Applied Energy 262 (2020) 114404

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

# Big data mining for the estimation of hourly rooftop photovoltaic potential and its uncertainty

Alina Walch[a,*], Roberto Castello[a], Nahid Mohajeri[b], Jean-Louis Scartezzini[a]

[a] Solar Energy and Building Physics Laboratory, Ecole Polytechnique Fédérale de Lausanne, Switzerland
[b] Environmental Change Institute, University of Oxford, Oxford, United Kingdom

## HIGHLIGHTS

- Hourly profiles of photovoltaic potential are estimated for 9.6 M Swiss rooftops.
- The data mining approach combines Machine Learning and Geographic Information Systems.
- Uncertainties are quantified and propagated throughout all stages of the estimation.
- Switzerland's annual rooftop PV potential is estimated at 24 ± 9 TWh.
- It may cover up to 43% of the national electricity demand (in 2018).

## ARTICLE INFO

*Keywords:*
Rooftop photovoltaic potential
Spatio-temporal modelling
Big data mining
Uncertainty estimation
Machine Learning

## ABSTRACT

The large-scale deployment of photovoltaics (PV) on building rooftops can play a significant role in the transition to a low-carbon energy system. To date, the lack of high-resolution building and environmental data and the large uncertainties related to existing processing methods impede the accurate estimation of large-scale rooftop PV potentials. To address this gap, we developed a methodology that combines Machine Learning algorithms, Geographic Information Systems and physical models to estimate the technical PV potential for individual roof surfaces at hourly temporal resolution. We further estimate the uncertainties related to each step of the potential assessment and combine them to quantify the uncertainty on the final PV potential. The methodology is applied to 9.6 million rooftops in Switzerland and can be transferred to any large region or country with sufficient available data. Our results suggest that 55% of the total Swiss roof surface is available for the installation of PV panels, yielding an annual technical rooftop PV potential of 24 ± 9 TWh. This could meet more than 40% of Switzerland's current annual electricity demand. The presented method for an hourly rooftop PV potential and uncertainty estimation can be applied to the large-scale assessment of future energy systems with decentralised electricity grids. The results can be used to propose effective policies for the integration of rooftop photovoltaics in the built environment.

## 1. Introduction

The decarbonisation of the energy system plays an important role in fulfilling the ambitious emission targets set by the Paris Agreement [1]. In this context, the large-scale deployment of rooftop-mounted photovoltaics (RPV) has attracted increasing attention in recent years [2]. A quantitative assessment of the potential electricity generation from RPV is essential to formulate effective incentive policies for their integration in the built environment. This requires accurate input data at a high spatial and temporal resolution in order to characterize regional differences and to assess the seasonal and intra-day variation of the potential generation [3]. In addition, the analysis of RPV potential involves several uncertainties, which need to be quantified to facilitate interpretation of the results for policy making.

Currently, there is no methodology that estimates the large-scale RPV potential at a high spatio-temporal resolution and also addresses the systematic propagation of uncertainties arising from the modelling process. This paper contributes to fill this gap by adapting state of the art methods for the assessment of RPV potential in order to quantify and combine different sources of uncertainty. We further use Machine Learning to incorporate information that is only available in parts of the study region of Switzerland.

* Corresponding author.
*E-mail address:* alina.walch@epfl.ch (A. Walch).

https://doi.org/10.1016/j.apenergy.2019.114404
Received 13 August 2019; Received in revised form 22 November 2019; Accepted 14 December 2019

Constrained by the availability of building and environmental data at a high spatial resolution, most existing studies on RPV potential are carried out at district or city scale only, including case studies in the US [4], Canada [5], Portugal [6] or Germany [7]. Few studies analyse an entire region or country, for example in the US [8,9], Spain [10], or Saudia Arabia [11]. Many PV assessment studies use monthly or yearly solar radiation data in order to derive a large-scale PV potential [12]. This data is used to quantify the available area to install PV [13] and to discuss the economic feasibility of PV scenarios [14]. While these granularities provide a relatively accurate estimation of the annual RPV potential, an hourly or higher temporal resolution is needed to assess the intra-day variation of the generation. Such temporal resolutions are used at the scale of cities or municipalities [15,16] and serve to validate the estimations against measurements [17], to compare PV technologies [18], or to simulate energy systems with high shares of PV [19]. However, an hourly resolution is used very rarely at regional or national scale, in which case the available area to install PV is not addressed [20].

