. Once a sample size of 2000 systems is achieved the bias error remains relatively constant across all months in the year. However, for these months the model tends to overestimate yield for the latter months in the year.

Figure 3.10 is a heatmap of the yield bias error containing only periods when the yield fell between 0.6 and 0.7. As in figure 3.9, the smaller sample sizes contain the largest vari-

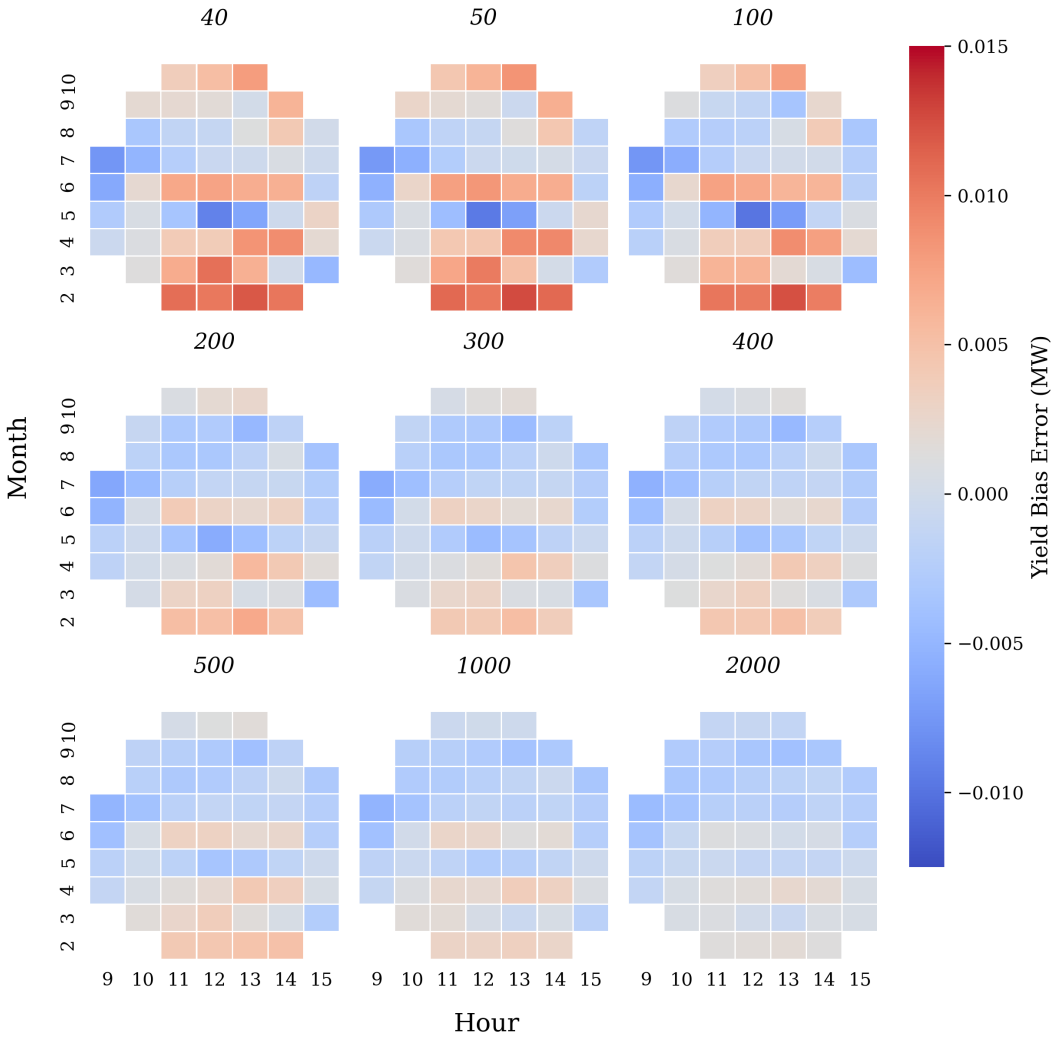


Figure 3.9: *Heatmaps of the bias error for periods when the actual PV yield fell between 0.5 and 0.6. One heatmap is shown for every sample size tested and missing values are shown in white.*

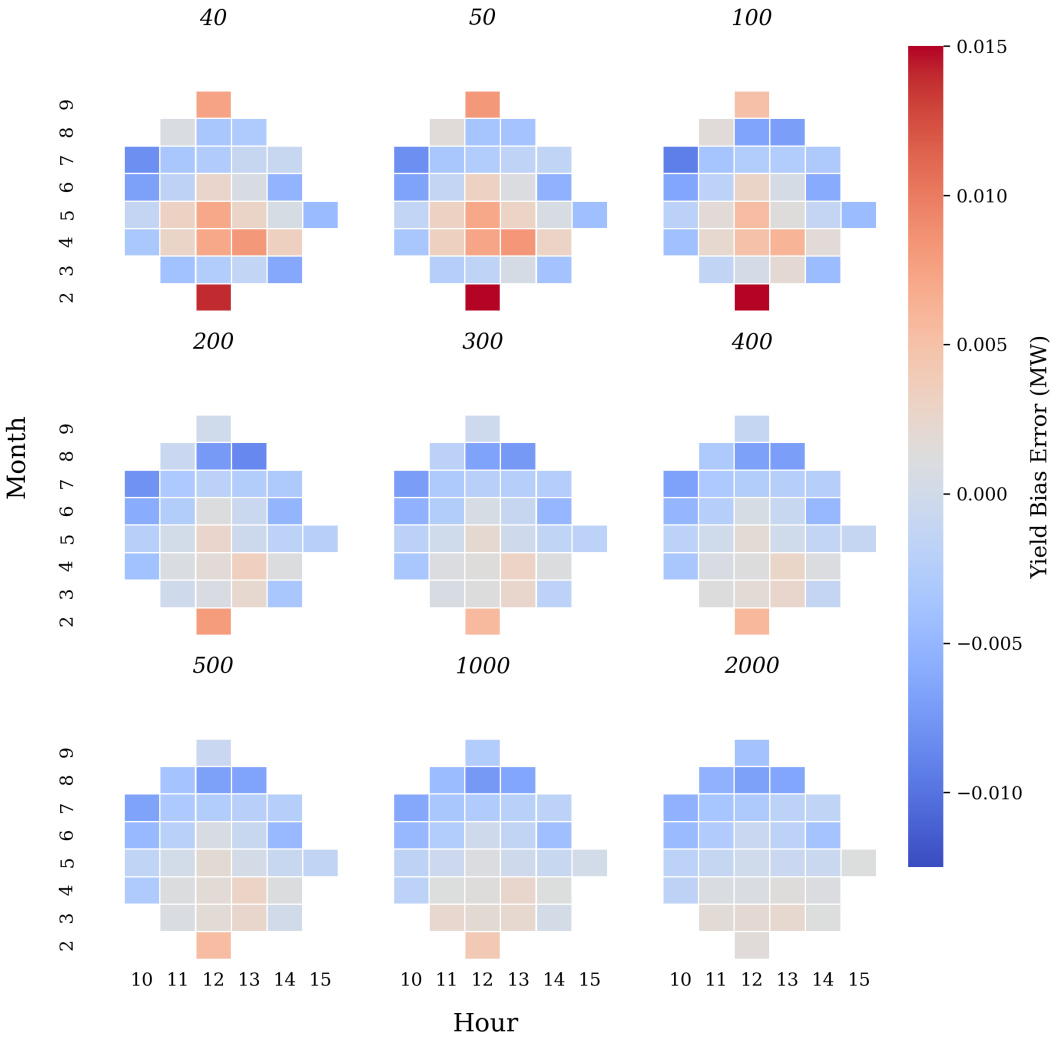


Figure 3.10: *Heatmaps of the bias error for periods when the actual PV yield fell between 0.6 and 0.7. One heatmap is shown for every sample size tested and missing values are shown in white.*

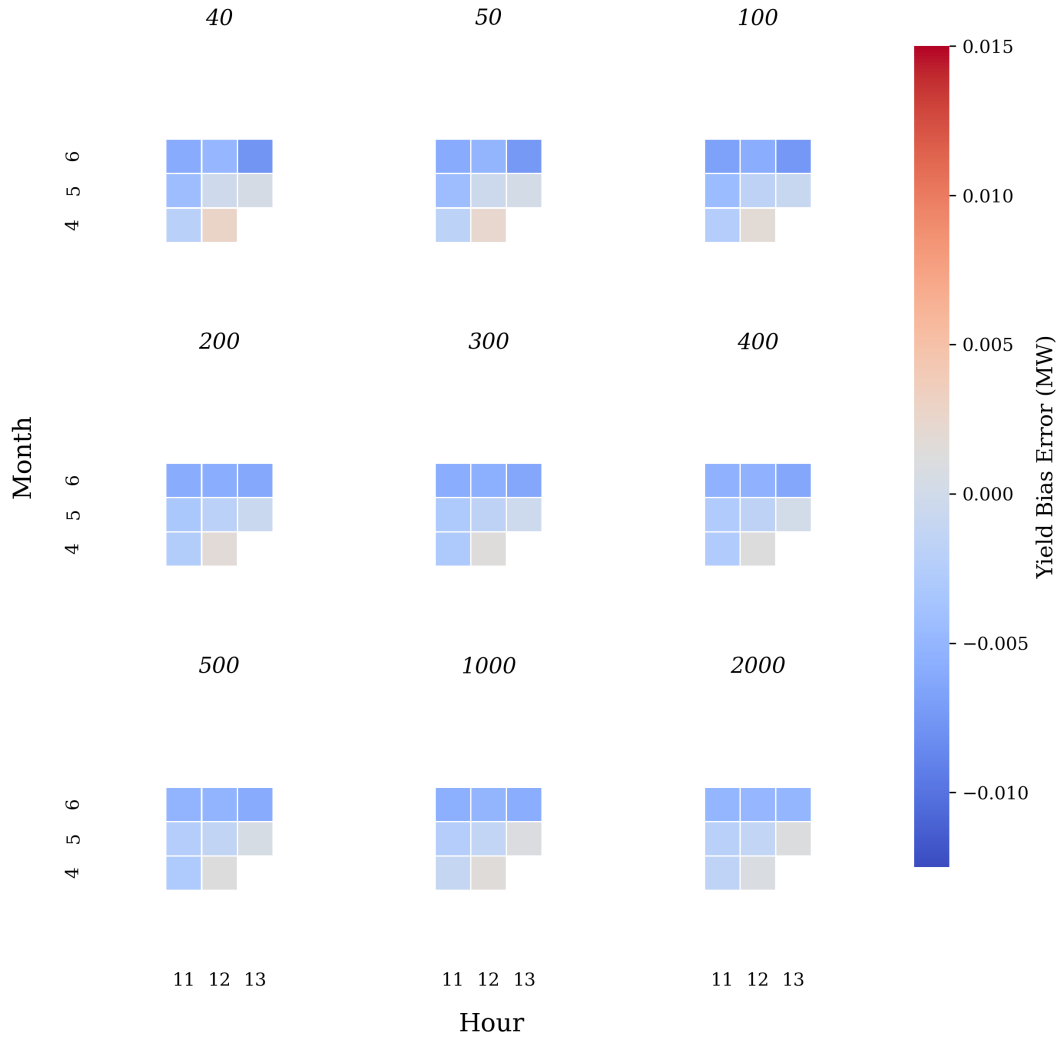


Figure 3.11: *Heatmaps of the bias error for periods when yield fell between 0.7 and 0.8. One heatmap is shown for every sample size tested and missing values are shown in white.*

ability in their bias error. There is no data for January with sufficient yield, but the tendency to overestimate the midday yield in February is consistent with the results in figure 3.10. The trend for overestimating PV yield in June is less pronounced but remains for the midday periods. Additionally, the PV yield model now appears to overestimate the yield in April for small sample sizes. As in figure 3.9, for sample sizes $geq2000$ the within month variability in bias error is constant and the model tends to underestimate yield in the latter months in the year.

Figure 3.11 is a heatmap of the yield bias error containing only periods for which the actual yield fell between 0.7 and 0.8. There is less data in this plot than in figure 3.9 and 3.10 because there are fewer periods with yield ≥ 0.7. There is also less variability in the yield bias error and sample size seems to have a weaker effect on the yield bias error for these estimates. However, as in figure 3.9 and 3.10, the PV yield model tends to underestimate PV yield in June.

3.4.3 Discussion

This study investigated the statistical model error associated with national GB yield model. To do this the yield model was computed many times for sample sizes between 50 and 1500 in a Monte Carlo style simulation. For sample sizes of > 6000 the yield model error is shown to be $\leq 1\%$.

The PV yield model error plateaus for sample sizes of ≥ 6000 and when that happens the absolute model error on the PV Live national PV yield model is $\pm 1\%$. This result is inline with previous studies [106, 64, 102, 97] which have shown that the error associated with estimating the yield of a fleet of distributed PV systems is less than 2%. When comparing the result of this study with existing literature the lower model error can be attributed to having a sample size which was two orders of magnitude larger, i.e. tens of thousands compared with 30 systems in Bright et al. [64] or 311 systems in Siant-Drenan [106].

The results from this study indicate that 6000 is a large enough sample to account for the spatial and temporal variation in GB PV power output. This result agrees to an order of magnitude with results on the spatial and temporal correlation of PV systems in Hoff and Perez's 2012 study [51] and with the analysis in the section 2.1. Both of these analyses found that for a half-hour period of observation, PV systems are correlated up to distances of 10-20 km's. Great Britain has a total area of $\sim 200,000\ km^2$ and figure 2.9 indicates that roughly half of Great Britain is covered in solar PV. Therefore, roughly 10000 PV systems

are needed for modelling solar PV output in Great Britain.

In both this study and the study by Bright et al. [64] the model error of $\pm 1\%$ relates to the observational error between the measured values of PV generation of the systems in the sample and the actual value of the PV generation of those systems. This observational error is random and to be expected from a real-world deployment of a fleet of PV systems. One source of random error could be systems going offline within a settlement period which is not picked up by the metering device. Leloux et al. [146] studied 6000 PV systems across 2015 and found that 10% of these systems experience faults with their PV generation. Some of these faults were only present over a very short time-frame and so it is possible that these faults would manifest as random noise in our sample data.

The national PV yield model has a small (-0.3%) negative bias error for large PV yields and the PV yield model shows an inter-annual trend for underestimating the PV yield toward the end of the year (2017). This study covered only one year because one year was deemed sufficient for the main purpose of the research and it was computationally prohibitive to consider a longer period. This interannual variation in the bias error could relate to inter-annual weather variability across Great Britain. One driver for this variability could be the Northern Atlantic Oscillation [147, 148, 149], which has been previously correlated with renewable power output. However, more data would be required for the preceding and following years to completely understand this relationship. Furthermore, since the PV yield bias error has been shown to be below 1% this source of uncertainty is considered of low importance for solar PV monitoring accuracy.

This study was limited by the fact that the same sample data was to test and validate the model. To try and minimise the effects of this the maximum sample size considered in this analysis was $15,000$ and the full sample size is $\sim 22,000$. In future, the sensitivity of the national PV yield model should be tested using a larger independent dataset. The data should come from a range of suppliers and should cover all types of system installation, domestic, commercial roof-top, and ground mount solar farms.

3.4.4 Conclusion

In this study the model error for the GB solar PV yield model has been shown to be $\leq 1\%$ providing that the number of systems used in the model is ≥ 6000. This result agrees with existing literature which also attempts to quantify error in upscaled measurements of national solar PV yield [106, 64, 102, 97] The PV yield model has been shown to underestimate

PV yield in the latter months of the year but more research is needed to identify the exact cause of this uncertainty. Future work should look to validate the PV yield model using larger independent data sources to try and identify any systematic errors inherent in the Passiv UK data.

3.5 Sample bias error

As well as statistical model error, the yield model will suffer from by sample bias if the yield of the systems in the sample is not representative of the yield of the GB PV fleet. The main difference between the sample and the GB PV fleet is that the sample comprises only domestic PV systems. Whereas, roughly half of the GB PV fleet capacity comes from commercial/utility systems. In this section the sample bias error will be investigated so that together with the statistical model error from section 3.4 the total error associated with the national yield estimates can be estimated. To do this I analyse the error associated with predictions made for commercial/utility PV systems by the GB PV yield model when it is computed using the sample of $20,000$ domestic systems from Passiv UK.

Selection bias occurs when the study samples are not representative of the target population. The most obvious source of selection bias in the PV model is that the sample data is exclusively provided by PV systems installed on domestic roof tops. The sample does not include any commercial/utility systems and these systems are different from domestic systems in several ways: domestic systems are always on roof-tops; commercial/utility systems can be on either the ground or on a roof-top, either mounted directly on a pitched roof or mounted in an array on a flat roof; commercial/utility systems that are mounted in an array will have an optimised orientation and tilt; some commercial/utiliity systems will be mounted in an array which mechanically moves the orientation and tilt of the panels to best track the local irradiance and weather conditions although trackers are uncommon in Great Britain; and commercial/utility systems tend to use cheaper PV panels than domestic PV systems.

In order to analyse the sample bias, a sample of $\sim$ 700 commercial/utility systems was provided by ElectraLink with meter readings, address, and capacity information. The dataset includes measurements of the half-hourly net electricity flow through all export MPANs $\geq$ 30 kW. ElectraLink have analysed the timeseries data associated with each MPAN to identify solar PV systems. They have then matched the address of the MPAN with capacity data from government subsidy scheme datasets, namely the REPD. The 30 kW cut

off for this dataset means that all the PV systems in this sample will be commercial PV systems.

It is important to note that the MPAN data from ElectraLink contains the net electricity flow through the electricity meter. Consequently, any self-consumption of electricity behind the meter will act to reduce the electricity flow through the meter and will degrade the PV generation data. Therefore, this analysis must identify and remove MPANs which have self-consumption so that the net electricity flow through the MPAN is equivalent to the PV generation of the PV system.

The Electralink data is sensitive because it contains information relating to MPANs and MPANs are considered personally identifiable data in the electricity industry. Therefore, this data set is subject to GDPR and as such the data must be aggregated before it can be analysed. In this analysis, the data will be aggregated by Grid Supply Point to anonymise the electricity data from each individual MPAN.

The regional PV yield model [30] will be evaluated against a sample of data from commercial ground mount solar PV systems provided by Electralink. The regional PV yield model will be simulated using the day-plus-one sample data containing data from ~ 20000 domestic PV systems to estimate the PV yield for each PV system in the Electralink sample. The results from the PV yield simulation will then be aggregated by Grid Supply Point and evaluated against the sum of the net electricity flow of all MPANs from the Electralink sample associated with each Grid Supply Point. The regional PV yield model is used here because it has a much higher spatial resolution than the national PV yield model and is more suited to estimating the yield of individual PV systems.

3.5.1 Method

The analysis in this study was split into three parts: cleaning and preparing the Electralink MPAN data, estimating the PV yield for every MPAN using the regional PV yield model trained on the Passiv sample of ~ 20000 domestic systems, and evaluating the bias error the yield estimates at each Grid Supply Point.

The MPAN data was cleaned to remove systems with self-consumption and poor-quality capacity data. To minimise the number of PV systems in the sample with self-consumption, only systems larger than 3 MW were included in this analysis. This cut off was chosen because systems larger than 3 MW are likely to be ground mount solar farms and therefore

they are unlikely to include significant self-consumption. Additionally, performance analysis of each individual MPAN was used to filter out MPANs with erroneous capacity data or self-consumption. For this pre-processing the yield of each MPAN was compared with it's neighbours. To do this, for each MPAN and for each Half Hour, the yield of each MPAN was compared to its nearest neighbours (minimum of 5) and only MPANs which fulfilled the criterion in equation 3.9 for more than 95% of measurements were included.

$$\overline{Y}_{NN} - 3\sigma_{NN} \leq Y \leq \overline{Y}_{NN} + 3\sigma_{NN} \tag{3.9}$$

Once the ElectraLink data had been cleaned, the regional PV yield model was computed using the historic Passiv sample of ~ 20000 domestic PV systems for training. Once the model was trained, it was used to predict the PV yield for each MPAN. The estimated PV yields of these MPANs were then aggregated by Grid Supply Point using geographic boundaries associated with each Grid Supply Point on the distribution grid [32]. An identical aggregation was also performed on the net electricity flow data for each MPAN provided by Electralink.

Correlation analysis has been performed to establish the empirical relationship between the modelled and measured PV yield. Correlation plots of the modelled against measured yield were created. For this analysis, the yield has been uses instead of power output because the bivariate analysis would be biased towards periods with larger PV ouptuts if the generation was used. Additionally, normalisation is required in order to compare the results from different Grid Supply Points with different volumes of installed solar PV capacity. For each correlation graph, a straight line was fitted using Ordinary Least Squares regression and the R Squared and RMSE of the line of fit were computed.