Reasons for the lack of national-scale studies at an hourly temporal resolution are the computational challenges associated with the processing of the required input datasets as well as the handling of missing data and data that is not available in the entire study region. To address these challenges, state of the art data processing techniques for estimating RPV potentials include physical models, geographic information systems (GIS), image processing and Machine Learning (ML). Physical models are used to compute the solar radiation on tilted surfaces, as well as the module and inverter efficiencies [21]. GIS is applied to estimate shading effects and the sky view factor [22,23]. The available area for installing PV is estimated in some studies using image recognition [24,25], while other methods use ML [26,27]. The recent computational and methodological advances enable the integration of all mentioned aspects in high-resolution PV assessments at the national scale. Missing data, in particular for predicting the available area for installing PV, is typically handled using constant coefficients and expert knowledge [28] or sampling techniques [29]. Data with partial spatial coverage, i.e. information that is available only in parts of the study area, is successfully used in [26,27] by applying ML. This method can be further improved by using different urban features and a larger training dataset.

To date, little research has quantified uncertainties for RPV potential assessments. Some studies address the uncertainties related to photovoltaic yield predictions for individual case studies [30,31]. In large-scale studies, confidence intervals are used in [20] to quantify the uncertainty related to the solar radiation, while they are used in [27] to assess the available area for installing PV. The combination of different sources of uncertainty has been addressed through a scenario-based analysis for a local case study [32] or by means of a sensitivity analysis [33]. Izquierdo et al. [34] provide a statistical propagation of uncertainty, which however focuses only on the available area for PV installation. To the best of our knowledge, currently no methodology exists which quantifies and combines different sources of uncertainty related to the solar radiation and the available area that yields an uncertainty estimate on the technical RPV potential.

Our big data mining approach contributes to the existing literature by providing a methodology for a large-scale RPV potential and uncertainty estimation in hourly temporal resolution and a spatial resolution of individual roof surfaces. For this purpose, we combine state of the art physical models and GIS processing techniques with ML in order to quantify (i) the spatio-temporal variation of the horizontal radiation, (ii) the effects of surrounding trees and buildings on roof shading and the sky view factor, (iii) the impact of roof geometry and roof superstructures, for example dormers and chimneys, on the available area for installing PV panels, and (iv) the temperature dependence of the PV module efficiency. We further present a systematic quantification of the uncertainties by treating each variable involved in the modelling process as a randomly distributed variable. This allows to combine the uncertainties using their statistical distribution in order to obtain a total uncertainty on the technical potential estimate. The application of our method to 3.7 million buildings in Switzerland results in the first national-scale dataset of PV potential and uncertainty in hourly temporal resolution for individual rooftops.

The paper is structured as follows. Section 2 introduces the datasets used in the study. Section 3 describes the methodology for the computation of the RPV potential and the uncertainty propagation. Section 4 presents the results of the individual processing steps as well as the final technical RPV potential. A sensitivity analysis is performed to identify the most impactful steps of the estimation process. Section 5 discusses the methodological and practical contribution of the work and outlines its limitations and applications. Section 6 presents the conclusions and gives an outlook to future applications of the developed method.

## 2. Data

Using big data mining techniques for the estimation of large-scale RPV potential requires the availability of accurate and high-resolution environmental and building datasets. Our approach combines large sets of meteorological data, building data and digital surface models at different resolutions and spatial coverage. Switzerland has been selected as case study area due to its high data availability.

### 2.1. Meteorological data

We use four types of meteorological data, namely hourly global, direct solar horizontal radiation (in W/m$^2$), daily surface reflectance (albedo), and daily maximum temperature (in °C), recorded during 12 years from 2004 to 2015. The specifications of all datasets are summarized in Table 1. The solar radiation and albedo data is derived from Meteosat Second Generation (MSG) satellite observations using the Heliomont algorithm [35]. Satellite data is preferred over data from measurement stations as it provides a better spatial coverage with an increased resolution, it has a very low missing data ratio (<1%) and it shows a negligible bias [36]. The daily maximum temperature data is the result of a gridded interpolation of near-surface air temperature measurements with errors below 1°C at urban altitudes [37]. As the PV panel efficiency decreases with temperature, using the daily maximum temperature corresponds to the least optimal PV performance.

We average the 12 years of data to obtain an annual mean dataset, i.e. 8760 time steps for solar radiation and 365 time steps for albedo and temperature. This reduces the variability of the meteorological data

**Table 1**
Meteorological data used in the study.

| Data | Spatial res. | Time | Range | Source |
|---|---|---|---|---|
| Global horizontal radiation | 1.25 deg. min.[a] | hourly | 2004–2015 | MeteoSwiss [38] |
| Direct horizontal radiation | 1.25 deg. min.[a] | hourly | 2004–2015 | MeteoSwiss [38] |
| Surface albedo | 1.25 deg. min.[a] | daily | 2004–2015 | MeteoSwiss [38] |
| Maximum temperature | 1 km$^2$ | daily | 2004–2015 | MeteoSwiss [39] |

[a] Deg. min. denotes degree minutes on a longitude-latitude grid. 1.25 deg. min. corresponds to approximately (1.6 × 2.3) km$^2$.

**Table 2**
Building data used in the study.

| Data | Coverage | Spatial res. | Creation | Source |
|---|---|---|---|---|
| Roof surfaces | Switzerland | Rooftops | 2010–2016 | Sonnendach [22] |
| Register of buildings | Switzerland | Buildings | 2015 | SwissStat [40] |
| Superstructures | Geneva Canton | Rooftops | 2005–2011 | SITG [41] |

and allows the estimation of long-term PV potential without bias due to extreme meteorological events of a specific year.

### 2.2. Building data

The computation of the RPV potential is based on a national dataset of building roofs. It contains around 9.6 million vector polygons representing all roofs in Switzerland's 3D building cadastre (LOD 2) with the roof tilt, aspect (both in degree) and area (in $m^2$). This dataset is combined with the national register of buildings and dwellings (RBD), which gives information on the building's footprint, construction period, number of floors and building type. To account for obstructing objects on rooftops (superstructures), which impede the installation of PV panels, we use a dataset derived from detailed city GML data (LOD 4) available only in the Canton of Geneva. It contains vector polygons that represent superstructures such as dormers and chimneys and covers nearly 38,000 roofs. Table 2 summarizes the building data.

### 2.3. Digital elevation models

Besides meteorological and building data, we use a Digital Terrain Model (DTM) and Digital Surface Models (DSM) derived from Light Detection and Ranging (LiDAR) data, which are summarized in Table 3. From the DTM, terrain characteristics including altitude, slope and curvature have been computed by Robert et al. [42]. The DSM is used to account for the effects of landscape characteristics on the PV potential, which requires the most up-to-date and accurate LiDAR data. We hence use a DSM at national scale in $(2 \times 2)$ $m^2$ resolution, denoted as $DSM_{2m}$, as well as a higher-resolution DSM of $(0.5 \times 0.5)$ $m^2$ available for the Canton of Geneva only, denoted as $DSM_{50cm}$.

## 3. Methodology

To estimate the RPV potential and its uncertainty, we propose a methodology that combines Machine Learning, GIS processing and physical models for the treatment of large spatio-temporal datasets, such as those presented in Section 2. Uncertainties are quantified from the statistical distribution of the variables involved in the potential estimation in the form of standard deviations. They are assessed for various modelling steps and combined in order to obtain an uncertainty on the PV potential.

The method, illustrated in Fig. 1, is based entirely on open-source software and adopts an hierarchical approach used in several related studies [21,26,34]. Its steps include (i) the *physical potential*, driven by the horizontal solar radiation, (ii) the *geographic potential*, accounting for the impact of the built environment, and (iii) the *technical potential*, defined as the potential electricity generation. The variables involved in each stage of the model and their relationships are explained below.

**Table 3**
Digital elevation models used in the study.

| Data | Coverage | Spatial resolution | Creation | Source |
|---|---|---|---|---|
| DTM | Switzerland | $(2 \times 2)$ $m^2$ | 2010–2016 | SwissTopo [43] |
| $DSM_{2m}$ | Switzerland | $(2 \times 2)$ $m^2$ | 2000–2008 | SwissTopo [44] |
| $DSM_{50cm}$ | Geneva Canton | $(0.5 \times 0.5)$ $m^2$ | 2013 | SITG [45] |

The following subsections will detail the methodological contributions for the individual processing steps and the quantification of uncertainty.