Additionally, the bias error between the modelled and measured PV yield has been calculated. The bias error was then binned according to the yield measured by ElectraLink in intervals of 0.1. Box plots of the bias error have then been plotted for each bin to show the relationship between the bias error and the measured PV yield.

3.5.2 Results

In figure 3.12, the half-hourly PV yield model estimates for each Grid Supply Point have been plotted against the actual PV yield per Grid Supply Point, as measured from the ElectraLink data. The data points are coloured according to their non-parametric number den-

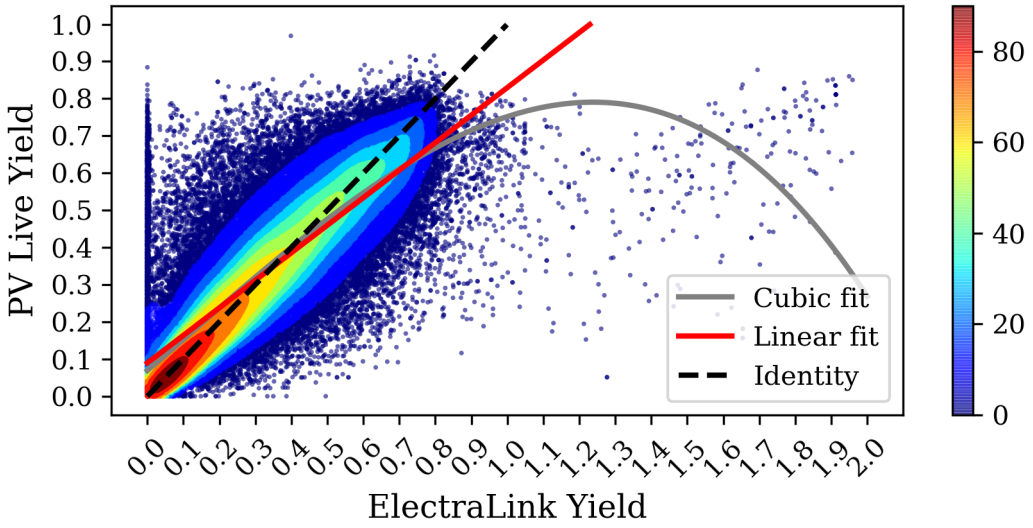


Figure 3.12: *Bivariate plot of the half-hourly PV yield model estimates for each grid supply point against the actual PV yield measurements. The data in the plot has been restricted between 10am and 2pm and the colouring represents a non-parametric number density.*

sity calculated from a Gaussian kernel density approximation and cubic and linear fits have been plotted using Ordinary Least Squares regression. The data should follow a linear fit, however, the cubic fit is more representative of the bulk of the density of the data illustrated by the area enclosed by the teal data points. Meaning that under clear-sky conditions, there is a negative bias error in the yield estimates for individual Grid Supply Points. Additionally, there are many periods with very low measured yields but medium-to-high modelled yields and there are some periods with PV yields much greater than 1 which are not physically possible.

Figure 3.13, shows the same plot as figure 3.12 except that periods with either a modelled or measured yield smaller than 0.01 or larger than 1 have been excluded. In this plot, the offset from the origin for the cubic and linear fits is smaller than in figure 3.12. This is because the very low measured yields have been excluded. However, the data still generally follows a cubic fit that is roughly linear for yields less than 0.5 and then for larger yields the modelled yield underestimates the actual yield. Implying that the model is unbiased for yields less than 0.5 and negatively biased for yields larger than 0.5 and also that the yield measurements greater than 1 were not the cause of the bias error in the yield model.

In figure 3.14, the bias error between the modelled and measured yield has been binned according to the measured yield into groups of width 0.1 and a box plot has been plotted for each bin. The data shows that the yield model is unbiased on average for yields smaller

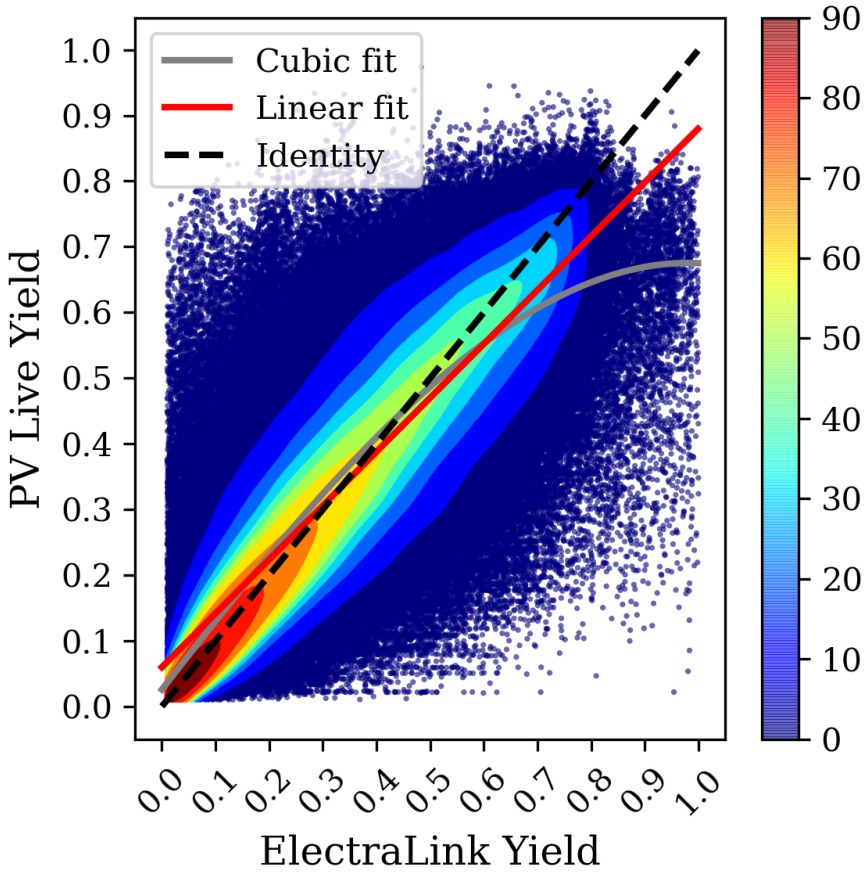


Figure 3.13: *The half-hourly yield averaged across all ElectraLink systems for each Grid Supply Point between 2014-11-01 and 2017-11-01. Only periods between 10am and 2pm are shown and the yield has been restricted between 0.01 and 1 for both the modelled and measured data. The colouring represents a non-parametric number density.*

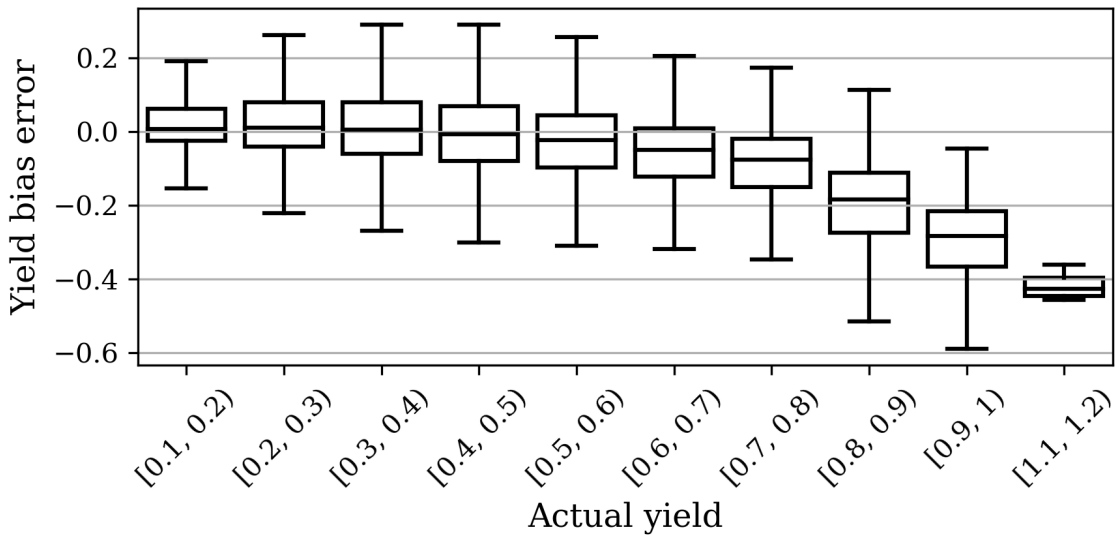


Figure 3.14: *The bias error of the GSP level PV yield estimates as a function of measured PV yield.*

than 0.6. However, for measured yields of 0.6 or larger the yield model is negatively biased. Meaning that the yield model is underestimating yield under clear-sky conditions.

In figure 3.15, the half-hourly PV yield averaged across all Grid Supply Points has been plotted against the actual PV yield averaged across all systems in the ElectraLink data. The data points are coloured according to their non-parametric number density calculated from a Gaussian kernel density approximation and cubic and linear fits have been plotted using Ordinary Least Squares regression. The data in this plot is much more densely packed around the identity line than the GSP level data and it is best represented by the linear fit with a gradient of 1.033. Figure 3.16 shows the bias error for the national average of the PV yield estimates across all Grid Supply Points and further demonstrates that when averaged nationally the yield estimates are unbiased.

3.5.3 Discussion

This study investigated the selection bias in the regional GB PV model caused by the input sample containing only domestic PV systems. The PV yield model was computed using a data from $\sim$ 20000 reference domestic PV systems supplied by Passiv Systems. The accuracy of PV yield model estimates were assessed using a sample of commercial/utility scale PV systems provided by Electralink.

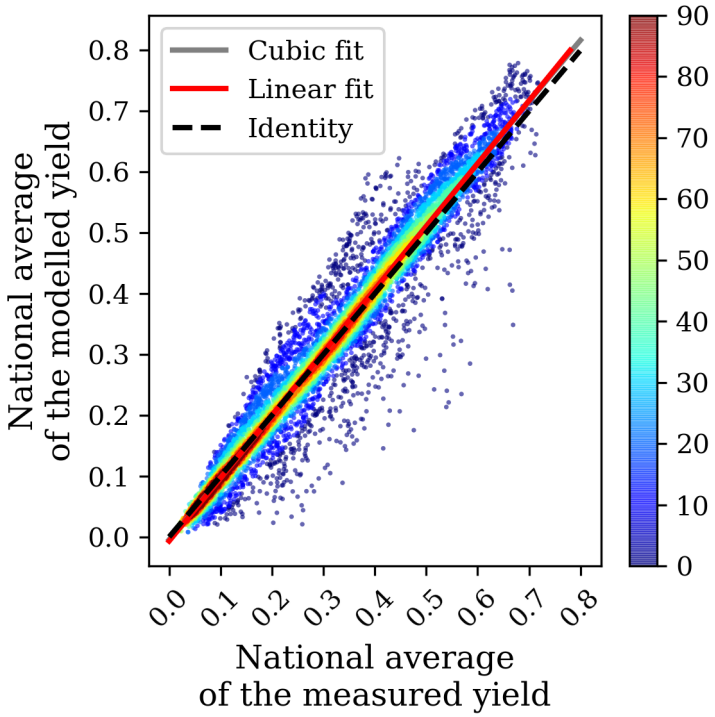


Figure 3.15: *The half-hourly national yield calculated by averaging the yield for all Grid Supply Points between 2014-11-01 and 2017-11-01 where the only systems present in the ElectraLink data set have been modelled. Only periods between 10am and 2pm are shown and the yield has been restricted between 0.01 and 1 for both the modelled and measured data. The colouring represents a non-parametric number density.*

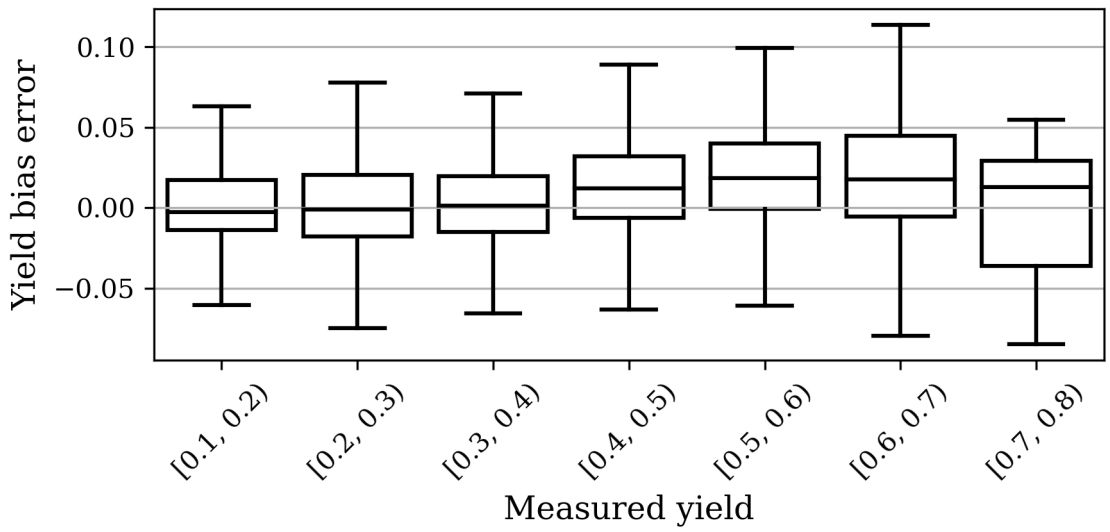


Figure 3.16: *The bias error for the national average of the GSP level yield estimates, plotted as a function of the measured yield.*

The results show that there is a significant bias error of up to -20% when the Passiv data is used to estimate the yield of commercial/utility systems at the regional level. However, there is no significant bias present when the data is averaged at the national level. Thus implying that for national yield estimates there is no significant sample bias error associated with the yield model when computed using a domestic sample.

In figure 3.13, the cubic fit best represents the distribution of the data and for yields smaller than 0.5 the cubic fit almost runs parallel to the identity line but with a small positive bias. Implying that for low yields there is a small positive sample bias error because domestic systems perform better than commercial/utility systems. This performance difference is likely because domestic system are roof mounted and therefore do not suffer from inter-array shading which, depending on the spacing between each set of panels in the commercial PV panel array, has been shown to result in between 1-30% of the surface of PV panels to be shaded [91]. Additionally, domestic systems are less likely to be subject to horizon shading because they are elevated to roof level.

For large yields in figure 3.13 the cubic fit falls below the identity line with quite a large negative bias. This relationship is also seen in figure 3.14 which shows that there is a $\sim -20\%$ bias error for yields greater than 0.7. Meaning that under clear-sky conditions the Passiv domestic systems have a 20% lower yield than the commercial/utility systems from the ElectraLink data. I believe this reduced performance is caused by higher panel temperatures for domestic systems under clear-sky conditions. Due to the fact that domestic systems are roof-mounted and therefore insulated on one side, whereas, commercial systems are ground mounted in an array, allowing for better air flow around the panels. The power output of PV systems de-rates by -0.4% per Kelvin increase in panel temperature [91]. Therefore, the -20% reduction in yield which we observe in this analysis corresponds to an increase of 50 degrees in cell temperature between domestic roof-mounted systems and ground-mount commercial systems.

Whilst the regional yield estimates have been shown to be positively biased for low yields and negatively biased for high yields. Figures 3.15 and 3.16 demonstrate that the national average of the yield across all Grid Supply Points is unbiased. The reduction in bias error occurs because the -20% bias is observed only under very high irradiance conditions and such conditions rarely occur uniformly across the UK. It is more common for irradiance to vary significantly between GSPs. Therefore, when averaged over the UK, this bias error shrinks significantly and it can be concluded that for national estimates of the PV yield there is no significant sample bias error for commercial/utility systems.

In the GSP level data in figure 3.12, measured yields much greater than one occur. The effect of these measurements is is visible in the lower right hand quadrant of the national fit in figure 3.15. One explanation for these data points is that some of the commercial/utility PV systems started generating before the "first generation" date which is recorded for them in the GB site list. Another explanation is that there is some other form of generation connected to the MPAN which is called upon when solar PV output is low. The meters at each MPAN only record the net flow of electricity, not the generation of the solar panel. Therefore, if another generator is connected to the same MPAN, then the measured yield will be incorrect. It is unclear, to what extent solar PV systems are paired with other generators and therefore it is difficult to determine how much of this net-flow electricity data has been corrupted by other generators. In future, a source of direct metered PV output data for a set of solar farms should be sought to rule out any impact from on-site consumption and generation.

3.5.4 Conclusion

The GB PV yield model was trained using a sample of domestic systems and used to predict commercial/utility systems for which data was provided by Electralink. When the measured PV yield was ≥ 0.5 the yield model estimates were negatively biased, by as much as -20%. The source of this error is likely a temperature difference between the cell temperatures between roof-mount domestic and ground mount commercial/utility. This error source has a significant impact on the sample bias at a regional level. However, because of the variability in weather across the country the effect of this error on the sample bias when estimating the national solar PV yield is negligible. Therefore, the sample bias error for the national PV yield model is insignificant and we can conclude that using domestic systems is acceptable for the purpose of modelling the national PV yield in Great Britain.

3.6 Chapter summary and conclusion

In this chapter, the accuracy of our national PV yield model for Great Britain, developed in partnership with National Grid ESO, is analysed. To do this three research questions were considered; how accurate are the intraday PV output estimates?; what is the statistical model error for the national PV yield model?; what is the sample bias error associated with the domestic Passiv UK sample?

Firstly, in section 3.3 the accuracy of the intraday PV output estimates were investigated. There are three differences between the intraday and historic model estimates: the site list; the metering method for the reference PV systems; and the number of systems in the reference PV sample. This analysis aimed to identify whether there were any systematic biases between the intraday and historic estimates of PV power output and identify the causes of any discrepancy.

The intraday case study identified a Root Mean Square Error of $\pm 640\ MW$ between the intraday model estimates and the historic model estimates. Controlling for the site list reduced the RMSE to $\pm 295\ MW$. Indicating that half of the error associated with the intraday model estimates comes from inaccuracies in the baseline site list used in the model. The source of this error is the simulated "uplift" capacity which is included in the site list to try and reduce error caused by the lag between system installation and reporting. Clearly, the calculation of uplift capacity is not error free. Further controlling for the metering mechanism had little effect on the intraday model estimates RMSE. Indicating that most of the remaining error is associated with the size of the reference system dataset used in the intraday model (~ 1000) compared with the historic model ($\sim 20,000$).

In section 3.4, the relationship between the statistical model error and sample size is investigated. The GB PV yield model was computed for sample sizes between 50 and 15,000 and compared with the model estimates computed using all $\sim\ 20,000$ available sample systems. The statistical model error for the GB PV yield model was shown to be $\leq 1\%$ for sample sizes greater than 6000 systems which is inline with similar results from existing literature [106, 64, 102, 97]. This experiment demonstrates that the yield error relating to sample size is low when a sufficient sample is achieved. However, a systematic bias may exist if the sample of systems is not representative of the wider GB PV fleet. For example, the sample used in the model only contains domestic PV systems so it is unclear if the performance of these systems will be representative of the performance of commercial/utility systems.

In section 3.5, the sample bias error for the prediction of commercial/utility systems was analysed via comparison with net metered demand data from ~ 700 solar farms in Great Britain. The regional PV yield model was trained using the domestic sample of systems from Passiv and used to estimate the PV yield of each solar farm in a sample of data provided by ElectraLink. Both the modelled and measured yields were then aggregated by GSP in order to anonymise the GDPR sensitive data from Electralink.

The results of sample bias error analysis indicates that there is a significant negative bias

for PV yields ≥ 0.5 of up to -20% for regional estimates of the yield of commercial/utility systems. However, in Great Britain the clear-sky conditions which cause this bias rarely occur over the whole country. Therefore, when the yield of these commercial systems is aggregated nationally, the sample bias error is insignificant.

In summary, once a sample of 6000 systems is achieved in Great Britain the statistical model error associated with the GB PV yield estimates is $\leq \pm 1\%$. Furthermore, even though all of the reference systems are domestic, there is no sample bias present in the yield estimates at a national level. However, the results in section 3.5 suggest that this might not hold true at a regional level. In the next chapter. the capacity error will be investigated in more detail.

Chapter 4

Accuracy of the national capacity estimate

We are running the most dangerous experiment in history right now, which is to see how much carbon dioxide the atmosphere can handle before there is an environmental catastrophe.

– Elon Musk

4.1 Introduction

Figure 4.1 shows the three processes associated with our approach for monitoring GB solar PV output. This chapter will investigate the error associated with the second part of this process; the capacity error. To do this the different error sources which affect solar PV capacity data will be defined and then their effect on the national solar PV capacity estimate will be simulated stochastically in a Monte Carlo styly analysis.