The physical potential is defined as the horizontal solar radiation at the earth's surface ($G_h$) for each time step $t$. The $G_h$ is composed of a direct beam component ($G_B$) and a diffuse component ($G_D$) such that [46]:

$$G_h(t) = G_B(t) + G_D(t) \quad (1)$$

The horizontal radiation is computed at a monthly-mean hourly (MMH) temporal resolution and a spatial resolution of $(200 \times 200)$ $m^2$, which is chosen as a trade-off between topographic detail and computational complexity as suggested in [27,34]. Each MMH time step represents an average value at a given hour of each month, across all days of the month. This leads to 288 distinct time steps, i.e. 24 h for each of the 12 months. Using MMH values instead of 8760 hourly time steps allows to reduce the computational cost by a factor of 30, while preserving the daily and seasonal patterns of the average PV potential. Any deviation from the MMH values is however covered by their uncertainty.

The geographic potential accounts for the rooftop geometry, for superstructures, for shading effects and for the sky visibility. It is estimated for each roof surface considering the tilted radiation ($G_t$) and the available roof area for PV panel installation ($A_{PV}$) [27], such that:

$$G_t(t) = (1 - S_{sh}(t)) * G_{Bt}(t) + SVF * G_{Dt}(t) + G_{Rt}(t) \quad (2)$$

$$A_{PV} = A_t * C_{pv} * (1 - C_{sh}) \quad (3)$$

where $G_{Bt}$, $G_{Dt}$ and $G_{Rt}$ are direct, diffuse and reflected tilted radiation components, $SVF$ is the sky view factor, $A_t$ is the tilted roof area, $C_{pv}$ is referred to as the *panelled area coefficient* and $C_{sh}$ and $S_{sh}$ are referred to as the *shaded area coefficient* and the *hourly shading fraction* of the rooftop, respectively. The $C_{sh}$ represents the fraction of roof surface that is unshaded in less than 40% of the daylight hours and is hence unsuitable for PV installation (see Section 3.5). The $S_{sh}(t)$ denotes the portion of the remaining roof area $(1 - C_{sh})$ that is shaded at each time step $t$. We treat $C_{sh}$ and $C_{pv}$, the proportion of roof area available to install PV panels, as independent factors. They are separately computed and it is assumed that no relevant overlap exists between the two.

The technical potential ($E_{PV}$) is the electricity output of each roof surface. It is obtained from the geographic potential, the panel efficiency ($\eta_{PV}$) and the performance factor ($PF$), which accounts for inverter efficiency and other losses such as soiling or degradation, such that [27]:

$$E_{PV} = G_t(t) * A_{PV} * \eta_{PV}(t) * PF(t) \quad (4)$$

Table 4 summarizes the variables introduced above and the method used to model each of them.

### 3.1. Physical models

State of the art hourly physical models are used to (i) obtain the tilted radiation components ($G_{Bt}$, $G_{Dt}$, $G_{Rt}$) and (ii) to quantify the module and inverter efficiency. Both models, shown as orange boxes in Fig. 1, are summarized in Appendix A. We use the anisotropic Perez model [47] to estimate $G_{Dt}$, which is the overall most accurate diffuse radiation model [46,48]. The $G_{Rt}$ is computed using monthly mean albedo data (see Section 3.3). This allows to account for snow cover in high-altitude locations, where a large albedo significantly increases the PV production on steep roofs [49]. The efficiencies are calculated for each time step $t$ from the tilted radiation and the ambient temperature using the *PVWatts* model [50]. We use daily maximum temperature (see Table 1) to obtain a conservative estimate of the panel efficiency.