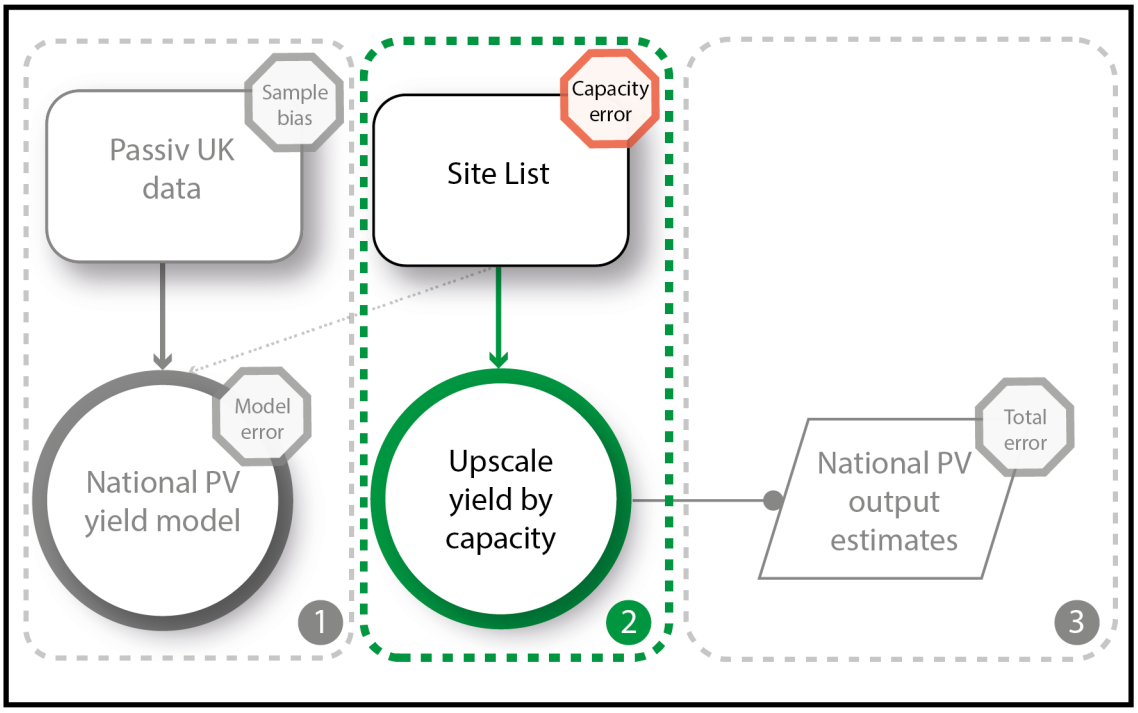


Figure 4.1: *The modelling approach of the GB solar PV monitoring service, PV Live.*

In the background section, a survey of national PV monitoring services was carried out which considered 27 national solar PV monitoring services from 20 different countries. Whilst all of these services use a baseline national site list to upscale their yield estimates, only The Dutch [84], German [121, 96], Australian [85], and Italian [126] services disseminate the source(s) for their baseline site list. The German and Italian governments have facilitated one asset register which records the installation of all PV systems in their region. In Germany this is the EEG register [79] and in Italy it is the GAUDI system [80]. The Dutch service describes collating PV asset data from multiple disparate data sources in their peer-reviewed publication [76]. Some of their data sources are public: the Nationaal Solar Trendrapport; the Dutch Central Bureau of Statistics (CBS); the Klimaatmonitor, and some of their data sources are provided by commercial organisation and are confidential: Solar Monkey, Eindhoven University of Technology (TU/e), and the Dutch transmission service

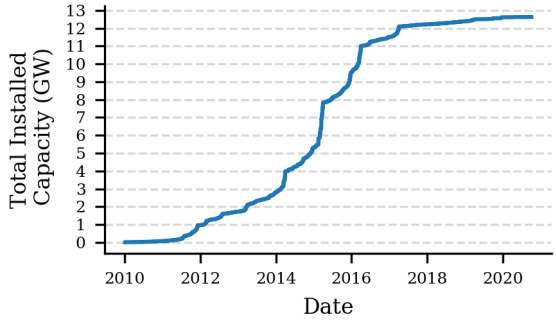


Figure 4.2: *The cumulative total installed capacity of the panels in the GB PV fleet e.g., the total Direct-Current capacity of all the panels in the GB PV fleet.*

operators (TSOs) Stedin and Alliander. Additionally, the Australian service describes on their website [150] their process for deriving a national baseline site list by taking all systems smaller than 100 kW from the Small-scale Generation Unit (SGU) database and all systems larger than 100 kW from the Large-scale Renewable Energy Target (LRET) database.

The GB baseline national site list, is prepared by collating multiple sources of capacity information: Government Feed In Tariff registers [151], Renewable Energy Planning Database [81], Renewable Obligation Certificate register [82], and a commercial capacity dataset from Solar Media [83]. These datasets include location and system size and are brought together, and cross checked to ensure that there is no double counting. A single comprehensive site list is compiled [33] which includes every reported system in Great Britain. Systems which occur in one or more source registers but do not have location data are allocated a location pseudo-randomly according to the likelihood of the systems location based on the capacity-weighted spatial kernel density estimate of PV systems with a known location. The growth in the GB capacity as measured by the site list compilation [33] is shown in figure 4.2.

Between 2010 and 2015 there was rapid growth in the installation of domestic systems due to release of the UK Feed in Tariff [151]. Since 2015, commercial systems have dominated the growth in installation of capacity and since the closure of the Feed-in-tariff, Renewable Obligation, and Contract for Difference schemes the reported capacity has been relatively stable. However, uncertainty in the reported capacity is growing as the PV fleet ages. One reason for this is that these capacity datasets have no update mechanism. Another reason is that registration is less likely than in the past since solar PV is now subsidy free and the subsidy schemes were the main sources of capacity information. Distribution network operators are starting to release embedded generation asset data through the Electricity Network Association asset register [152], but this is limited to the larger systems and

so far there has been no attempt to validate the accuracy of this data. Furthermore, the ENA asset is not currently used in the baseline site list derivation.

In the second process in figure 4.1 the yield, as calculated using the method in section 3.2, is scaled by the capacity associated with every system in the GB PV site list. Any error in the recorded capacity translates directly into an error in the estimated PV output and there are many failure modes for GB capacity data:

- ***Transcription error***: a transcription error might have occurred when the capacity was recorded.
- ***Unreported***: some systems might be missing from the site list altogether.
- ***Offline***: some systems might be offline at the time of observation due to system maintenance or fault.
- ***Revision***: the capacity of the system might have been revised since the initial installation by either adding or removing panels.
- ***Decommissioned***: some systems have been decommissioned because they are at end-of-life or they have been destroyed in a fire. Additionally, some systems are removed by new owners on property sale.
- ***Network outages***: if the local electricity network which connects a solar PV system to the wider electricity grid is experiencing an outage then the power generated from that system will be lost.

In this chapter, to investigate the error in the national capacity estimate associated with these capacity error sources, the baseline site list is systematically iterated using a Monte Carlo approach to evaluate the combined influence of each source of uncertainty. For each source of uncertainty upper and lower bounds on the enhancement and/or reduction in capacity over the baseline site list are chosen using relevant literature studies. For each iteration of the Monte Carlo analysis values of the uncertain input parameters for each source of uncertainty are chosen randomly from their priors and by sampling thousands of times a statistical distribution of the likelihood of the total capacity is computed.

Sources of uncertainty in the capacity data and site list are presented and discussed in section 4.2. My approach to estimating the uncertainty in PV capacity is qualitatively like a failure mode effects analysis [153] and has been used in similar energy system studies when little data exists from which to quantitatively characterise the uncertainty of input

parameters [154, 155]]. Potential "failure modes" or, they are called here - sources of uncertainty, are listed (table 4.2) and then in-turn, independent information sources are used to estimate the likely potential lower and upper bounds of the impact of each source on the total capacity.

Some sources of variation, such as a transcription error in system capacity or location, apply to a system over the entire operation lifetime of the system. Some sources of uncertainty in capacity, such as a system being offline temporarily due to a network outage or a local trip due to over voltage protection circuits, vary temporally, occurring for a period before disappearing. To simplify the modelling approach, sources with temporal variation have been substituted as a system-to-system variation for different time windows in a combined analysis. Consequently, the repeats in the Monte Carlo experiment can be thought of as a set of independent days where any error is assumed to occur over the whole day.

The Monte Carlo method for estimating the likely distribution of national capacity is detailed in section 4.3. The sources of uncertainty are sampled 10 thousand times in a Monte Carlo approach with upper and lower limits assigned to each of the different uncertain input parameters. The model outputs a range of likely possible national PV capacities.

4.2 Sources of uncertainty in GB solar PV capacity

Historically, the installation of GB solar PV systems have been recorded through one or more parallel registration databases. Distinct registration schemes existed for three types of system: domestic, commercial, and utility. In this section we will characterise the different sources of uncertainty which affect solar PV capacity data. We consider domestic and commercial/utility installations separately and summarise the different sources of uncertainty in table 4.2. Where $10\ kW$ is chosen for the cut off between domestic and commercial/utility systems.

Most domestic systems are accredited through a trade organisation scheme known as the Microgeneration Certification Scheme (MCS). For a system to be accredited, the installer must be certified with the MCS and a fee must be paid upon registration. The MCS has a register containing data on all accredited systems, but this data is not public and therefore it has not been used explicitly in any GB capacity register or in our site list.

Between 2010 and 2019 a second parallel registration scheme that allows owners to claim a feed in tariff (FIT) [151] has been operational. The FIT database is administered

by the UK Office For Gas and Electricity Markets and made available publicly via the Government Department for Business, Energy and Industrial Strategy [156]. To be eligible for FIT registration, the installer of the system must be registered with the Microgeneration Certification Scheme. In February 2021 there were 477 companies registered on the Microgeneration Certification Scheme database. However, studying UK companies house registrations highlights hundreds of installation companies that are not registered with the Microgeneration Certification Scheme. Hence, there is probably a significant domestic PV capacity that is not included in the FIT register, either due to the installer not being MCS registered, or the system not being registered for the FIT.

There are also multiple registers for commercial/utility scale PV systems. The Renewable Obligation Certificate incentive operated between 2002 and 2017 for commercial systems with capacity $> 50\ kW$ and published its own register. The Renewable Energy Planning Database is a synthesis of PV system data extracted from 126 local authority planning datasets and is assembled quarterly by a private company under a contract issued by the Department for Business, Energy and Industrial Strategy.

The Renewable Obligations Certificates dataset tracks the declared net capacity of the inverter as opposed to the total installed capacity of the panels. For modelling PV yield the total installed capacity of the panels is required and so the renewable obligations datasets is not used in the site list [33]. The REPD tracks the total installed capacity of the panels for all PV systems which applied for planning permission from their local authority. In the UK, this should be all non-domestic PV systems, but it may include systems for which planning has been granted but were never installed or commissioned. The REPD provides a broad overview of the current and future deployment pipeline for non-domestic renewable energy systems. However, it includes large discrepancies between the planned and installed capacity of solar PV systems.

Solar Media Ltd. are a company offering commercial insights for the solar and storage sector in Great Britain. They compile system information for all PV systems larger than 30kW for PV systems in Great Britain and Northern Ireland. To research system meta data they cross check multiple data sources and often speak directly with the PV asset owners. For this reason, the Solar Media data is considered the best quality data for PV system information in Great Britain.

Information on individual PV systems is often duplicated across multiple PV capacity data sets. For example, every ground mount solar system must apply for planning permission so their will be at least one entry for these systems in the Renewable Planning

Database. Sometimes planning permission will have been applied for in two parts resulting in two entries in the Renewable Planning Database. The same system might have also applied for to either the Renewable Obligation Certification scheme or the Contracts for Difference scheme. However, this application might relate to both entries in the Renewable Planning Database. Additionally, if the system is larger than 30kW then the system should also be found in the Solar Media data. After the system has been installed it is possible that the PV system owner will have applied for more planning permission, resulting in more entries in the Renewable Planning Database and possibly more entries in the subsidy data sets.

To understand the accuracy of the recorded system information the individual system records have been cross-referenced across multiple data sets. To do this PV system data has been cross-referenced across three GB capacity datasets: The Renewable Planning Database (REPD), Renewable Obligations (RO), and Solar Media (SM). To cross reference the entries for each PV system across each dataset, the datasets were sorted in ascending order and the 676 largest systems were matched. To start with each data set was searched internally to group together multiple entries for the same PV system. Then all of the entries associated with an individual system were located across every data set. Entries were matched both internally and between datasets by manually searching each using a combination of system name, location, capacity, system age, and size. Once every system had been located in each data set, the row identifiers relating to each system from each database were collated and a lookup table was created.

When comparing the recorded capacity between the three data sets, it has been assumed that the Solar Media capacity data is reliably accurate because their approach of speaking directly with PV asset owners after system installation is the most robust.Therefore, the capacity error associated with the capacity information in the RO and REPD datasets is quantified by calculating the normalised residual given by the residual between the capacity measured in RO or REPD and the capacity recorded in Solar Media divided by the capacity registered in Solar Media.

In figure 4.3, distributions of the normalised residuals of the capacity recorded in the Renewable Obligation Certificate's data set and Renewable Energy Planning Database. The Renewable Energy Planning Database is shown to systematically over report system capacity by 1.5% and Renewable Obligations dataset is shown to underestimate system capacity by 18.4%. In conclusion, several percent error exists between the system capacity recorded in the different Government capacity registers. Additionally, there is several percent more

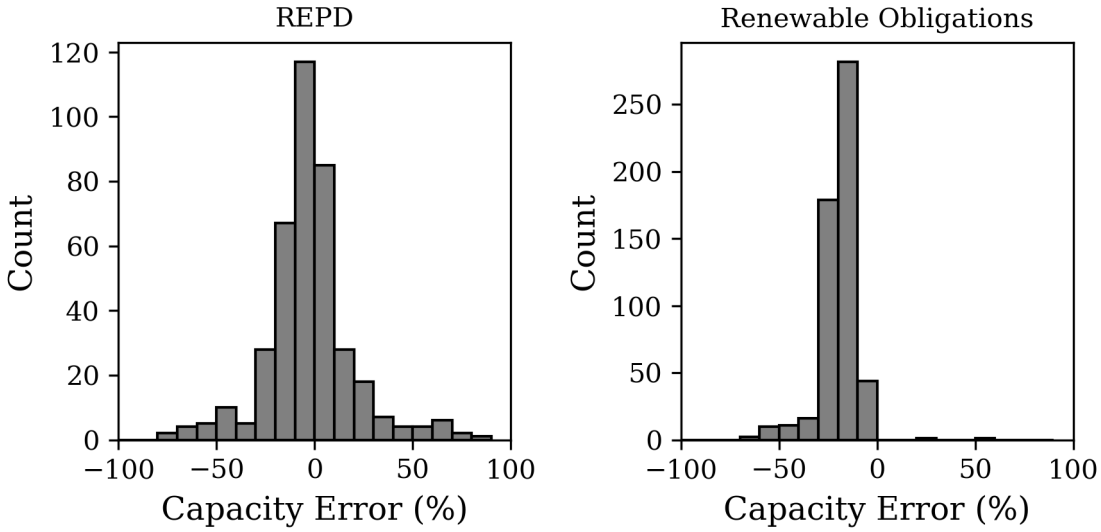


Figure 4.3: *The error in the individual system capacity as recorded in the Renewable Energy Planning Database (REPD) and the Renewable Obligations (RO) database. The error was evaluated by calculating the normalised bias error with respect the system capacity recorded in the Solar Media dataset.*

uncertainty in terms of capacity that is not reported at all. Our estimates of the likely range of these uncertainties are presented below.

4.2.1 Unreported

There are some domestic PV systems missing from one or both of the MCS and FIT registers. Figure 4.4 shows the decline of the FIT rate along with data from the Department for Business Energy and Industrial Strategy (BEIS) tracking the number of FIT registered and unregistered systems over time. To track FIT registered systems BEIS have cross-referenced the MCS and the FIT databases and identified systems which are in the MCS but did not the FIT. As the FIT rate falls the number of FIT registered systems also falls. Historically, the main motivation to register with the MCS was to qualify for the FIT, and so I believe it is also reasonable to assume that MCS registration falls. Hence, overtime the number of systems which are not registered with either the FIT or the MCS (aka. unreported) increases.

To estimate a benchmark for the number of systems that are neither FIT nor MCS registered (unreported) we use, as a first estimate, the same probability that a system is unregistered for the FIT given that it is registered for the MCS. This is shown in figure 4.4 as the cumulative ratio of the number of systems unregistered for the FIT to the number of

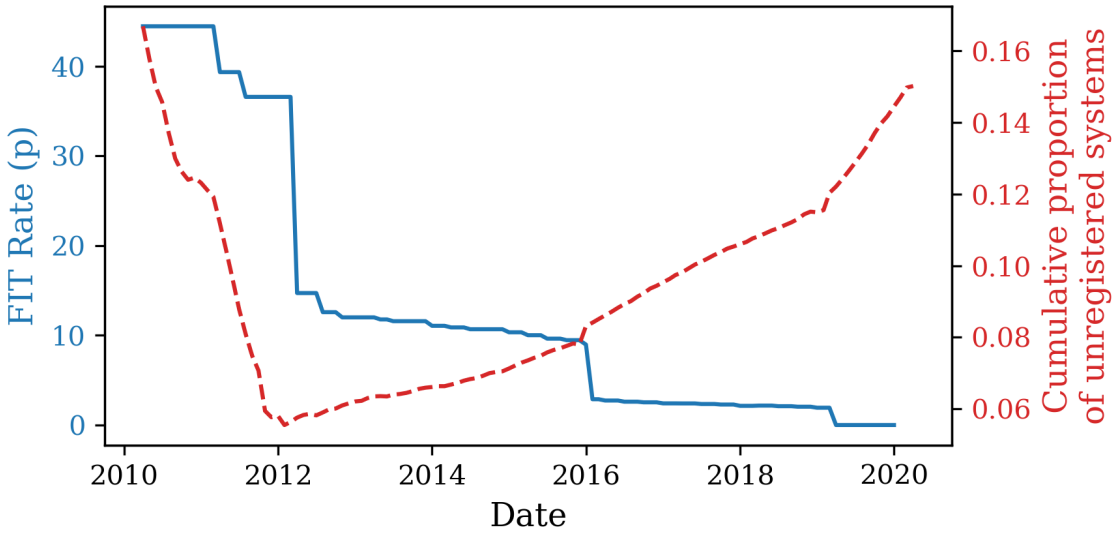


Figure 4.4: *The dashed red line gives the cumulative proportion of unregistered systems and the blue line gives the average FIT rate available for each month. This data has been taken from table 2 of the UK governments' solar photovoltaics deployment tracker [157].*

systems registered for the FIT, as calculated from government data [157]. For example, if in a given period 10 systems registered only for the MCS and 100 systems registered for the MCS and the FIT then we would estimate the probability of a system being unreported as 10%. Using BEIS data the monthly probability of a system being unreported is calculated. In figure 4.5, unreported probability is plotted as a function of the FIT rate available at the time of installation. In choosing high and low limits for the probability of a system being unreported (i.e. non-MCS and non-FIT) in the Monte Carlo investigation these benchmarks were varied by $\pm 10\%$ to give the high and low limits for the probability that a system was unreported.

In addition to the stochastically selecting the unreported probability between the high and low limits identified above. The stochastically generated unreported probability is multiplied by a scaling factor linked to system size to reflect the fact that most of the unreported systems are small domestic installations. The benchmarks are scaled by 0.5, 0.5, 0.25, 0.05, 0 for systems between $0\ to\ \leq 4kW$, $4\ to\ \leq 10kW$, $10\ to\ \leq 50kW$, $50\ to\ \leq 5MW$, $\geq 5MW$, respectively. The scaled probabilities are then used to calculate the number of unreported systems and the results of this are shown in figure 4.6 and table 4.1.

The benchmark number of unreported systems is estimated to be 215,000, 4400, 1400, 30 and 0 for systems of size $0\ to\ \leq 4kW$, $4\ to\ \leq 10kW$, $10\ to\ \leq 50kW$, $50\ to\ \leq 5MW$, $\geq 5MW$, respectively. Assuming the unreported systems follow the same capacity distri-

bution as the reported systems the unreported systems provide: 660 MW, 29 MW, 43 MW, 67 MW for systems of size $0\ to\ \leq 4kW$, $4\ to\ \leq 10kW$, $10\ to\ \leq 50kW$, $50\ to\ \leq 5MW$, respectively. Thus, it is estimated that there is currently a total of 0.8 GW of unreported capacity in Great Britain.