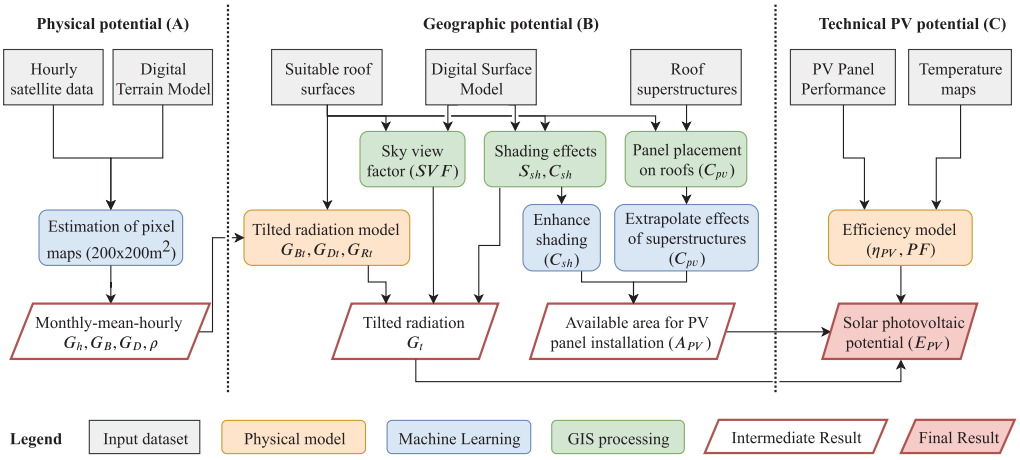



**Fig. 1.** Workflow for modelling of hourly solar PV potential.

**Table 4**
Summary of the parameters used in the estimation of technical RPV potential. The dimension refers to $P$ for pixels of $(200 \times 200)$ m$^2$, $R$ for roof surfaces and $t$ for time steps. PM denotes the use of physical models.

| | $G_{h,B,D}$ | $G_{Bt,Dt,Rt}$ | $S_{sh}$ | $SVF$ | $C_{sh}$ | $C_{pv}$ | $\eta_{PV}$, $PF$ |
|---|---|---|---|---|---|---|---|
| Dimension | $P \times t$ | $R \times t$ | $R \times t$ | $R$ | $R$ | $R$ | $R \times t$ |
| Unit | W/m$^2$ | W/m$^2$ | – | – | – | – | – |
| Method | ML | PM | GIS | GIS | GIS + ML | GIS + ML | PM |
| Section | 3.3 | 3.1 | 3.5 | 3.6 | 3.5 | 3.4 | 3.1 |

### 3.2. Machine Learning

Machine Learning is used here (i) to obtain the horizontal radiation and the albedo at the $(200 \times 200)$ m$^2$ output grid from lower-resolution satellite data, (ii) to estimate $C_{sh}$ and (iii) to account for the roof area covered by superstructures in the estimation of $C_{pv}$. We use supervised regression algorithms, which learn from a training set that contains pairs of inputs (features) and their known output values (targets). The models are then applied to predict unknown outputs for a new set of features.

We consider two ML algorithms, Random Forests (RF) [51] and Extreme Learning Machine Ensembles (ELM-E) [52,53]. Both are ensemble algorithms which predict a target variable by averaging the results of multiple estimators (decision trees/ELMs) that have been trained on random re-samples of the training data [54]. Ensembles are well-suited for the estimaiton of uncertainties, which give a useful indication of the model's accuracy [55]. The ELM-E is very efficient at modelling large datasets [56], so we use this algorithm to estimate the hourly solar radiation and the monthly albedo. Preliminary work has shown that the RF outperforms the ELM-E for higher-dimensional datasets with less training data. The RF is hence used to estimate $C_{sh}$ and $C_{pv}$.

To optimize the performance of each ML model, it is necessary to firstly select its features, and secondly to choose the parameters defining the structure of each model through hyper-parameter tuning. In the following sections, we will focus on the feature selection and the performance of the optimized ML models. Information regarding the tuning of the algorithms is provided in Appendix B.

### 3.3. Horizontal solar radiation and albedo

We propose an ML-based approach to estimate the horizontal radiation and the surface albedo for pixel maps of $(200 \times 200)$ m$^2$ resolution, as presented in the blue box in Fig. 1-A. The ML models yield the global ($G_h$) and direct ($G_B$) horizontal radiation for each MMH time step $t$ and the albedo ($\rho$) as monthly mean values. The diffuse radiation ($G_D$) is obtained from the estimated $G_h$ and $G_B$ by applying Eq. (1). The use of ML allows to model the complex spatial patterns while being more efficient than other spatial interpolation techniques such as kriging [57].

The targets for the ML models of $G_h$ and $G_B$ are the satellite-derived global and direct radiation, which are split into 12 monthly subsets. The initially considered set of features are the mean hour of each month, the geographic coordinates (x, y) as well as several terrain features including altitude (z), terrain slope and curvature [42]. A DTM is used to derive these terrain features for the satellite data and for the