Anecdotal accounts of unreported PV systems are becoming increasingly common in the PV industry. For example, a Flemish magazine article highlights 5000 unreported systems which Fluvius (the electricity and gas network operator in Flanders) identified using aerial imagery in Flanders, Belgium [158]. However, only one previous study exists that attempts to estimate this unreported capacity for Great Britain. Stowell et al. [144] used satellite imagery and a public information campaign which utilised the open street map community and machine learning to locate and quantify the capacity of $270k$ domestic and commercial systems across Great Britain. They conclude that the installed base of solar PV in Great Britain could already be as large as $16\ GW$, $3.5\ GW$ higher than the accepted value at the end of 2020. I see this as an upper limit due to the methodological limitations: requiring impossible knowledge of PV module sizes a priori and assuming that all observed systems were connected and operational. However, this upper bound is the only other academic report of GB capacity against which we can compare this work and it is consistent with these findings.

System Size	**MCS accred. and FIT unaccred.**	**MCS and FIT unaccred.**	**Total**
$0\ to \leq 4\ kW$	141,796	76018	217.814
$4\ to \leq 10\ kW$	0	4371	4371
$10\ to \leq 50kW$	0	1447	1447
$50\ to \leq 5\ MW$	0	32	32
$> 5\ MW$	0	0	0
Total	**147,96**	**81,868**	**223,664**

Table 4.1: Results for the simulation of the total number of unreported PV systems in GB as of January 2020, broken down by system size (the same as in [157]). The number of MCS accredited and FIT unaccredited systems is taken from BEIS' reporting on solar PV deployment in the UK [157].

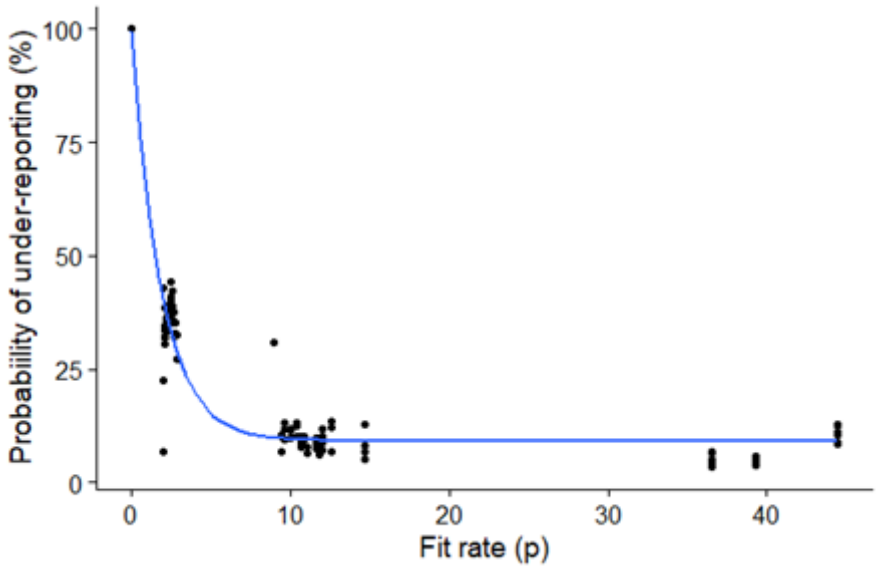


Figure 4.5: *The probability that a system is installed and is unaccredited for the FIT. The unaccredited probability is plotted as a function of the FIT rate available at installation. The black points are measured data points taken from [1] and calculated as the ratio between number of systems which were installed and accredited with the MCS but not with the FIT and the number of systems which were installed and accredited with both the FIT and the MCS. The blue line denotes an exponential line of best fit which has been plotted using non-linear least squares regression.*

4.2.2 Transcription error

The accuracy of the reported capacity of PV systems is also uncertain. For domestic systems, an example of this can be seen in by investigating the capacity of systems with reported capacity equal to the band cut off in the feed in tariff. Feed in tariff bandings incentivise installation of specific sizes of systems to maximise return by minimising per kWp costs within a band. For systems smaller than 50kW, the FIT band classification is determined by the Total Installed Capacity (TIC) of the PV panels. We believe that many systems will have had their Declared Net Capacity of the inverter reported instead of the typically higher Total Installed Capacity to ensure band compliance. There is evidence for this in the Government's own photovoltaics deployment tracker [157] which details that for the MCS systems there is a 3.7% difference between the DNC and the TIC. It is expected, for an individual system that the difference between the DNC and the TIC is roughly 10 %. Therefore, it is likely that a significant number of domestic systems have incorrectly recorded either the TIC. Additionally, figure 4.7 shows a bimodal histogram of the peak power output from a sample 190 4 kWp domestic systems. From these 4kWp systems 15% have peak outputs greater than 4kW which is beyond what can reasonably be expected for a 4 kWp system

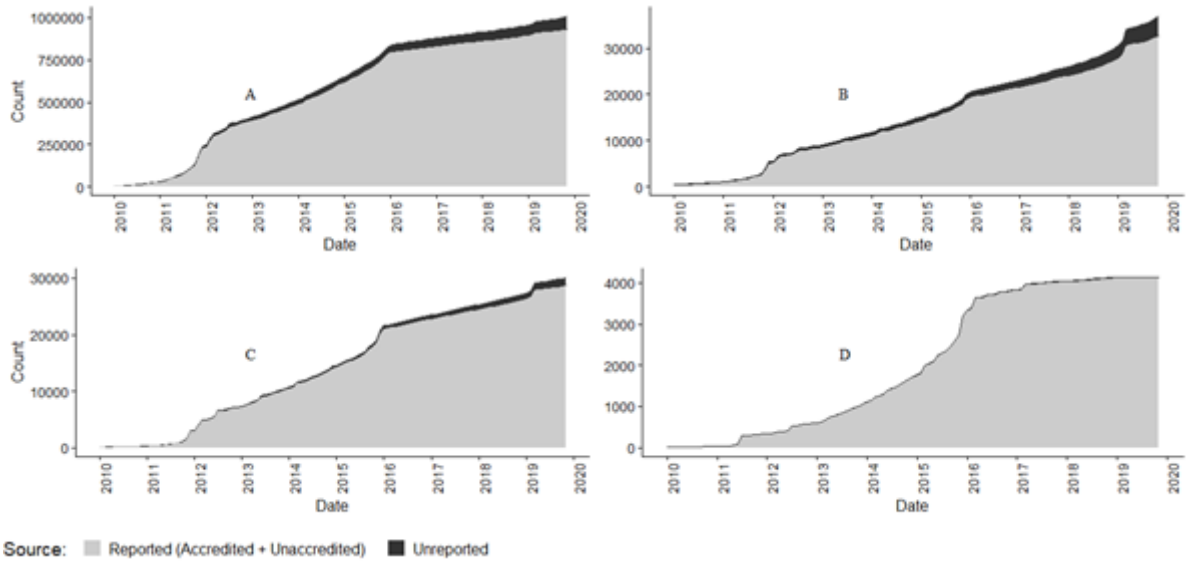


Figure 4.6: *Cumulative count of reported and unreported solar PV systems in GB broken down by system size: Plot A includes systems with capacity* $0\ to \leq 4\ kW$*, plot B includes* $4\ to \leq 10\ kW$*, plot C includes* $10\ to \leq 50kW$*, and D includes* $50\ to \leq 5\ MW$*. The grey shows the growth in the total count of systems which are reported for the FIT and the black shows the additional number of simulated unreported systems.*

and is consistent with recording DNC instead of TIC.

Figure 4.7 highlights that often the reported capacity for solar PV systems does not match exactly the capacity installed. Since there are roughly 1 million domestic solar PV systems in Great Britain these errors will exist in both directions in roughly equal proportions. Therefore, in our Monte Carlo model we will assume a rounding error with high and low values of $\pm 0.5\ kW$ which will be simulated using a normal distribution with mean of 0 and standard deviation of $0.167\ kW$.

To estimate the uncertainty in commercial PV system data, the Renewable Planning Database has been cross referenced with the Solar Media data as presented in figure 4.3. The cross-referencing results show that only 40% of the REPD systems had the correct capacity recorded. Figure 4.3 shows there was a negative bias error of -2.75% for recorded capacity of the the 60% of systems which had an incorrect capacity.

The probability that a commercial/utility system suffers a transcription error will be simulated using a uniform distribution between $55\tilde{}65\%$. Given a transcription error, the system capacity is scaled by a random number generated from a truncated normal distribution with a mean of -99.25% and standard deviation of 26% and bounds of [0, 100].

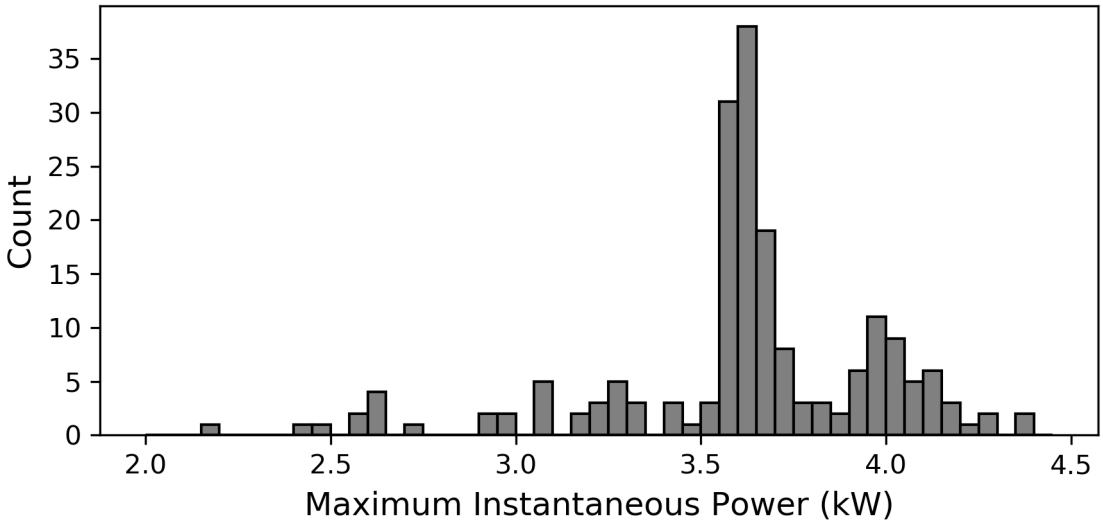


Figure 4.7: *The distribution of maximum instantaneous power readings from 190 PV systems with maximum recorded capacity of 4kW. The data was collected using a 2-minutely temporal resolution for reading the instantaneous.*

4.2.3 Revision and decommissioning

Although less significant than uncertainty in system capacity and unreported capacity, we also expect an influence from: revisions to system capacities (either the addition or removal of panels); and system decommissioning. System decommissioning can happen at: end-of-life, which is unlikely since most systems are < 10 years old; during change of ownership of a property; or following discrete events such as fire or inverter failure. As of May 2018, the BRE report that there have been 80 fires due to PV systems [159], so whilst uncommon they have been included. Of these 80 fires, 37 were domestic, 37 were non-domestic, and 6 were solar farms. We believe that a reasonable estimate for the number of decommissioned domestic and commercial/utility solar PV systems is ~ 1000 and ~ 100, respectively. Moreover, we believe it likely that an order of magnitude more systems have had their capacity revised than have been decommissioned.

For both domestic and commercial/utility systems, we believe it is reasonable to assume that most revision increases PV system capacity. Therefore, in the Monte Carlo model if a PV system is revised, its capacity will be scaled by a factor drawn from a truncated normal distribution centred on 1.2 and with standard deviation of 0.15, and with bounds $[0, 2]$.

4.2.4 Offline system capacity

Data on offline domestic systems can be found in PV system fault literature. Leloux et al. [146] performed fault analysis on 6000 European solar PV systems. They found that 10% of the systems exhibited faults, half of which are attributable to the system being offline or having a faulty inverter. Firth et al. [160] also performed fault analysis on 27 solar PV systems in the UK. They categorised faults as; sustained zero efficiency, brief zero efficiency, shading, and non-zero efficiency and non-shading. Sustained zero efficiency and non-zero efficiency and non-shading are equivalent to the offline and string outage errors referred to in this work. Firth identified that these faults occurred in PV systems for roughly 7-8% of the hours for which the PV systems were operational.

Additionally, we have analysed the number of missed readings from a sample of roughly $\sim$ 1000 PV systems distributed across the UK which provided 2-minutely instantaneous PV power readings. A missed reading was classified by a missing 2-minute reading in a day in which a PV system provided data. If the system was offline for the whole day, it was excluded from this analysis. The average percentage of missed readings across every system in the analysis, and the number of systems providing readings are plotted for 2019 in supplementary figure 4.

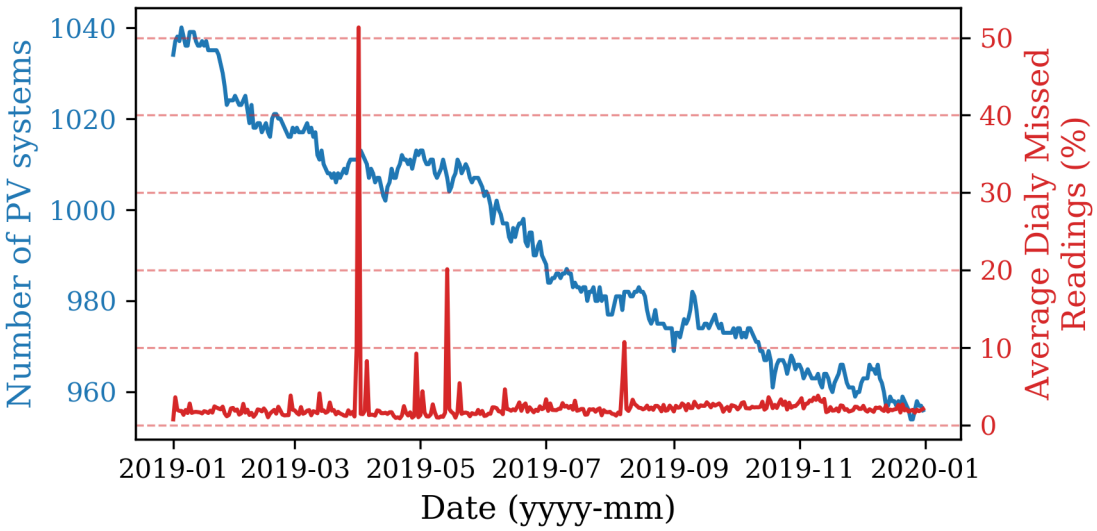


Figure 4.8: *The number of PV systems providing data during 2019: the left axis shows the number of PV systems providing data each day (blue) and the right axis shows the mean percentage of missed readings (red) across all PV systems which provided readings for each day.*

Figure 4.9 shows the distribution of the percentage of missed readings across all systems

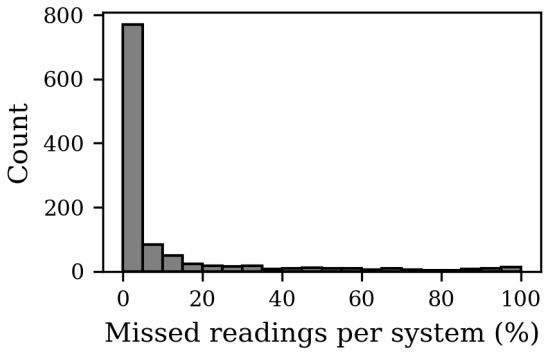


Figure 4.9: *The distribution of the missed readings for each system in the sample across 2019. This distribution is long-tailed and therefore cannot be described accurately using the mean and standard deviation because these estimators are skewed by the large values in the long tail. Instead, the interquartile range and the median provide better illustration of the distribution of the data. The 25th percentile is 1.1, the median is 1.7 and the 75th percentile is 6.9.*

in 2019. The distribution is long-tailed and therefore it is not well represented by a normal distribution. Instead, the interquartile range and the median provide better illustration of the distribution of the data. The 25th percentile is 1.1, the median is 1.7 and the 75th percentile is 6.9. Thus, broadly agreeing with the data from Leloux et al [146] and Firth et al. [160]. It is important to note that the cause of the missed readings were not identified and could be unrelated to the output of the PV system e.g., a problem with the metering system. Additionally, any system fault which occurs over > 24 hours is not included in this analysis. The data from firth et al. suggests that sustained zero efficiency faults could contribute roughly a third of all PV system faults, so this will be significant.

Considering the results from my own analysis on missed readings and those from Firth et al. [160] and Leloux et al. [146], the uncertainty in offline PV systems will be modelled using a normal distribution with a mean of 4% and a standard deviation of 0.5%. This uncertainty range was reached by considering the percentage of operational hours which were affected by sustained zero efficiency faults from Firth et al [160]. Furthermore, I believe that both the data in Leloux et al. [146] and my own analysis justifies this estimation for the likelihood that a domestic PV system is offline.

Data on the likelihood of commercial systems being offline can be found in year-end reports for renewable energy investment funds. For example, in their yearly reports Foresight Solar Investment Fund details the effect outages have on the yield of their portfolio. Their 2016 report states that due to offline systems their fleet output reduced by 1.5% [161]. Furthermore, in their 2019 report Next Energy Solar Fund state that offline outages reduced their portfolio output by 0.1% in 2019 and 1.8% in 2018 [162]. Using this data offline com-

mercial PV systems will be modelled using a normal distribution with a mean of 1.5% and a standard deviation of 0.5% was used to simulate if commercial systems were offline.

4.2.5 Network outages

To simulate network outages on domestic solar PV systems, probabilities are taken from PV performance literature. Leloux et al. [146], found that faults occur with 10% of domestic PV systems and half of these faults were attributable to network outages. The authors of this paper did not specify the average time to fix these faults, just the likelihood that a particular PV system would be affected by a fault. We believe that a network outage is likely to be a brief fault since the network companies are constantly monitoring their network for faults and have engineers available 24 hours a day to respond to fix any faults which occur. Firth et al. [160] found that domestic PV systems are affected by brief zero efficiency faults between 0.1% and 1.3% of their operational hours. Therefore, in this analysis the likelihood that a network outage affects a domestic system will be simulated using a normal distribution with a mean of 0.7% and a standard deviation of 0.2%.

Data on network outages for commercial systems can be found in year-end reports for renewable energy investment funds. Foresight Solar Investment Fund state that in 2016 a grid outage resulted in a 3.8% reduction in their portfolio output [161]. Next Energy Solar Fund reported that network outages reduced their portfolio output by 1% in 2019 and 0.9% in 2018 [162]. Therefore, for the probability that a network outage occurs to a commercial system, we will use a normal distribution a mean of 2% and a standard deviation of 1%.

4.2.6 Summary of uncertainties

In summary, I have estimated that for domestic systems: there are currently between 200k and 240k unreported systems with a total capacity of between 600 and 700 MW; the reported capacity of domestic systems is subject to a ±0.5kW rounding error; between 100 and 1800 systems have been decommissioned; between 1k and 18k systems have had their capacity revised since they were installed; and domestic systems are affected by network outages between 0.1 – 1.3% of the time.

I have also estimated that for commercial/utility systems: there are currently between 1.3k and 1.6k unreported systems with a total capacity of between 100 and 120 MW; between 18.5k and 22k systems have inaccurate capacity data with an error between -75 % and 81

%; between 10 and 190 systems have been decommissioned; between 100 and 1.9k systems have had their capacity revised since they were installed, and commercial/utility systems are affected by network outages between 0 – 5% of the time.

Table 4.2: Upper and lower limits on the different sources of uncertainty in the capacity of domestic and commercial/utility PC systems in Great Britain. To refers to uncertainty at initial system installation and Tt uncertainty relates to post installation changes. [80]–[86].

Domestic systems		
Unreported	Systems missing from the capacity registers and site list.	The number of unreported domestic systems has been simulated once for each of the 10 thousand simulations. Section 3.2 details the benchmark for the number of unreported systems. For each of the 10 thousand simulations, a scaling factor is generated using a uniform distribution between [0.9, 1.1] and the number of systems in the benchmark of unreported systems is scaled using this randomly generated number. The unreported systems are created by sampling the baseline site-list using sampling with replacement. The site list for the simulation is given by the combination of the baseline site list and the randomly selected unreported sample.
Decommissioned	Inaccuracies in recorded capacity.	The probability that any domestic system has been decommissioned is chosen once for each of the 10 thousand simulations. To do this a truncated normal distribution with $\mu = 0.1\%$, $\sigma = 0.3\%$, and bounds of [0,100] have been used for the probability of decommissioning. For each system, a random number is generated from a uniform distribution between [0, 100] and if it is less than the decommissioning probability then it's capacity is scaled by zero.
Transcription error	Inaccuracies in recorded capacity.	A rounding error has been simulated for each domestic system for each of the 10 thousand simulations. The size of the rounding error has been simulated by drawing from a normal distribution with $\mu = 0\ kW$, $\sigma = 0.167\ kW$. Once the rounding error has been selected for each system its capacity is modified by taking the sum of the original capacity and the rounding error.
Revision	Systems whose capacity has been modified since installation.	The probability that any domestic system has had its capacity revised is chosen once for each of the 10 thousand simulations. The probability of revision was generated using a truncated normal distribution $\mu = 1\%$, $\sigma = 0.3\%$, and bounds [1, 00]. Then for each system, a random number is generated from a uniform distribution between [0,100] and if this number is smaller than the generated probability of revision then the system has its capacity scaled by a number generated from another normal distribution with $\mu = 1.2$, $\sigma = 0.15$.