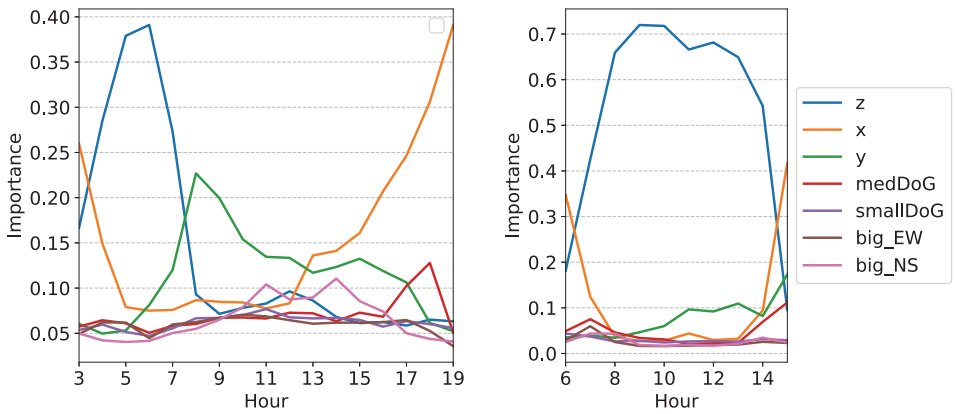



**Fig. 2.** (a) Feature importance per hour for $G_h$ for June and (b) December for six terrain features, namely altitude (z), longitude (x), latitude (y), medium-scale curvature (medDoG), small-scale curvature (smallDoG), terrain slope in east-west (big_EW) and north-south (big_NS) direction.

**Table 5**
Test mean-squared error (MSE) for the estimation of $G_h$ and $G_B$.

| MSE | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G_h$ | 0.10 | 0.07 | 0.07 | 0.06 | 0.06 | 0.04 | 0.06 | 0.05 | 0.12 | 0.10 | 0.10 | 0.06 |
| $G_B$ | 0.17 | 0.12 | 0.18 | 0.17 | 0.21 | 0.14 | 0.18 | 0.12 | 0.26 | 0.16 | 0.21 | 0.11 |

(200 × 200) $m^2$ output grid. To better model the spatio-temporal patterns of $G_h$ and $G_B$, we analyze the spatial features for each time step using the RF feature importance score [58]. Fig. 2 shows their pattern for the months of June and December. The feature importance exhibits a strong intra-day variation for x, y and z, which may be explained by the impact of the Swiss alpine terrain. All other features have an overall low importance throughout the day. Including these in the model does not improve its performance further. Four features are hence selected for the final model: hour, x, y, z. Fig. 2 refers to $G_h$, but $G_B$ shows a similar trend.

Separate ELM-E are trained for each month for $G_h$ and $G_B$. Their tuning procedure and the resulting hyper-parameter values of the models are provided in Appendix B. Table 5 shows the mean-squared error (MSE) between the targets and predictions for a random 20% of the satellite coordinates (test set), which are excluded from training. Each entry represents one trained ELM-E. The model performance is overall better for $G_h$ than for $G_B$, as it is harder to model $G_B$ from the given set of features. To estimate $\rho$ from the satellite-derived daily albedo measurements, we train one ELM-E with similar features as described above, namely x, y, z and month. The model is tuned in a similar fashion as that of $G_h$ and $G_B$ and yields a test MSE of 0.10, which lies in the range of the values obtained for $G_h$.

### 3.4. Available area for PV panel installation

The *panelled area coefficient* $C_{pv}$ describes the available roof area for PV installation considering the roof geometry and superstructures. To compute $C_{pv}$, we use a geospatial algorithm in combination with ML, as represented by the green and blue box yielding $C_{pv}$ in Fig. 1-B. The geospatial algorithm, detailed in Appendix C, virtually installs PV panels by projecting rectangular polygons onto the tilted roofs as shown in Fig. 3. The $C_{pv}$ is then obtained from the number of installed panels. The panels are installed in both horizontal and vertical alignments, as no alignment has technical advantages over the other and both are widespread. The configuration with a higher number of panels is selected for each roof.

The data on roof superstructures is however only available in one of Switzerland's 26 cantons (Geneva). We hence train an ML model to estimate the change in $C_{pv}$ due to the area covered by superstructures, which is applied to the rooftops of the remaining 25 cantons. In our case, the training of the ML model is performed using the rooftop dataset with superstructure information in the Canton of Geneva (see Section 2). The fraction of their area on which virtual panels are installed provides the target for the ML algorithm ($C_{pv}^T$). In addition, we define $C_{pv}^F$ as the fraction of the area on which virtual panels may be installed if no superstructures are removed from the roof polygons. This $C_{pv}^F$ is one feature of the ML model. The full set of the considered features is listed in Table 6. Fig. 4 shows the feature importance for all inputs. While the top six features have the highest importance, the complete set of features is used for modelling $C_{pv}$ as this has been found to improve the estimation precision.

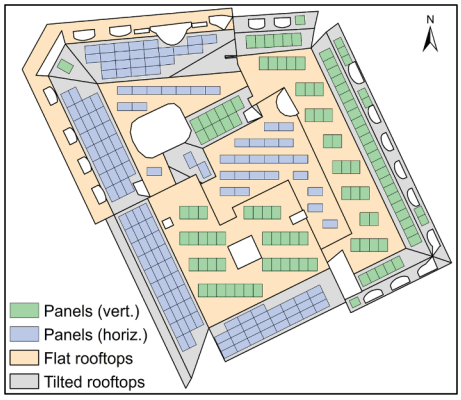



**Fig. 3.** Output of the virtual installation of PV panels after removing roof superstructures. The best configuration of vertically and horizontally oriented panels is selected. Panels on flat roofs are placed in south-facing rows and tilted at an optimal angle of 30°.

**Table 6**
Features considered in the ML models to estimate $C_{pv}$ and $C_{sh}$. [1]The building density is computed as the number of building coordinates within a 100 m radius of each roof. [2]The roof perimeter, shape index (perimeter per area) and vertex count are derived from the polygon data.

| Computed features | Roof features | Building features |
|---|---|---|
| $C_{pv}^F$/ $C_{sh}^F$ | Tilted area | Footprint area |
| Panel tilt | Tilt angle | Building type |
| (not used for $C_{sh}$) | Aspect angle | Construction period |
| Build. density[1] | Perimeter[2] | Number of floors |
| | Shape index[2] | |
| | Vertex count[2] | |

Table 7 reports the residuals between $C_{pv}^F$ and $C_{pv}^T$ (baseline), as well as the cross-validation error between the estimated $C_{pv}$ and the target $C_{pv}^T$ in the case of additive bias correction (MBE+) and for using the tuned RF model (see Appendix B for details). The bias correction is performed by adding the mean bias error (MBE) between $C_{pv}^F$ and $C_{pv}^T$ to all samples. We compare the root mean squared error (RMSE), mean absolute error (MAE), MBE and the $R^2$ coefficient of determination. All methods remove the negative baseline bias of 17%. The RF outperforms the bias correction, as it achieves lower errors and a higher $R^2$. It is hence used to estimate the $C_{pv}$ used in Eq. (3).

### 3.5. Shading effects

To quantify shading effects from surrounding buildings and trees, we use a shadow casting approach [20–23]. In contrast to the existing studies, we account for the shading effects in two ways: First, strong shading effects may render parts of a roof unsuitable for installing PV.

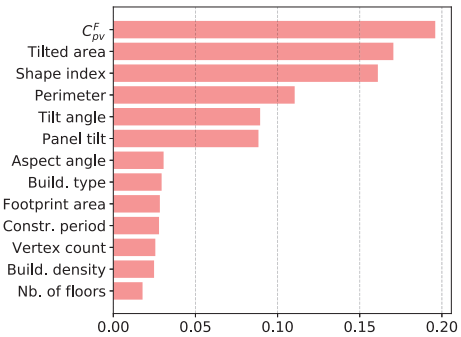



**Fig. 4.** Feature importance of all features considered for the estimation of $C_{pv}$.

**Table 7**
Errors for estimating $C_{pv}$. We compare the residuals between $C_{pv}^{F}$ and $C_{pv}^{T}$ (baseline) with the cross-validation error between the estimated $C_{pv}$ and the target $C_{pv}^{T}$ in the case of bias correction (MBE+) and for using the Random Forest model (RF).

| | RMSE | MAE | MBE | $R^2$ |
|---|---|---|---|---|
| Baseline | 0.23 | 0.17 | −0.17 | −0.10 |
| MBE+ | 0.15 | 0.12 | 0.00 | 0.52 |
| RF | 0.12 | 0.09 | 0.00 | 0.69 |

Second, some of the suitable area may be shaded at particular hours and reduce the PV potential for these hours. We hence propose to distinguish between the *shaded area coefficient* ($C_{sh}$), the fraction of roof area which is unsuitable due to shading, and the *hourly shading fraction* ($S_{sh}(t)$), which is computed for each time step as the shaded portion of the remaining roof.