Table 4.2: Upper and lower limits on the different sources of uncertainty in the capacity of domestic and commercial/utility PC systems in Great Britain. To refers to uncertainty at initial system installation and Tt uncertainty relates to post installation changes. [80]–[86].

Offline	Systems which are offline due to a local system fault, e.f. a component failure, maintenance, or a variation in grid frequency.	The probability that any domestic system is offline chosen once for each of the 10 thousand simulations. The probability of a system being offline is generated using a truncated normal distribution with $\mu = 4\%$, $sigma = 0.5\%$, and bounds [0, 100]. For each system, a random number is generated using a uniform distribution between [0, 100] and if this random number is less than the probability that a system is offline then it has its capacity scaled by zero.
Network outage	Inverter shutdown during a power cut, planned network outage or an unplanned network outage.	The probability that any domestic system is affected by a network outage is simulated once for each of the 10 thousand simulations. The probability of the network outage is generated using a truncated normal distribution with $\mu = 0.2\%$, $\sigma = 0.7\%$, and bounds of [0, 100]. For each system, a random number is generated from a uniform distribution with bounds was used to simulate if a domestic system was affected by a network outage.
Commercial/utility systems		
Unreported	Systems missing from the capacity registers and site list.	The number of unreported domestic systems has been simulated once for each of the 10 thousand simulations. Section 3.2 details the benchmark for the number of unreported systems. For each of the 10 thousand simulations, a scaling factor is generated using a uniform distribution between [0.9, 1.1] and the number of systems in the benchmark of unreported systems is scaled using this randomly generated number. The unreported systems are created by sampling the baseline site-list using sampling with replacement. The site list for the simulation is given by the combination of the baseline site list and the randomly selected unreported sample.
Decommissioned	Systems included in capacity registers and site list but no longer installed.	The probability that any non-domestic system is decommissioned is simulated once for each of the 10 thousand simulations. The probability that a non-domestic system has been decommissioned is generated using a truncated normal distribution with $\mu = 0.3\%$, $\sigma = 0.09\%$, and bounds [0, 100]. Then for each system, a random number is generated using a uniform distribution between [0, 100] and if this random number is less than the probability that a system has been decommissioned then the system capacity is scaled by zero.

Table 4.2: Upper and lower limits on the different sources of uncertainty in the capacity of domestic and commercial/utility PC systems in Great Britain. To refers to uncertainty at initial system installation and Tt uncertainty relates to post installation changes. [80]–[86].

Transcription error	Inaccuracies in recorded capacity.	The probability that non-domestic systems have had their capacity recorded incorrectly is chosen once for each of the 10 thousand simulations. The probability of incorrect capacity transcription is generated using a uniform distribution between [55, 65]. Then for each system, a random number is generated using a uniform distribution between [0, 100] and if this random number is less than the probability of incorrect transcription then the system capacity is scaled using another random number generated using a truncated normal distribution with $\mu = 99.9725\%$, $\sigma = 26\%$, and bounds of [0, 200].
Revision	Systems whose capacity has been modified since installation.	The probability of revision for non-domestic systems is chosen once for each of the 10 thousand simulations. The probability of revision is generated using a truncated normal distribution with $\mu = 3\%$, $\sigma = 0.9\%$, and bounds [0, 100]. Then for each system, a random number is generated using a uniform distribution between [0, 100] and if this random number is less than the probability of revision then the system capacity is scaled by another random number generated using a truncated normal distribution with $\mu = 1.2$, $\sigma = 0.15$, and bounds [0, 200].
Offline	Systems which are offline due to a local system fault, e.g. component failure, maintenance, variation in grid frequency or voltage.	The probability of a non-domestic system being offline is chosen once for each of the 10 thousand simulations. The offline probability is generated using a truncated normal distribution with $\mu = 1.5\%$, $\sigma = 0.5\%$, and bounds [0, 100]. Then for each system, a random number is generated using a uniform distribution between [0, 100] and if this random number is less than the offline probability then the system capacity is scaled by zero.
Network outage	Inverter shutdown during a power cut, planned network outage, or an unplanned network outage.	The probability of a non-domestic system being affected by a network outage is chosen once for each of the 10 thousand simulations. The network outage probability is generated using a truncated normal distribution with $\mu = 2\%$, $\sigma = 1\%$, and bounds [0, 100]. Then for each system, a random number was generated using a uniform distribution between [0, 100] and if this random number was less than the network outage probability then the system capacity was scaled by zero.

4.3 Monte Carlo Model

In this section the method for the Monte Carlo model for national solar PV capacity is described. The model takes as input a baseline site list containing system information for all known installed PV systems in Great Britain. The baseline site list, used in this research has been curated by NGESO using the python software [33]. The Monte Carlo model for national solar PV capacity then modifies the baseline site list by randomly simulating the effect of each of the 6 sources of capacity uncertainty on each site in the site list as described in table 4.2.

The Monte Carlo model is evaluated 10 thousand times, and an estimate of the probability density function of the national effective capacity is calculated using equations 4.1 and 4.2. Where: y_i is the national effective capacity of each Monte Carlo simulation, n_s is the number of Monte Carlo simulations, $\widehat{E}(y)$ is the estimated value of the national effective capacity, and $\widehat{\sigma}$ is the estimated standard deviation of the national effective capacity.

$$\widehat{E}(y) = \frac{1}{n_s}\sum_{i=1}^{n_s} y_i \tag{4.1}$$

$$\widehat{\sigma}(y) = \sqrt{\frac{1}{n_s}\sum_{i=1}^{n_s}\Big(y_i - \widehat{E}(y)\Big)^2} \tag{4.2}$$

To investigate the effects of these different errors, three different Monte Carlo experiments have been computed: simulating only the transcription error, simulating the transcription error and operational errors, and simulating all errors (e.g. transcription, operational, and unreported). In figure 5 the distribution of the national capacities resulting from each of the 10 thousand simulations for each experiment have been plotted.

4.4 Results

Figure 4.10 shows the results of the capacity error analysis for the three different experiments: one simulating only the transcription error, one simulating the transcription and the operational errors, and one simulating all errors (transcription, operational, and unreported systems). The national capacity from the baseline site list is shown as a dashed black line.

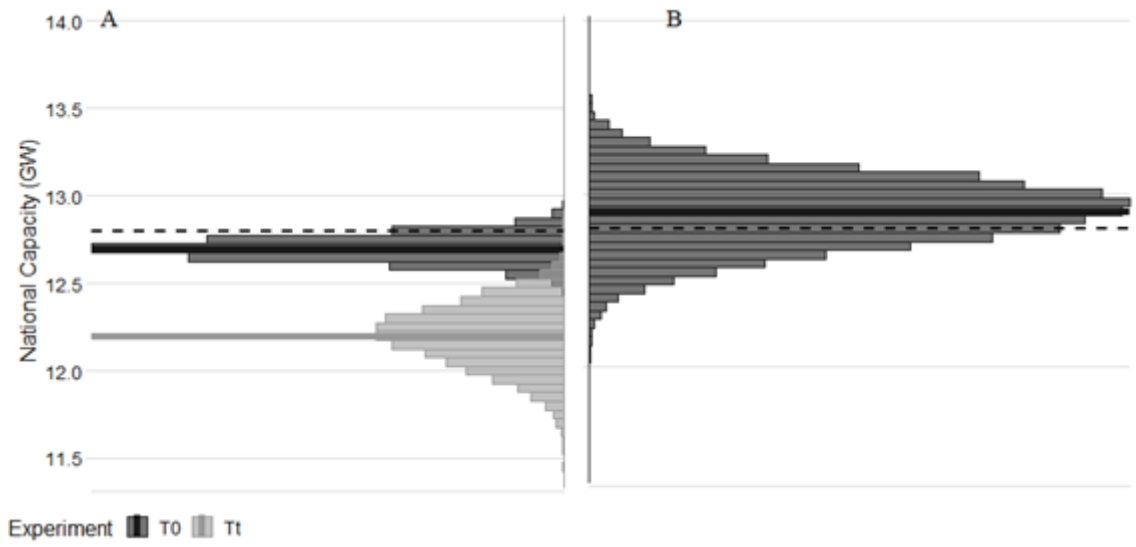


Figure 4.10: *Histograms of the national PV capacity, for 1000 simulations of the Monte Carlo model of national capacity. The benchmark capacity is shown by the dashed line and is the sum of all capacity in the initial site list.*

The dark grey histogram in left hand side of figure 4.10 shows the results of the transcription error Monte Carlo experiment. The mean and standard deviation of this distribution are 12.6 GW and 0.07 GW, respectively. Corresponding to an uncertainty in the national capacity of $\pm 1.5\%(\pm 3\sigma)$. This distribution represents the uncertainty in the national capacity estimate caused by error in the transcription of capacity for individual systems and acts as a baseline uncertainty relating to the inherent quality of the underlying capacity data.

The light grey histogram on the left hand side of figure 4.10 shows the results of the transcription and operational error Monte Carlo experiment. The mean and standard deviation of this experiment are $12.18\ GW$ and $0.19\ GW$, respectively. Hence, the mean national capacity from this experiment is $0.44\ GW$ smaller than for the transcription error experiment and $0.62\ GW$ smaller than the baseline site list. This result is expected since the net effect of the operational error sources is to reduce the national capacity. Additionally, the standard deviation in national capacity is three times larger than the transcription error experiment. The uncertainty in the national PV capacity from this experiment is $\pm 4.7\%(\pm 3\sigma)$.

Finally, in the histogram on the right hand side of figure 4.10 the system list was augmented with unreported systems before being passed through the Monte Carlo model and simulating the transcription and operational errors. Unsurprisingly, the mean of these results is $0.68\ GW$ larger than the transcription and operational experiment. The standard deviation increased from $0.19\ GW$ to $0.21\ GW$. Corresponding to an uncertainty of

$\pm 4.9\%(\pm 3\sigma)$ and indicating that whilst unreported systems increase the size of the national capacity by $0.68\ GW$, they have only a small effect on the total uncertainty of national capacity.

4.5 Discussion

This chapter set out to answer the research question: "what is the error and uncertainty associated with the national solar PV capacity derived from the GB baseline national site list?" Where the baseline site list applies to deployed capacity in Great Britain on the 1st of January 2020. It, has been modified by stochastically simulating the effect of each of the 6 sources of capacity uncertainty on each site in the site list as described in table 4.2. The mean and standard deviation of the distribution of national capacities from these modified site lists are $12.9\ GW$ and $0.2\ GW$ respectively. Corresponding to a total uncertainty in the national solar PV capacity for Great Britain of $\pm 4.9\%(\pm 3\sigma)$.

The mean national capacity from the transcription and operational errors is $0.7\ GW$ less than baseline national capacity indicating that on average 5% of PV capacity at any time does not contribute electricity to the grid. This difference represents the real world operation of electricity systems and highlights that it is unreasonable to expect 100% uptime from PV systems. However, despite this, baseline site lists are still used for estimating the real-time solar PV power production across the world.

Leloux et al. [146] analysed metered PV generation data from 6000 European domestic PV systems and found that they were offline 10% of the time. This translates to $1.3\ GW$ of offline PV capacity for GB. The difference between this paper ($0.7\ GW$) and Leloux's inferred estimate can be explained by reduced downtime for non-domestic PV systems and metering faults.

This analysis estimated that there is $\pm 4.9\%$ uncertainty in grid-connected and operational national capacity for any given period. This is a large uncertainty with respect to the accuracy of monitoring and forecasting electricity demand and supply. For example, the Mean Absolute Percentage Error for state of the art for day-ahead (24 hour forecast horizon) PV forecasts is $\pm 2 - 3\%$ [49]. Additionally, with $/sim 13\ GW$ of installed PV capacity in Great Britain, this uncertainty corresponds to $\pm 0.65\ GW$ of uncertainty in solar PV capacity available for power production for any given period.

When unreported systems were stocastically included in the Monte Carlo model, na-

tional capacity derived from the baseline site list is $0.06\ GW$ smaller than the mean national capacity from the experiment which simulated the transcription, operational, and unreported system errors. Indicating that currently, capacity from unreported PV systems is masking real-time suppression in solar PV capacity. Consequently, the overall bias error associated with the baseline site list is small. However, this balance is unlikely to remain; as the electricity system ages we can expect the overall error in the the national capacity to increase.

The Monte Carlo model for national capacity has been chosen because the aim of the present research is not to generate a method for modelling national capacity, but to compare the uncertainty arising from different sources of uncertainty and their impact on the overall error in PV output measurement. However, in future, and for both investigative and operational purposes, more sophisticated analyses of national PV capacity will be required. In this paper an uncertainty quantification has been performed. In future, a more robust method such as local sensitivity analysis could be considered, in which the effect of each parameter is considered in isolation.

The most significant and controllable error is the number of unreported systems. To reduce the number of unreported systems governments must facilitate accurate system registers, as they have done in Germany with the "EEG Register" and have started in Great Britain with the System Wide Resource Register [152]. However, currently only systems with capacity greater than $1MW$ are included the GB System Wide Resource Register. Therefore, this registration system will be missing roughly half of the current installed PV fleet capacity.

It is imperative that regulation around registers motivate system owners to accurately register and update information on their assets. Germany has incentivised asset owners to sign up by withholding subsidy payments for non-compliance. Countries with less generous subsidy schemes cannot rely on this tactic and will have to employ other means to ensure full compliance with asset registers. Additionally, mechanisms for recording decommissioning of PV systems are required. The German EEG register contains a mechanism for recording decommissioning. However, currently none of the PV systems in the EEG register are recorded as having been decommissioned. Given the size and age of the German PV market it is unlikely that this is the truly the case and this further highlights that the challenge is not only to facilitate accurate recording mechanisms but to incentivise/enforce accurate record updates from system owners.

In addition to improved system registers, a more holistic and open approach to energy

data and models must be adopted to overcome the challenges with monitoring solar PV capacity and generation. Uncertainty from real-time errors can be minimised by moving towards a real-time model of the electricity grid. In such a model, electricity flows from individual generators would be superimposed on a topological grid model. This would allow for the interactions between the real-time network state and the volume of available PV capacity. However, such a model will only be realised when a the energy industry has access to a data information schema which facilitates interoperability between all the dispatate energy datasets and a set of common industry standards which regulate data quality and provision. This recommendation, aligns with a recent publication from Bauer et al. [163] in which they call for the creation of digital twins to improve Earth-system science.

4.6 Conclusion

In this chapter, the error and uncertainty associated with the national capacity estimate for the GB PV fleet were investigated. Six failure modes for solar PV system capacity information were researched and their likely effect on solar PV capacity was quantified through a set of prior probabilities with high and low limits. A Monte Carlo model for national capacity was then computed by modifying the GB PV site list $10,000$ times and each time stochastically simulating the effect of each failure mode on every system in the site list. The total bias error on the baseline site list which is currently used in the GB PV monitoring service was estimated to be $0.06\ GW$ (0.4%) and the total uncertainty in was estimated to be $\pm 4.9\%$.

The bias error is small because the operational errors affecting solar PV systems, such as offline PV systems and network outages, are cancelled out by an error in the baseline site list relating to unreported systems. In future, we expect that this balance will not remain and the total bias error will increase and the PV fleet develops over time. To minimise the number of unreported systems and transcription errors for reported systems, Governments must facilitate accurate PV system registers with mechanisms for updating PV system capacity retrospectively. It is unlikely that any such system would improve operational errors such as offline PV systems and network outages. Instead, to track the operational capacity of distributed technologies like solar PV, Governments and Transmission System Operators should create information schemas which facilitate interoperability between the many disparate energy datasets.

Chapter 5

Total error in PV output estimates

Climate change does not respect border; it does not respect who you are — rich and poor, small and big. Therefore, this is what we call 'global challenges,' which require global solidarity

– Ban Ki-moon

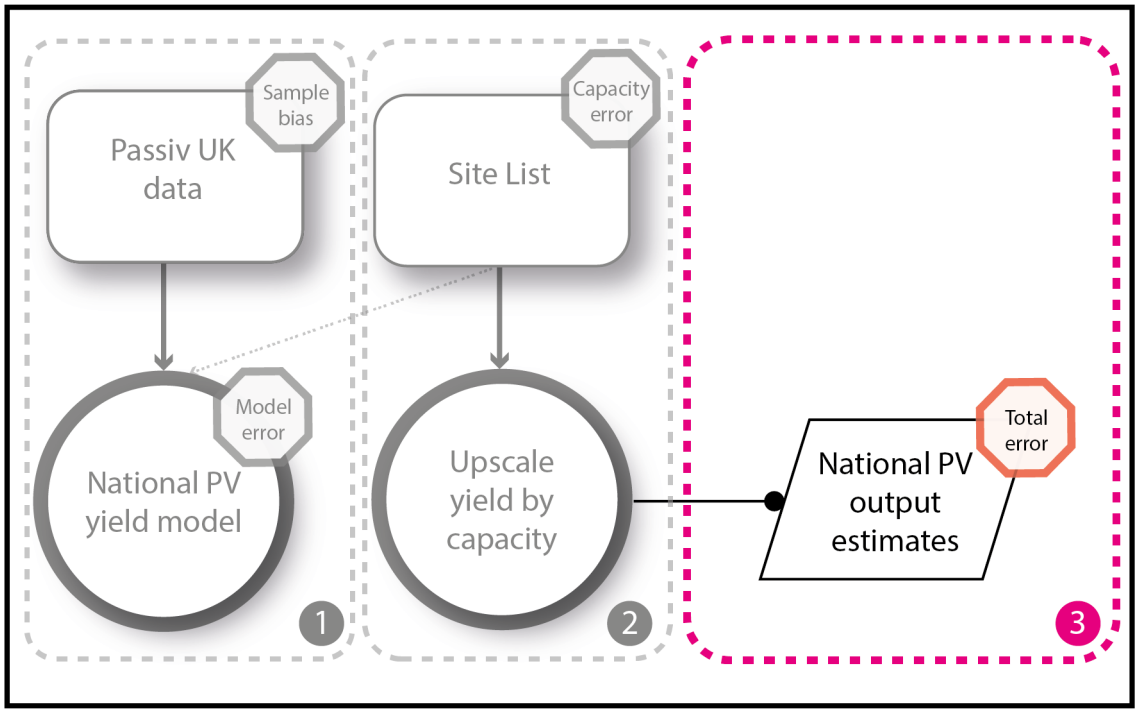


Figure 5.1: *The processed involved in the GB national PV Live, PV output monitoring methodology.*

5.1 Introduction

It is prudent to return once again to figure 5.1 which shows the three processes involved in the GB PV monitoring methodology: 1) modelling the national PV yield, 2) estimating the total installed national solar PV capacity, and 3) calculating the total national solar PV output by upscaling the modelled PV yield with the estimate of national solar PV capacity. In chapter 3 the national PV yield model was shown to have $\leq \pm 1\%$ error on it's yield estimates. Then in chapter 4 the national solar PV capacity data was shown to have a $\pm 5\%$ uncertainty and a $0/4\%$ bias error. In this chapter these errors will be combined in quadrature to estimate the total error associated with the estimates of national solar PV output from the PV Live model.

In this chapter it will be assumed that the two errors from the PV yield model and the national solar PV capacity data are independent. This is done to simplify the maths involved, since if the error sources are dependent then the solution involves covariance terms which are difficult to quantify.

In practise, the errors from the yield model and national capacity data are not independent. This is because, as shown diagrammatically in figure 5.1 by the dashed arrow from

the site list to the yield model, the yield model makes use of the national site list data when calculating the representative sample. Therefore, errors in the capacity data could result in a sample with a spatial distribution which does not follow the spatial distribution of the installed and operational PV fleet.

There are rare circumstances in which the contribution from error in the site list to the yield error might be significant. Namely, when a significant part of the distribution network is offline in a local area which renders a large volume of commercial/utility solar PV capacity offline. However, since the yield error is small compared with the error in national capacity any contribution to the overall error is likely to be insignificant. Additionally, the capacity error derivation depends on many difficult to quantify prior probabilities, such as the likely number of unreported systems. Until the prior probabilities from these error sources are defined more precisely there is little to be gained from a more mathematically robust calculation which includes the covariance between the yield and capacity errors.

5.2 Methods

Solar PV output is calculated as the product of the modelled yield and the national capacity estimate. Hence, the total error in solar PV output estimates is given by the combination of the statistical uncertainty in the yield model, $\leq \pm 1\%$ (chapter 3.4), the sample bias error, 0%, (chapter 3.5), and the capacity error and uncertainty, $0.4\% \pm 5.1\%$ (chapter 4). The sample bias error for national estimates has been shown to be insignificant and the total error in the national capacity is very small (0.4%) because currently unreported systems cancel out operational errors in our capacity information. Therefore, the sample bias error and the capacity error will be ignored and the total uncertainty in the GB solar PV output will be calculated by combining the capacity and yield uncertainties.

For simplicity I treat the yield error and capacity error as independent variables. Therefore, the combination of their uncertainties can be expressed by equations 5.1 and 5.2. In equation 5.2, the covariance term has been excluded since this term is zero for independent variables. The total uncertainty in solar PV generation estimates has been evaluated using equation 5.2. Where σ_Y is the standard deviation of the yield, σ_{C^*} is the standard deviation of the capacity, μ_Y is the modelled PV yield for a given period, and μ_{C*} is the capacity for the same period. In this approach we have assumed that all of the variance in the yield measurement comes from the variation in input sample, and that all the variance in capacity is caused by the failure modes detailed in table 4.2 and which are included in the Monte

Carlo model in section 4.3.

$$\sigma(Y, C^*) = +cov(Y^2, C^{*2}) + \left[\sigma(Y) + \widehat{E}(Y)^2\right] \times \left[\sigma(C^*) + \widehat{E}(C^*)^2\right] - \left[cov(Y, C^*)) + \widehat{E}(Y)^2 + \widehat{E}(C^*)^2\right] \tag{5.1}$$

$$\sigma(Y, C^*) = \sigma(Y)\sigma(C^*) + \sigma(Y)^2\widehat{E}(C^*) + \sigma(C^*)^2\widehat{E}(Y) \tag{5.2}$$

5.3 Results

The total uncertainty in the national solar PV output estimates has been evaluated by taking the expected capacity in January 2020 as $12.86\ GW$ with one standard deviation of $0.21\ GW$ from section 4.3 and one standard deviation of the yield to be 0.33% from section 3.4. Using equation 5.2 the standard deviation of the solar PV output is calculated to be 1.7%. Hence, assuming that the uncertainty in solar PV output is normally distributed, the total uncertainty in national solar PV output is $\pm 5.1\%$ (i.e., $\pm 3\sigma$).

In chapter 4, uncertainty in national capacity data was estimated to be $\pm 5\%(\pm 3\sigma)$ and in chapter 3 the yield model error was estimated to be $\pm 1\%(\pm 3\sigma)$. We can therefore conclude that most of the uncertainty in solar PV output modelling arises from uncertainty in our knowledge of the installed and operational base of solar PV systems. This has significant policy and research implications; we must focus on tracking solar PV capacity if we are to minimise system operation errors caused by its intermittent weather dependent generation. Using this estimate for the total error on the solar PV output, the total solar PV output has then been visualised with uncertainty bounds in figures 5.2, 5.3, and 5.4.

In figure 5.2 the growth in solar PV output has been visualised between 2015 and 2020. To create this graph the GB PV output model was simulated for each half-hour period between the 1st of January 2015 and the 1st of January 2020. For each period high and low uncertainty bounds were generated by scaling the output by $\pm 5.1\%$. The daily maximum output was then selected and a rolling mean with a period of 180 days was computed and is plotted in figure 5.2. The blue line shows the rolling mean and the grey area shows the uncertainty associated with this data as calculated. The graph shows that the absolute uncertainty in solar PV output is greatest in the summer months when solar PV output is largest. It also shows that in recent years for the summer months there has been roughly $1\ GW$ of uncertainty in the modelled solar PV output.

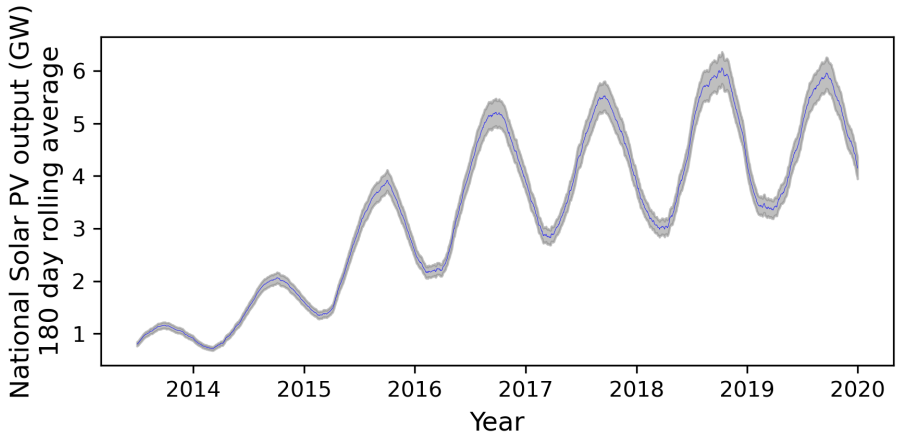


Figure 5.2: *The growth in solar PV generation with error bars showing the* 5.1% *uncertainty range associated with the PV output estimates.*

In figure 5.3 the solar PV output has been plotted for 14th of May 2019 which as of the 1st of January 2020 was the day with the largest single half-hour of national solar output. On the graph the blue line shows the solar PV output calculated by the GB PV Live model and the grey area shows the $\pm 5.1\%$ uncertainty in this calculation. The peak generation on this day as measured by the PV Live model is $9.5\ GW$. However, the graph shows that there is a 1 GW total uncertainty associated with this estimate and that it would have been more informative to report the solar PV output as $9.5 \pm 0.5\ GW$.

In figure 5.4, four graphs have been plotted which break down the total electrical generation on the GB grid into solar and non-solar components for typical weekday and weekend days in winter and summer. The data for the non-solar generation comes from the Elexon, BMReports database and the solar PV output has been computed using our GB PV output model.

By comparing the graphs in figure 5.4 with the cardinal points in figure 1.2 for National Grid ESO's (ESO), national demand forecast. We can evaluate the effect of the uncertainty in solar PV output on each cardinal point. In the summer months the solar PV output is significant between roughly 8am and 6pm. Meaning that solar PV output affects the 2A, 2B, 3B, 3C, and 4A cardinal points in the demand forecast. Whereas in the winter shorter daylight hours mean that the solar PV output is significantly smaller and is more concentrated around the middle of the day. Consequently, the solar PV output only affects the 2A, 2B, and 3B cardinal points.

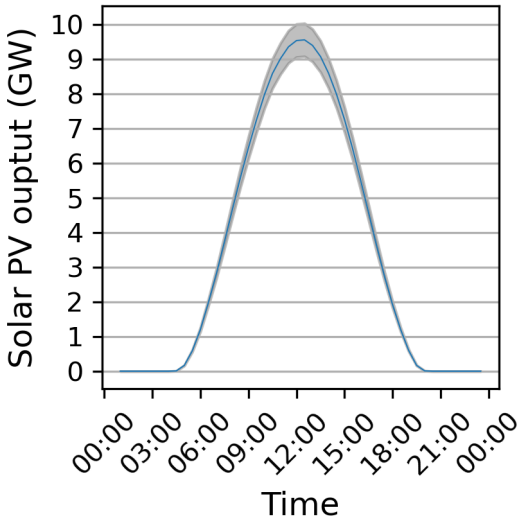


Figure 5.3: *The national solar PV output for the 14th of May 2019 which was the day with the largest single half hour period of solar PV output as of the 1st of January 2020. The blue line is the solar PV output as calculated by the PV Live model computed using the historic reference PV dataset with* $\sim 20,000$ *systems and the grey area shows the* $\pm 5.1\%$ *uncertainty bounds for this estimate of solar PV output.*

5.4 Discussion

In this chapter, the total error for the solar PV output estimates was evaluated. It was assumed that the yield and capacity errors are independent and therefore they were combined in quadrature using equation 5.2 to calculate the total error for the solar PV output. The total error in the solar PV output estimates is $\pm 5.1\%$. Indicating that contribution from the capacity error ($\pm 5\%$) is far more significant than the yield error ($\pm 1\%$) contribution.

Figures 5.2 and 5.3 illustrate the real-world effect of this uncertainty on solar PV output data. Figure 5.3 shows that when GB solar PV output is at it's highest, there is a 1 GW uncertainty associated with the solar PV output estimate. In particular, it is worth nothing that figure 5.3 indicates that Great Britain may have already experienced 10 GW of solar PV power output for a single half-hour period. A milestone which is informally considered to be significant by many in the GB solar PV sector.

Figure 5.4 shows that in summer the solar PV output affects the 2A, 2B, 3B, 3C and 4A cardinal points in ESO's national demand forecast. Although there is PV output across the 2A - 4A cardinal points, the uncertainty idnetified in this thesis is significant only for the 2A, 2B, and 3C cardinal points when the absolute uncertainty in solar PV output is

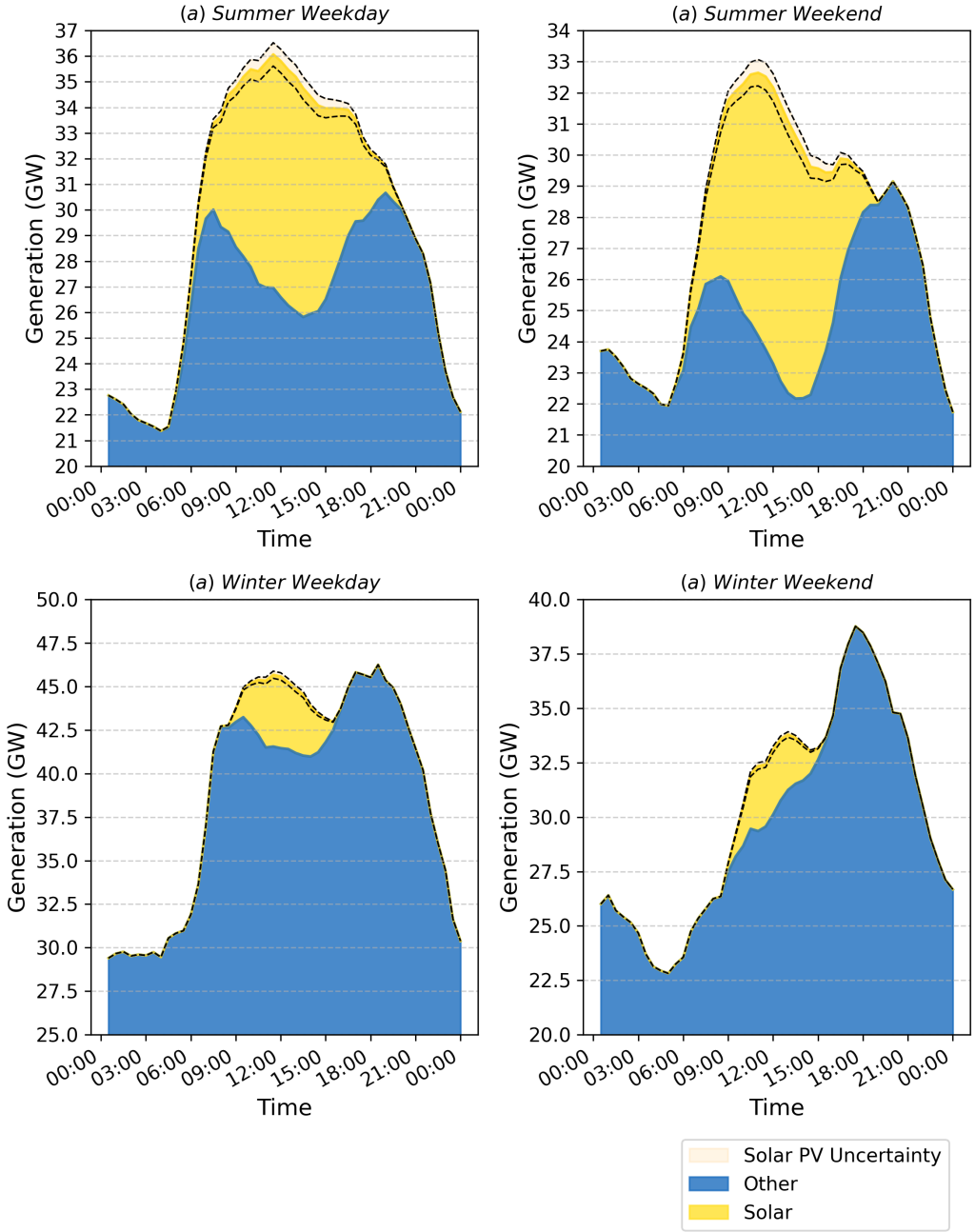


Figure 5.4: *Graphs showing the total electricity generation on the GB electricity grid. The blue are represents all the non-solar generation and the yellow and orange area represents the solar PV generation. Where the orange area denotes the uncertainty range for the solar PV output.*

maximised. On a sunny summer weekend day with 20 GW of demand across lunch time, as in figure 5.4. The $1\ GW$ uncertainty in the solar PV output estimate represents a 5% uncertainty in ESO's national demand forecast for the 2A, 2B, and 3C cardinal points. This uncertainty in the solar PV output data will result in increased error in day-ahead solar PV forecasts which ESO use to procure generation units in the advanced electricity markets. Consequently, ESO will have to call on more power systems at short notice within their role as the System Balancer. Until the electricity system is completely decarbonised much of this power production will come from dirty coal and gas generators which clearly is sub-optimal from both a cost and a carbon perspective.

In the winter, solar PV output is reduced and is more concentrated over the midday hours as shown in figure 5.4. Additionally, since demand is increased in winter compared with summer, the solar PV output makes up a less significant portion of national demand. Therefore, in winter the uncertainty in the solar PV output data is much less important than in the summer for ESO's demand profile.

5.5 Conclusion

Tot total error in the GB solar PV output estimates is calculated to be 5.1%. Significantly the capacity error (4.9%) dominates by an order of magnitude the yield error ($\leq 1\%$). Meaning that poor knowledge of the installed base of PV systems limits the accuracy of the PV output data and services that rely on it such as net system demand forecasting. Consequently, higher volumes of backup power production capacity are needed to ensure security of electricity supply. Since this backup capacity must have short ramp times it is often provided by lignite coal, CCGT, and pumped-hydro [87] and this is sub-optimal from both a cost and carbon perspective.

Chapter 6

Thesis summary and future work

We really need to kick the carbon habit and stop making our energy from burning things. Climate change is also really important. You can wreck one rainforest then move, drain one area of resources and move onto another, but climate change is global.

– Sir David Attenborough

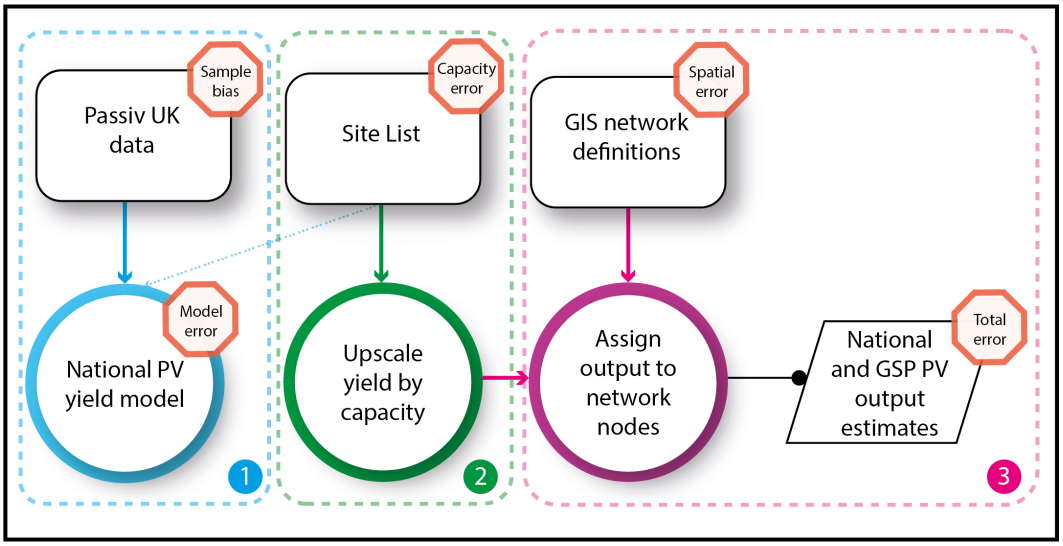


Figure 6.1: *The GB PV Live modelling approach.*

Thesis summary

This thesis has investigated the uncertainty associated with solar PV output estimates from national solar PV monitoring services, with a specific focus on Great Britain. The fundamental research question of this thesis was, what is the total uncertainty associated with the GB national solar PV estimates from the PV Live real time PV monitoring service. This thesis answered this question by breaking down the contributions from uncertainty in the national yield model and uncertainty in the knowledge of the installed base of PV systems and evaluated the total uncertainty in the GB national solar PV estimates to be $\pm 5.1\%$.

The thesis started with a review of 27 national solar PV monitoring services from 20 different countries. The review highlighted the fact that all across the world, techniques for monitoring national solar PV output follow two general steps. First the national PV yield is modelled, and then the modelled yield is scaled by an estimate of national solar PV capacity to calculate the solar PV output. This review highlighted the importance of solar PV capacity data internationally for accurate monitoring and forecasting of solar PV output.

Using the GB PV monitoring service as a case study, this thesis then went on to estimate the total uncertainty associated with the half-hourly solar PV output estimates for GB. To do the this problem was broken down into three distinct questions in order to identify the error associated with each of the three processes in the GB PV Live model as shown in figure 6.1. The questions were: what is the uncertainty associated with the PV yield model?; what

is the uncertainty associated with the national capacity data for Great Britain?; and finally, what is the total uncertainty associated with the national GB solar PV output estimates.

As shown in 6.1, the error in the PV yield model can be broken down into the statistical model error and the sample bias error. In chapter 3, the statistical model error for the national PV yield model was shown to be $\leq \pm 1\%$ when a sample size of 6000 was reached and the average yield estimates were shown to be unbiased for when compared with net metering data from a sample of 800 ground mount solar farms. Implying that differences in system performance caused by different system configurations (tilt, orientation, panel type, shading, etc.), cancel out for the national yield due to central limit theorem and the law of large numbers. Therefore, the error associated with modelling the PV yield is $\leq \pm 1\%$. In this analysis the effect of the capacity error on the accuracy of the PV yield model was ignored because it is insignificant compared to the other errors under investigation in this thesis.

In chapter 4, the national capacity estimate was shown to have a bias error of 0.4% and an uncertainty of $\pm 5\%$. Then in chapter 5, ignoring the small sample bias error and capacity error, the total uncertainty in the GB solar PV output estimates was calculated to be $\pm 5.1\%$. Solar PV forecasts typically have a $2-3\%$ error in their day-ahead estimates [49]. Therefore, this error represents a significant source of uncertainty for national solar PV forecasts and in the summer this uncertainty will negatively impact the accuracy of National Grid ESO's (ESO's) 2A, 2B, 3C, and 4A cardinal point demand forecasts. Importantly, the yield error ($\pm 1\%$) was shown to be far less important than the capacity error ($\pm 5\%$).

Future research

If I had the opportunity to repeat this thesis, there are two ways I would seek to improve it's quality. Firstly, I would improve the quantification of the sample bias error in chapter 3. Secondly, I would develop a more sophisticated national capacity model to facilitate performing a sensitivity analysis so that the contribution of each error source can be determined independently.

The sample bias error is important because it contributes towards the yield error and is critical in assessing the total error associated with solar PV output. To better quantify the sample bias error, an independent sample of domestic PV systems would be needed. The yield distribution of the independent sample should be compared with the Passiv sample, accounting for geographic and temporal variability. The research should seek to answer

the research question are the yield distributions statistically different from each other. This research would contribute towards a more robust understanding of the yield modelling error for estimates of national solar PV yield and is needed because the domestic sample in this thesis was only compared against a sample of ground-mount solar farm systems.

This primary novel contribution of this thesis is documentation of the 6 sources of uncertainty which affect the accuracy of recorded data for PV system capacity and the Monte Carlo model for quantifying their net effect on the national capacity estimate. The Monte Carlo approach used in the national capacity model did not investigate the relative impact of each different source of uncertainty. Therefore, it was not possible to draw any conclusions about which source of uncertainty is most significant. Given the opportunity to repeat this PhD I would develop a more sophisticated model of national capacity to allow a local sensitivity analysis to be performed. This research is needed to inform system operators and government officials on which areas they should focus on when keeping records of installed PV system capacity.

This thesis also presents some follow up research questions which academia and industry should work to solve. Firstly, how does uncertainty in solar PV output vary regionally and secondly, can national and regional capacity be derived independently from capacity registers.

Section 3.5 suggests the sample bias error is a more significant source of uncertainty at a regional level than at a national level. In section 3.5 the sample bias between the Passiv domestic sample and a sample of 800 solar farms was investigated. The yield estimates were shown to be biased at a regional level, however, when the national yield was evaluated the regional effects cancelled out and the yield estimates were unbiased. Hence, this result implies that the uncertainty in PV output may vary dynamically at a regional level and that sample bias error might be more significant than it is at a national level. Furthermore, this thesis has considered the capacity and yield errors to be independent. The capacity and yield errors are not independent if the reported distribution of installed capacity differs significantly from the true distribution. At a national level this assumption is valid because adding new solar PV systems to the capacity data set does little to change the overall capacity density. However, at a regional level and especially in regions with small volumes of installed solar PV capacity this assumption breaks down. Therefore, future research must also seek to understand the relationship between the error in the yield model and the error in capacity information to allow for a complete understanding of the regional uncertainty in PV output.

This thesis, for the first time, highlights the importance of accurate information on the installed base of solar PV capacity. Furthermore, the challenges associated with keeping an accurate and up to date list of all solar PV systems has been described. In future, as more and more domestic and commercial rooftop, and ground-mount solar is deployed, keeping a site list of all installed solar PV systems is only going to become more complicated. Therefore, future research must investigate methods for tracking solar PV capacity from alternative data sources. For example, dissaggregating smart meter data or using satellite imagery to identify installed solar PV systems.

Policy recommendations

The Monte Carlo capacity model in chapter 4 identified a $\pm 5\%$ uncertainty in the national capacity estimate. In the Monte Carlo model there were two general sources of uncertainty. One affecting the accuracy of the capacity when it is first reported, e.g. the transcription and unreported errors and another affecting the operational capacity estimate. For example, changes in the grid-connected capacity capable of producing power due to network outages, offline systems, decommissioned systems, or revisions to the capacity of a system since it was first installed.

Policy is needed for both facilitating better mechanisms for recording solar PV capacity and incentivising and regulating the use of any such register. Germany has done this by introducing the EEG register and withholding subsidy payments until assets are registered. This approach will work where subsidy is high, but new approaches are needed for tracking subsidy free solar capacity. Moving forward most solar PV capacity will be subsidy free so it is important to facilitate good policy which tackles these issues.

Whilst better registers will help to track the installed capacity more accurately, tracking the operational state of solar PV systems is a much more complicated task which will require a holistic approach to merge many disparate datasets so that the operational state of power flows from distributed assets can be tracked in real-time. The ElectraLink data in chapter 3 highlighted the challenges faced in this respect.

For example, ElectraLink manages the storage of MPAN readings for all 12 district network operators. Therefore, they have a database which records MPAN number, address and electricity meter readings. Since they do not have capacity information recorded for each MPAN they had to look this information up from other government data sources, namely the FIT and REPD databases.

There is no global key which identifies each PV system across every database. Therefore, it is only possible to cross-reference based on address. However, the address of a PV system is ambiguous and different databases will almost always have different addresses for the same system. Sometimes the farmhouse will be given, sometimes entrance to the field will be specified, and in some cases simply the postcode of the nearest road will be recorded and for rural locations postcodes can cover very large areas.

If the system can be identified based on address then the capacity for the whole system can be looked up from the government dataset which it has been cross referenced with. However, there might be multiple MPANs for one system and under the current system it is not possible to determine how much capacity is available at each MPAN. Additionally, it is common for large PV systems to be installed in parts with different first generation dates for each part of the system so there may also be multiple entries in the linked capacity register.

Once the capacity and location information for a system is known the next challenge is to understand where this power is located on the electricity system. To do this the DNOs have been working to map the areas served by their grid supply points [32]. These maps can be used to lookup the most likely Grid Supply Point given the location of the PV system. However, if the network deviates from its standard operating conditions, for example under maintenance or fault conditions, then this information might be incorrect. The only way to know which part of the network each system is connected to is to cross reference each PV system with DNO records and again the lack of unique system identifiers makes this process difficult.

To illustrate this point I will consider a hypothetical scenario relating to congestion on the distribution network. A domestic PV and battery system is connected to the LV network and signed up to an aggregator service. It is a sunny summer weekend day with low demand and a high PV forecast and a constraint is forecasted for the LV network. National Grid ESO calls on the aggregator to help manage the constraint so the battery is set to charge. However, a local solar farm is offline because the transformer which it is connected to is under maintenance. If there are many battery systems in the local area all operating in a similar manner, potentially from different aggregator service providers, then this scenario could result in a demand shortfall. Risking power cuts to domestic properties in the local area.

To manage such a scenario there are many different data sources, some confidential, which must come together to facilitate stable grid operation. The identity, capacity, loca-

tion and grid connection of the solar farm all must be known. The maintenance schedule of the transformer must be available. All of the domestic solar PV systems in the area must be identified with capacity, location and grid connection information. Any solar PV systems which are signed up to an aggregator service must be identified separately from non-aggregator PV systems which will probably involve cross referencing a general PV register with a register from each aggregator and black-listing the PV capacity in one of the datasets. The solar PV forecast data for each PV system must be available alongside forecasts of the local demand. All of this data will come from different sources and currently, no machine-readable information schema exists which could identify each of these electricity assets and power flows across these different datasets.

Conclusion

In conclusion, national PV monitoring services currently operate with an error which is comparable to the forecast error. However, whilst the PV forecast error is thoroughly discussed, the PV monitoring error is not widely understood. At regional and local level the precision of the information associated with PV system locations and network connections is insufficient for managing local constraints and balancing. Improved recording of local electricity generation and storage assets is needed to manage the future electricity system. In this regard, accurate system registers are required that facilitate the development of information schemas which allow assets to be identified across multiple disparate government and commercial datasets. If this does not happen then the significance of the total error in the PV output will increase as the PV fleet ages. Furthermore, if we cannot facilitate interoperability between different disparate energy datasets then similar errors will be introduced for other types of distributed energy resources. Consequently, the total error in the demand forecast will increase significantly, resulting in large costs to the consumer.

Bibliography

[1] TF Stocker et al. *IPCC, 2013: Summary for Policymakers*. Tech. rep. Cambridge.

[2] NASA. *Expert credibility in climate change*. July 2010. DOI: 10.1073/pnas.1003187107. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1003187107.

[3] Gunnar Luderer et al. "Residual fossil CO2 emissions in 1.5-2 °c pathways". In: *Nature Climate Change* 8.7 (July 2018), pp. 626–633. ISSN: 17586798. DOI: 10.1038/s41558-018-0198-6. URL: https://doi.org/10.1038/s41558-018-0198-6.

[4] Joeri Rogelj et al. "Scenarios towards limiting global mean temperature increase below 1.5 °c". In: *Nature Climate Change* 8.4 (Apr. 2018), pp. 325–332. ISSN: 17586798. DOI: 10.1038/s41558-018-0091-3. URL: https://doi.org/10.1038/s41558-018-0091-3.

[5] Arnulf Grubler et al. "A low energy demand scenario for meeting the 1.5 °c target and sustainable development goals without negative emission technologies". In: *Nature Energy* 3.6 (June 2018), pp. 515–527. ISSN: 20587546. DOI: 10.1038/s41560-018-0172-6. URL: https://doi.org/10.1038/s41560-018-0172-6.

[6] Joeri Rogelj et al. *Mitigation Pathways Compatible with 1.5°C in the Context of Sustainable Development*. Tech. rep. IPCC, 2018.

[7] Oliver Ruhnau et al. "Direct or indirect electrification? A review of heat generation and road transport decarbonisation scenarios for Germany 2050". In: *Energy* 166 (Jan. 2019), pp. 989–999. ISSN: 0360-5442. DOI: 10.1016/J.ENERGY.2018.10.114. URL: https://www-sciencedirect-com.sheffield.idm.oclc.org/science/article/pii/S0360544218321042.

[8] Jesse D Jenkins and Samuel Thernstrom. *Deep decarbonization of the electric power sector insights from recent literature*. Tech. rep. Energy Innivation Reform Project, 2017. URL: http://innovationreform.org/wp-content/upload

s/2017/03/EIRP-Deep-Decarb-Lit-Review-Jenkins-Thernstrom-March-2017.pdf.

[9] Elmar Kriegler et al. "The role of technology for achieving climate policy objectives: overview of the EMF 27 study on global technology and climate policy strategies". In: *Climatic Change* 123 (2014), pp. 353–367. DOI: 10.1007/s10584-013-0953-7.

[10] Geoffrey M Morrison et al. "Comparison of low-carbon pathways for California". In: (). DOI: 10.1007/s10584-015-1403-5.

[11] OFGEM. *Licensable activities*. URL: https://www.ofgem.gov.uk/licences-industry-codes-and-standards/licences/licensable-activities.

[12] Elexon. *What is the BSC?* URL: https://www.elexon.co.uk/knowledgebase/about-the-bsc/.

[13] National Grid ESO. *Connection and Use of System Code (CUSC)*. URL: https://www.nationalgrideso.com/industry-information/codes/connection-and-use-system-code-cusc.

[14] Electralink. *DCUSA*. URL: https://www.dcusa.co.uk/.

[15] National Grid ESO. *Grid Code (GC)*. URL: https://www.nationalgrideso.com/industry-information/codes/grid-code.

[16] Energy Networks Association. *DCode*. URL: http://www.dcode.org.uk/.

[17] GemServ. *Master Registration Agreement*. URL: https://www.mrasco.com/mra-products/master-registration-agreement/.

[18] GemServ. *Smart Energy Code*. URL: https://smartenergycodecompany.co.uk/.

[19] Energy Innovation Centre, Energy Systems Catapult, and Cornwall Insight. *An Introductory guide to the GB energy industry: Chapter 1, the importance of the GB energy sector*. Tech. rep. 2018.

[20] Department for Business Energy and Industrial Strategy and OFGEM. *Proposals for a Future System Operator role*. URL: https://www.gov.uk/government/consultations/proposals-for-a-future-system-operator-role.

[21] Elexon. *What is the balancing mechanism?* URL: https://www.elexon.co.uk/knowledgebase/what-is-the-balancing-mechanism/.

[22] Luis Boscán and Rahmatallah Poudineh. "Business Models for Power System Flexibility: New Actors, New Roles, New Rules". In: *Future of Utilities - Utilities of the Future: How Technological Innovations in Distributed Energy Resources Will Reshape the Electric Power Sector*. Elsevier Inc., Mar. 2016, pp. 363–382. ISBN: 9780128043202. DOI: 10.1016/B978-0-12-804249-6.00019-1.

[23] J. A.Peças Lopes et al. "Integrating distributed generation into electric power systems: A review of drivers, challenges and opportunities". In: *Electric Power Systems Research* 77.9 (July 2007), pp. 1189–1203. ISSN: 03787796. DOI: 10.1016/j.epsr.2006.08.016.

[24] Energy Innovation Centre, Energy Systems Catapult, and Cornwall Insight. *An introductory guide to the GB energy industry: Chapter 5, GB energy networks*. Tech. rep. 2018.

[25] OFGEM. *The GB electricity distribution network*. URL: https://www.ofgem.gov.uk/electricity/distribution-networks/gb-electricity-distribution-network.

[26] Ofgem. *Independent Distribution Network Operators*. URL: https://www.ofgem.gov.uk/electricity/distribution-networks/connections-and-competition/independent-distribution-network-operators.

[27] Jeremy Caplin. *Demand Forecasting*. Tech. rep. National Grid.

[28] Solar Power Portal. *Solar forecasting accuracy improves by 33% through AI machine learning*. July 2019. URL: https://www.solarpowerportal.co.uk/ne

ws/solar_forecasting_accuracy_improves_by_33_through_ai_machine_learning.

[29] Iain Staffell et al. "Electric Insights Quarterly: January to March 2021". In: *Drax Electric Insights* (2017). URL: https://s3-eu-west-1.amazonaws.com/16058-drax-cms-production/documents/170811_Drax_Q2_Report_06.pdf.

[30] National Grid Electricity System Operator and National Grid Electricity Transmission. *PV Monitoring Phase 2*. URL: https://www.smarternetworks.org/project/nia_nget0170.

[31] Sheffield Solar. *PV Live*. 2018. URL: https://www.solar.sheffield.ac.uk/pvlive/.

[32] National Grid ESO. *GIS Boundaries for GB Grid Supply Points*. URL: https://data.nationalgrideso.com/system/gis-boundaries-for-gb-grid-supply-points#.

[33] Sheffield Solar. *PV-Deployment-Tracker*. URL: https://github.com/SheffieldSolar/PV-Deployment-Tracker.

[34] Samuel Bimenyimana. "Traditional Vs Smart Electricity Metering Systems: A Brief Overview". In: *Journal of Marketing and Consumer Research* 46.June (2018), pp. 1–7.

[35] George Casella, Stephen Fienberg, and Ingram Olkin. *An Introduction to Statistical Learning*. 2013, p. 618. ISBN: 9780387781884. DOI: 10.1016/j.peva.2007.06.006. URL: http://books.google.com/books?id=9tv0taI8l6YC.

[36] Frank P.M. Kreuwel et al. "Analysis of high frequency photovoltaic solar energy fluctuations". In: *Solar Energy* 206 (Aug. 2020), pp. 381–389. ISSN: 0038092X. DOI: 10.1016/j.solener.2020.05.093.

[37] R. Z. Wang and T. S. Ge. *Advances in Solar Heating and Cooling*. Elsevier Inc., June 2016, pp. 1–577. ISBN: 9780081003022. DOI: 10.1016/C2014-0-03661-1.

[38] Soteris A. Kalogirou. *Solar Energy Engineering: Processes and Systems: Second Edition*. Elsevier Inc., 2014, pp. 1–819. ISBN: 9780123972705. DOI: 10.1016/C2011-0-07038-2.

[39] Francesco Calise et al. *Solar Hydrogen Production: Processes, Systems and Technologies*. Elsevier, Aug. 2019, pp. 1–560. ISBN: 9780128148549. DOI: 10.1016/C2017-0-02289-9.

[40] Jan Kleissl. *Solar Energy Forecasting and Resource Assessment*. Elsevier Inc., 2013, pp. 1–416. ISBN: 9780123971777. DOI: 10.1016/C2011-0-07022-9.

[41] Benjamin Y.H. Liu and Richard C. Jordan. "The interrelationship and characteristic distribution of direct, diffuse and total solar radiation". In: *Solar Energy* 4.3 (July 1960), pp. 1–19. ISSN: 0038-092X. DOI: 10.1016/0038-092X(60)90062-1. URL: https://www.sciencedirect.com/science/article/pii/0038092X60900621.

[42] N. A. Engerer and F. P. Mills. "KPV: A clear-sky index for photovoltaics". In: *Solar Energy* 105 (July 2014), pp. 679–693. ISSN: 0038092X. DOI: 10.1016/j.solener.2014.04.019.

[43] F. Antonanzas-Torres et al. *Clear sky solar irradiance models: A review of seventy models*. June 2019. DOI: 10.1016/j.rser.2019.02.032. URL: https://doi.org/10.1016/j.rser.2019.02.032.

[44] R.W. Mueller et al. "Rethinking satellite-based solar irradiance modelling: The SOLIS clear-sky module". In: *Remote Sensing of Environment* 91.2 (May 2004), pp. 160–174. ISSN: 0034-4257. DOI: 10.1016/J.RSE.2004.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0034425704000690.

[45] R. W. Mueller et al. "The CM-SAF operational scheme for the satellite based retrieval of solar surface irradiance - A LUT based eigenvector hybrid approach". In: *Remote Sensing of Environment* 113.5 (May 2009), pp. 1012–1024. ISSN: 00344257. DOI: 10.1016/j.rse.2009.01.012.

[46] Pierre Ineichen. "A broadband simplified version of the Solis clear sky model". In: *Solar Energy* 82.8 (Aug. 2008), pp. 758–762. ISSN: 0038-092X. DOI: 10.1016/J.SOLENER.2008.02.009. URL: https://www.sciencedirect.com/science/article/pii/S0038092X08000406.

[47] Christian A. Gueymard. "REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset". In: *Solar Energy* 82.3 (Mar. 2008), pp. 272–285. ISSN: 0038-092X. DOI: 10.1016/J.SOLENER.2007.04.008. URL: https://www.sciencedirect.com/science/article/pii/S0038092X07000990.

[48] Christelle Rigollier, Olivier Bauer, and Lucien Wald. "On the clear sky model of the ESRA - European Solar Radiation Atlas - With respect to the Heliosat method". In: *Solar Energy* 68.1 (Jan. 2000), pp. 33–48. ISSN: 0038092X. DOI: 10.1016/S0038-092X(99)00055-9.

[49] Utpal Kumar Das et al. *Forecasting of photovoltaic power generation and model optimization: A review*. Jan. 2018. DOI: 10.1016/j.rser.2017.08.017.

[50] Richard Perez et al. "Short-term irradiance variability: Preliminary estimation of station pair correlation as a function of distance". In: *Solar Energy* 86.8 (Aug. 2012), pp. 2170–2176. ISSN: 0038-092X. DOI: 10.1016/J.SOLENER.2012.02.027. URL: https://www.sciencedirect.com/science/article/pii/S0038092X12000928.

[51] Thomas E. Hoff and Richard Perez. "Modeling PV fleet output variability". In: *Solar Energy* 86.8 (Aug. 2012), pp. 2177–2189. ISSN: 0038-092X. DOI: 10.1016/J.SOLENER.2011.11.005. URL: https://www.sciencedirect.com/science/article/pii/S0038092X11004154.

[52] Thomas E. Hoff and Richard Perez. "Quantifying PV power Output Variability". In: *Solar Energy* 84.10 (Oct. 2010), pp. 1782–1793. ISSN: 0038-092X. DOI: 10.1016/J.SOLENER.2010.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0038092X10002380.

[53] Ameena Saad Al-Sumaiti et al. “Stochastic PV model for power system planning applications”. In: *IET Renewable Power Generation* 13.16 (Dec. 2019), pp. 3168–3179. ISSN: 17521424. DOI: 10.1049/iet-rpg.2019.0345.

[54] H. Suehrcke and P. G. McCormick. “Solar radiation utilizability”. In: *Solar Energy* 43.6 (Jan. 1989), pp. 339–345. ISSN: 0038092X. DOI: 10.1016/0038-092X(89)90104-7.

[55] M. Jurado, J. M. Caridad, and V. Ruiz. “Statistical distribution of the clearness index with radiation data integrated over five minute intervals”. In: *Solar Energy* 55.6 (Dec. 1995), pp. 469–473. ISSN: 0038092X. DOI: 10.1016/0038-092X(95)00067-2.

[56] A. Skartveit and J. A. Olseth. “The probability density and autocorrelation of short-term global and beam irradiance”. In: *Solar Energy* 49.6 (Dec. 1992), pp. 477–487. ISSN: 0038092X. DOI: 10.1016/0038-092X(92)90155-4.

[57] Richard Perez et al. “Parameterization of site-specific short-term irradiance variability”. In: *Solar Energy* 85.7 (July 2011), pp. 1343–1353. ISSN: 0038-092X. DOI: 10.1016/J.SOLENER.2011.03.016. URL: https://www.sciencedirect.com/science/article/pii/S0038092X11000995.

[58] J. Tovar et al. “Dependence of one-minute global irradiance probability density distributions on hourly irradiation”. In: *Energy* 26.7 (July 2001), pp. 659–668. ISSN: 03605442. DOI: 10.1016/S0360-5442(01)00024-X.

[59] Achim Woyte, Ronnie Belmans, and Johan Nijs. “Fluctuations in instantaneous clearness index: Analysis and statistics”. In: *Solar Energy* 81.2 (Feb. 2007), pp. 195–206. ISSN: 0038092X. DOI: 10.1016/j.solener.2006.03.001.

[60] P. Ashwini Kumari and P. Geethanjali. “Parameter estimation for photovoltaic system under normal and partial shading conditions: A survey”. In: *Renewable and Sustainable Energy Reviews* 84.October 2017 (2018), pp. 1–11. ISSN: 18790690. DOI: 10.1016/j.rser.2017.10.051. URL: https://doi.org/10.1016/j.rser.2017.10.051.

[61] Björn Müller et al. “Yield predictions for photovoltaic power plants: empirical validation, recent advances and remaining uncertainties”. In: *PROGRESS IN PHOTOVOLTAICS: RESEARCH AND APPLICATIONS*. Vol. 29th. Amsterdam, 2015, pp. 1114–1129. DOI: 10.1002/pip.

[62] A. Hammer et al. “Short-term forecasting of solar radiation: a statistical approach using satellite data”. In: *Solar Energy* 67.1-3 (July 1999), pp. 139–150. ISSN: 0038-092X. DOI: 10.1016/S0038-092X(00)00038-4. URL: https://www.sciencedirect.com/science/article/pii/S0038092X00000384.

[63] Stefan Pfenninger and Iain Staffell. “Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data”. In: *Energy* 114 (Nov. 2016), pp. 1251–1265. ISSN: 0360-5442. DOI: 10.1016/J.ENERGY.2016.08.060. URL: https://www.sciencedirect.com/science/article/pii/S0360544216311744?via%3Dihub.

[64] Jamie M. Bright et al. “Improved satellite-derived PV power nowcasting using real-time power data from reference PV systems”. In: *Solar Energy* (Nov. 2017). ISSN: 0038-092X. URL: https://doi.org/10.1016/j.solener.2017.10.091https://www.sciencedirect.com/science/article/pii/S0038092X17309714.

[65] Kenji Otani, Jyunya Minowa, and Kosuke Kurokawa. “Study on areal solar irradiance for analyzing areally-totalized PV systems”. In: *Solar Energy Materials and Solar Cells* 47.1-4 (Oct. 1997), pp. 281–288. ISSN: 09270248. DOI: 10.1016/S0927-0248(97)00050-0.

[66] E. Wiemken et al. “Power characteristics of PV ensembles: experiences from the combined power production of 100 grid connected PV systems distributed over the area of Germany”. In: *Solar Energy* 70.6 (Jan. 2001), pp. 513–518. ISSN: 0038092X. DOI: 10.1016/s0038-092x(00)00146-8.

[67] Norihiro Kawasaki et al. “An evaluation method of the fluctuation characteristics of photovoltaic systems by using frequency analysis”. In: *Solar Energy Materials and Solar Cells* 90.18-19 (Nov. 2006), pp. 3356–3363. ISSN: 09270248. DOI: 10.1016/j.solmat.2006.02.034.

[68] Akinobu Murata, Hiroshi Yamaguchi, and Kenji Otani. "A method of estimating the output fluctuation of many photovoltaic power generation systems dispersed in a wide area". In: *Electrical Engineering in Japan* 166.4 (Mar. 2009), pp. 9–19. ISSN: 04247760. DOI: 10.1002/eej.20723. URL: http://doi.wiley.com/10.1002/eej.20723.

[69] Matthew Lave and Jan Kleissl. "Solar variability of four sites across the state of Colorado". In: *Renewable Energy* 35.12 (2010), pp. 2867–2873. ISSN: 09601481. DOI: 10.1016/j.renene.2010.05.013.

[70] Manajit Sengupta. "Measurement and modeling of solar and PV output variability". In: *40th ASES National Solar Conference 2011*. Vol. 1. January 2011. 2011, pp. 486–490. ISBN: 9781618394279.

[71] Marco Pierro et al. "Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical weather prediction data". In: *Solar Energy* 158 (Dec. 2017), pp. 1026–1038. ISSN: 0038092X. DOI: 10.1016/j.solener.2017.09.068.

[72] Antonio Luque and Steven Hegedus. *Photovoltaic Science Handbook of Photovoltaic Science*. 3rd Ed. 2011. ISBN: 9780470721698. DOI: 10.1002/9780470974704.

[73] Jamie Taylor et al. "Performance of Distributed Pv in the Uk: a Statistical Analysis of Over 7000 Systems". In: *31st European Photovoltaic Solar Energy Conference and Exhibition* 53.9 (2016), pp. 1689–1699. ISSN: 1098-6596. DOI: 10.13140/RG.2.1.2019.6568.

[74] Sven Killinger et al. "On the search for representative characteristics of PV systems: Data collection and analysis of PV system azimuth, tilt, capacity, yield and shading". In: *Solar Energy* 173.August (2018), pp. 1087–1106. ISSN: 0038092X. DOI: 10.1016/j.solener.2018.08.051.

[75] Nicholas A. Engerer and James Hansard. "Real-time simulations of 15,000+ distributed PV arrays at sub-grid level using the Regional PV Simulation System (RPSS)". In: *ISES Solar World Congress 2015, Conference Proceedings* December 2018 (2015), pp. 1589–1598. DOI: 10.18086/swc.2015.06.02.

[76] Veikko Schepel et al. "The Dutch PV portal 2.0: An online photovoltaic performance modeling environment for the Netherlands". In: *Renewable Energy* 154 (July 2020), pp. 175–186. ISSN: 18790682. DOI: 10.1016/j.renene.2019.11.033.

[77] Electralink. *About Us*. 2018.

[78] Laura Sandys et al. *A strategy for a Modern Digitalised Energy System Energy Data Taskforce report*. Tech. rep. BEIS, OFGEM, Innovate UK, 2019. URL: https://es.catapult.org.uk/wp-content/uploads/2019/06/Catapult-Energy-Data-Taskforce-Report-A4-v4AW-Digital.pdf.

[79] Bundesnetzagentur. *EEG register data and reference values for payment*. 2018. URL: https://www.bundesnetzagentur.de/EN/Areas/Energy/Companies/RenewableEnergy/Facts_Figures_EEG/Register_data_tariffs/EEG_registerdata_payments_node.html.

[80] Terna. *Renewable sources*. URL: https://www.terna.it/en/electric-system/dispatching/renewable-sources.

[81] BEIS. *Renewable Energy Planning Data*. Mar. 2021. URL: https://www.gov.uk/government/collections/renewable-energy-planning-data.

[82] BEIS. *Renewables Obligation: certificates and generation*. Mar. 2021. URL: https://data.gov.uk/dataset/1f0e6bc9-e37a-4f2b-b89e-dcb2d02bb1aa/renewables-obligation-certificates-and-generation.

[83] Solar Media. *Solar Media*. URL: https://solarmedia.co.uk/.

[84] TU Delft. *Dutch PV Portal*. URL: https://www.tudelft.nl/en/ewi/over-de-faculteit/afdelingen/electrical-sustainable-energy/photovoltaic-materials-and-devices/dutch-pv-portal.

[85] Australian PV Institute (APVI). *Solar Map, funded by the Australian Renewable Energy Agency*. URL: https://pv-map.apvi.org.au/live.

[86] OPSD. *Open Power System Data*. 2018. URL: https://open-power-system-data.org/background/.

[87] R. Ahmed et al. *A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization*. May 2020. DOI: 10.1016/j.rser.2020.109792.

[88] Sobrina Sobri, Sam Koohi-Kamali, and Nasrudin Abd Rahim. "Solar photovoltaic generation forecasting methods: A review". In: (2017). DOI: 10.1016/j.enconman.2017.11.019. URL: https://doi.org/10.1016/j.enconman.2017.11.019.

[89] EUMETSAT. *Rapid Scanning Service*. URL: https://www.eumetsat.int/rapid-scanning-service.

[90] Open Climate Fix. *Home*. URL: https://openclimatefix.org/.

[91] Ahmed Bilal Awan et al. "Comparative analysis of ground-mounted vs. rooftop photovoltaic systems optimized for interrow distance between parallel arrays". In: *Energies* 13.14 (2020). ISSN: 19961073. DOI: 10.3390/en13143639.

[92] Augustin McEvoy, Tom Markvart, and Luis Castaner. *Practical Handbook of Photovoltaics*. Elsevier Ltd, 2012. ISBN: 9780123859341. DOI: 10.1016/C2011-0-05723-X.

[93] Sameer Al-Dahidi et al. "Ensemble approach of optimized artificial neural networks for solar photovoltaic power prediction". In: *IEEE Access* 7 (2019), pp. 81741–81758. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2923905.

[94] C. Schwingshackl et al. "Wind effect on PV module temperature: Analysis of different techniques for an accurate estimation". In: *Energy Procedia*. Vol. 40. Elsevier Ltd, Jan. 2013, pp. 77–86. DOI: 10.1016/j.egypro.2013.08.010.

[95] Alberto Dolara, Sonia Leva, and Giampaolo Manzolini. "Comparison of different physical models for PV power output prediction". In: *Solar Energy* 119 (Sept. 2015), pp. 83–99. ISSN: 0038092X. DOI: 10.1016/j.solener.2015.06.017. URL: http://dx.doi.org/10.1016/j.solener.2015.06.017.

[96] SMA Solar. *PV electricity produced in Germany*. URL: https://www.sma.de/en/company/pv-electricity-produced-in-germany.html.

[97] Sebastian Schierenbeck et al. "Ein distanzbasiertes Hochrechnungsverfahren für die Einspeisung aus Photovoltaik". In: *Energiewirtschaftliche Tagesfragen* 60.12 (2010), pp. 60–64.

[98] Anastasios Golnas et al. "Performance assessment without pyranometers: Predicting energy output based on historical correlation". In: *Conference Record of the IEEE Photovoltaic Specialists Conference*. 2011, pp. 002006–002010. ISBN: 9781424499656. DOI: 10.1109/PVSC.2011.6186347.

[99] Vincent P.A. Lonij et al. "Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors". In: *Solar Energy* 97 (Nov. 2013), pp. 58–66. ISSN: 0038092X. DOI: 10.1016/j.solener.2013.08.002.

[100] Vincent P. Lonij et al. "Analysis of 80 rooftop PV systems in the Tucson, AZ area". In: *Conference Record of the IEEE Photovoltaic Specialists Conference*. 2012, pp. 549–553. ISBN: 9781467300643. DOI: 10.1109/PVSC.2012.6317674.

[101] Michael Till Beck et al. "Estimating photo-voltaic power supply without smart metering infrastructure". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8343 LNCS. Springer Verlag, 2014, pp. 25–39. ISBN: 9783642551482. DOI: 10.1007/978-3-642-55149-9{_}3. URL: https://link.springer.com/chapter/10.1007/978-3-642-55149-9_3.

[102] Y. M. Saint-Drenan et al. "Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method". In: *Solar Energy* 135 (Oct. 2016), pp. 536–550. ISSN: 0038092X. DOI: 10.1016/j.solener.2016.05.052.

[103] Shin ichi Inage. "Development of an advection model for solar forecasting based on ground data first report: Development and verification of a fundamental model". In: *Solar Energy* 153 (Sept. 2017), pp. 414–434. ISSN: 0038092X. DOI: 10.1016/j.solener.2017.05.019.

[104] Renewablesninja. *Renewablesninja*. URL: https://www.renewables.ninja/.

[105] TransnetBW. *ERNEUERBARE ENERGIEN: EINSPEISUNG FOTOVOLTAIK*. Tech. rep. URL: www.ise.fraunhofer.de,.

[106] Yves-Marie Saint-Drenan. "A Probabilistic Approach to the Estimation of Regional Photovoltaic Power Generation using Meteorological Data". PhD thesis. 2015.

[107] International Renewable Energy Agency (IRENA). *RENEWABLE CAPACITY STATISTICS 2021*. 2021. ISBN: 9789292603427. URL: www.irena.org.

[108] Elexon. *Copyright: Licence to use BMRS open data*. URL: https://www.elexon.co.uk/operations-settlement/bsc-central-services/balancing-mechanism-reporting-agent/copyright-licence-bmrs-data/.

[109] Elia. *Elia Open Data License*. URL: https://www.elia.be/en/grid-data/elia-open-data-license.

[110] Bundesnetzagentur. *Imprint*. URL: https://www.smard.de/en/impressum.

[111] Energinet. *CONDITIONS FOR USE OF DANISH PUBLICSECTOR DATA*. 2017. URL: https://www.energidataservice.dk/static/pdf/Conditions_for_use_of_Danish_public_sector_data-License_for_use_of_data_in_ED.pdf.

[112] RTE. *Legal notice*. URL: https://www.rte-france.com/en/legal-notice.

[113] Electric Power Company. *Photovoltaic Power Generation Performance Graph*. URL: http://agora.ex.nii.ac.jp/earthquake/201103-eastjapan/energy/electrical-japan/power-solar/.

[114] Fraunhofer ISE. *Installed power*. URL: https://www.energy-charts.de/power_inst.htm.

[115] Fraunhofer ISE. *Publishing Notes*. URL: https://energy-charts.info/publishing-notes.html?l=en&c=DE.

[116] TenneT. *Actual and forecast solar energy feed-in.* URL: https://www.tennet.eu/electricity-market/transparency-pages/transparency-germany/network-figures/actual-and-forecast-solar-energy-feed-in/.

[117] ENTSO-E. *ENTSO-E Transparency Platform.* URL: https://transparency.entsoe.eu/dashboard/show.

[118] Amprion. *Photovoltaic infeed.* URL: https://www.amprion.net/Grid-Data/Photovoltaic-Infeed/.

[119] 50Hertz (Elia). *Photovoltaics.* URL: https://www.50hertz.com/en/Transparency/GridData/Production/Photovoltaics.

[120] SMA Solar. *Corporate Information.* URL: https://www.sma.de/en/corporate-information.html.

[121] SMARD. *Market data visuals.* URL: https://www.smard.de/en/marktdaten?marketDataAttributes=%7B%22resolution%22:%22day%22,%22from%22:1612134000000,%22to%22:1614553199999,%22moduleIds%22:%5B1004068%5D,%22selectedCategory%22:null,%22activeChart%22:true,%22style%22:%22color%22,%22region%22:%22AT%22%7D.

[122] Bundesnetzagentur. *The Bundesnetzagentur's Duties.* URL: https://www.bundesnetzagentur.de/EN/General/Bundesnetzagentur/About/Functions/functions_node.html.

[123] California ISO. *Supply.* URL: http://www.caiso.com/TodaysOutlook/Pages/supply.aspx.

[124] Solar Energy Industries Association. *State Solar Spotilight.* Tech. rep. 2021. URL: https://www.thesolarfoundation.org/solar-jobs-census/states/.

[125] California ISO. *Key Statistics - February 2012.* URL: http://www.caiso.com/Documents/KeyStatistics-Feb2021.pdf.

[126] Terna. *Actual generation.* URL: https://www.terna.it/en/electric-system/transparency-report/actual-generation.

[127] Sheffield Solar. *PV Live API User Guide V3.* URL: https://docs.google.com/document/d/e/2PACX-1vTvanhq7tLEpHKdnGcI-fH8tCqqz52XiDUrpeyM9WFDMHkuTTid2wuI5x_hKcTqwJME0COcQXlZI81q/pub.

[128] Drax. *Electric Insights.* URL: https://www.electricinsights.co.uk/#/homepage?&_k=pw4q62.

[129] ELEXON. *Actual Aggregated Generation Per Type (B1620).* URL: https://www.bmreports.com/bmrs/?q=actgenration/actualaggregated.

[130] Red Eléctrica de España. *Generation.* URL: https://www.ree.es/en/datos/generation/generation-structure.

[131] RTE France. *eco2mix mix énergétique.* URL: https://www.rte-france.com/en/eco2mix/eco2mix-mix-energetique-en.

[132] Elia Belgium. *Solar power generation.* URL: https://www.elia.be/en/grid-data/power-generation/solar-pv-power-generation-data.

[133] CEPS. *All Data.* URL: https://www.ceps.cz/en/all-data#GenerationRES.

[134] Austrian Power Grid. *Where does electricity come from?* URL: https://www.apg.at/en/Where-does-elecitricity-come-from.

[135] Austrian Power Grid (APG). *Terms of Use.* URL: https://www.apg.at/en/About-us/Impressum.

[136] Polskie Sieci Elektroenergetyczne (PSE). *Generation of Wind Farms and Solar Farms.* URL: https://www.pse.pl/web/pse-eng/data/polish-power-system-operation/generation-in-wind-farms.

[137] Energinet. *ENERGI DATA SERVICE.* URL: https://www.energidataservice.dk/.

[138] REN. *Actual Generation*. URL: https://www.mercado.ren.pt/EN/Electr/MarketInfo/Gen/Pages/Actual.aspx.

[139] Mavir. *Generation Forecasts and Actual Data for Solar*. URL: https://www.mavir.hu/web/mavir-en/generation-forecasts-and-actual-data-for-solar.

[140] Plc. (SEPS) Slovak electricity transmission system. *Daily operational data*. URL: https://sepsas.sk/en/control-centre/daily-operational-data/.

[141] Elering. *Elering Live*. URL: https://dashboard.elering.ee/et.

[142] Gareth James et al. *An introduction to Statistical Learning*. Vol. 7. 10. 2000, pp. 995–1039. ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7.

[143] Henri Felten, Anne Kreutzmann, and Philippe Welter. "Increase in grid-connected pv system power in Germany". In: *4th World Conference on Photovoltaic Energy Conversion*. IEEE, 2006, pp. 2494–+. ISBN: 1424400163. URL: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=23&SID=F4MYNBrDxH1DKu9Lzeh&page=3&doc=21.

[144] Dan Stowell et al. "A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK". In: *Scientific Data* 7.1 (Dec. 2020), pp. 1–15. ISSN: 20524463. DOI: 10.1038/s41597-020-00739-0. URL: www.nature.com/scientificdata.

[145] National Grid. *Electricity System Operator Forward Plan*. Tech. rep. 2018. URL: www.nationalgrid.com/uk/.

[146] Jonathan Leloux et al. "Performance to Peers (P2P): A benchmark approach to fault detections applied to photovoltaic system fleets". In: *Solar Energy* 202 (May 2020), pp. 522–539. ISSN: 0038092X. DOI: 10.1016/j.solener.2020.03.015.

[147] Bas van Zuijlen et al. "Cost-optimal reliable power generation in a deep decarbonisation future". In: *Applied Energy* 253 (Nov. 2019), p. 113587. ISSN: 03062619. DOI: 10.1016/j.apenergy.2019.113587.

[148] Hannah C Bloomfield et al. "Characterizing the winter meteorological drivers of the European electricity system using targeted circulation types". In: *Meteorol Appl* (2020), p. 27. DOI: 10.1002/met.1858.

[149] Andrew N. Commin et al. "The influence of the North Atlantic Oscillation on diverse renewable generation in Scotland". In: *Applied Energy* 205 (Nov. 2017), pp. 855–867. ISSN: 03062619. DOI: 10.1016/j.apenergy.2017.08.126.

[150] Australian Photovoltaic Institute. *Mapping Australian Photovoltaic installations*. URL: https://pv-map.apvi.org.au/historical.

[151] OFGEM. *Feed-in Tariffs (FIT)*. URL: https://www.ofgem.gov.uk/environmental-programmes/fit.

[152] OFGEM. *DCP350 – Creation of Embedded Capacity Registers*. URL: https://www.ofgem.gov.uk/publications-and-updates/dcp350-creation-embedded-capacity-registers.

[153] D.R. Kiran. "Failure Modes and Effects Analysis". In: *Total Quality Management*. Elsevier, Jan. 2017, pp. 373–389. DOI: 10.1016/B978-0-12-811035-5.00026-X. URL: https://linkinghub.elsevier.com/retrieve/pii/B978012811035500026X.

[154] Amalia Pizarro-Alonso, Hans Ravn, and M. Münster. "Uncertainties towards a fossil-free system with high integration of wind energy in long-term planning". In: *Applied Energy* 253 (Nov. 2019), p. 113528. ISSN: 03062619. DOI: 10.1016/j.apenergy.2019.113528.

[155] Mi Dong et al. "Uncertainty and global sensitivity analysis of levelized cost of energy in wind power generation". In: *Energy Conversion and Management* 229 (Feb. 2021), p. 113781. ISSN: 01968904. DOI: 10.1016/j.enconman.2020.113781.

[156] GOV. *Department for Business, Energy & Industrial Strategy*. URL: https://www.gov.uk/government/organisations/department-for-business-energy-and-industrial-strategy.

[157] GOV. *Solar photovoltaics deployment*. URL: https://www.gov.uk/government/statistics/solar-photovoltaics-deployment.

[158] Alan Hope. "Aerial photos reveal more than 5,000 illegal solar installations". In: *The Brussels Times* (May 2021). URL: https://www.brusselstimes.com/belgium/174213/consultative-committee-agenda-relaxations-coronavirus-measures-restaurants-bars-cafes-social-bubbles-events-weddings-funeral-alexander-de-croo/.

[159] BRE and National Solar Centre. *Fire and Solar PV Systems-Literature Review*. Tech. rep. Cornwall: BRE National Solar Centre, 2017.

[160] S.K. Firth, K.J. Lomas, and S.J. Rees. "A simple model of PV system performance and its use in fault detection". In: *Solar Energy* 84 (2009), pp. 624–635.

[161] Foresight Solar Group. *AUDITED CONSOLIDATED ANNUAL REPORT AND FINANCIAL STATEMENTS*. Tech. rep. 2016. URL: https://fsfl.foresightgroup.eu/investor-relations/publications/annual-results/.

[162] NextEnergy Solar Fund Limited. *Annual Report and Audited Financial Statements for the year ended 31 March 2019*. Tech. rep. 2019. URL: https://www.nextenergysolarfund.com/.

[163] Peter Bauer et al. "The digital revolution of Earth-system science". In: *Nature Computational Science* 1.2 (Feb. 2021), pp. 104–113. ISSN: 2662-8457. DOI: 10.1038/s43588-021-00023-0. URL: http://www.nature.com/articles/s43588-021-00023-0.