The shadow casting approach (green box in Fig. 1-B) models shadows on building roofs for each pixel of a DSM at a given sun position. It is used to produce hourly shadow maps for the representative day of each month (close to day 15 [23]) and hence yields results at the same temporal resolution as the solar radiation. Fig. 5a shows an example of these maps. They are binary maps, with a 0 (dark areas) representing a cast shadow and a 1 (yellow areas) indicating direct sun exposure. The shadow maps are averaged for all daylight hours, resulting in the mean illumination map shown in Fig. 5b. Its values represent the percentage of time steps for which each pixel is unshaded.

All areas with a mean illumination lower than the average value across all roofs (red areas) are considered unsuitable. We compute this average value as 42.7% and hence use a rounded threshold of 40% mean illumination to exclude unsuitable areas. Dividing the red areas in Fig. 5b by the total roof area yields $C_{sh}$, while $S_{sh}$ is the fraction of zeros in Fig. 5a on the suitable (non-red) surfaces at each time $t$. The GIS algorithm for the shading effects is given in Appendix C.

Two datasets are available to compute $C_{sh}$ and $S_{sh}$: the national $DSM_{2m}$ and the $DSM_{50cm}$ for the canton of Geneva (see Table 3). Our method aims at enhancing the national-scale $C_{sh}$ and $S_{sh}$ from the $DSM_{2m}$ based on knowledge extracted from the $DSM_{50cm}$, which is more recent and detailed. For this task, we denote the values extracted from the $DSM_{2m}$ as $C_{sh}^{F}$, $S_{sh}^{F}$ (features) and those obtained from the $DSM_{50cm}$ as $C_{sh}^{T}$, $S_{sh}^{T}$ (targets). Their comparison yields an MBE of 2.7% between $S_{sh}^{F}$ and $S_{sh}^{T}$ and of 8.9% between $C_{sh}^{F}$ and $C_{sh}^{T}$. As the MBE for $S_{sh}$ is small compared to the order of magnitude of other uncertainties, we use an additive bias correction (MBE+) to correct $S_{sh}^{F}$ at the national scale, which is performed separately for each time step.

The $C_{sh}$ shows a larger bias, so $C_{sh}^{F}$ systematically underestimates the shaded area coefficient. This leads to an overestimation of available area. As a mean bias correction (MBE+) increases the MAE (see Table 8), we apply an ML model (RF) to estimate $C_{sh}$, illustrated by the blue box yielding $C_{sh}$ in Fig. 1-B. The features that are used to estimate

**Table 8**
Errors for estimating $C_{sh}$. Baseline is the error between $C_{sh}^{F}$ and $C_{sh}^{T}$. The considered models are MBE+ and RF.

| | RMSE | MAE | MBE | $R^2$ |
|---|---|---|---|---|
| Baseline | 0.23 | 0.13 | 0.09 | 0.12 |
| MBE+ | 0.21 | 0.14 | 0.00 | 0.25 |
| RF | 0.18 | 0.12 | 0.00 | 0.44 |

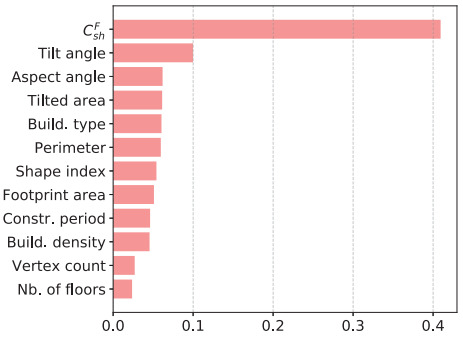



**Fig. 6.** Feature importance of all features considered for the estimation of $C_{sh}$.

$C_{sh}$ are listed in Table 6. Fig. 6 shows the feature importance for $C_{sh}$. Interestingly, the building type has a high importance. This may be due to the different objects that cause shading on different types of buildings.

The errors for estimating $C_{sh}$ are shown in Table 8. While the improvement of the MAE when using ML is relatively small, we do observe a large improvement in the $R^2$ coefficient. Unpredictable factors, such as discrepancies between the $DSM_{2m}$ and the $DSM_{50cm}$ or mismatches between the roof polygons and the DSMs, keep even the improved $R^2$ coefficient rather low. These errors will be reflected in the results by an increased uncertainty.

### 3.6. Sky view factor

The sky view factor (SVF), representing the visible proportion of the sky, is computed by combining the vertical elevation angles of the horizon for a discretized set of directions [59]. We find 32 equally spaced directions to be sufficient for a precise estimation of the SVF, which we have implemented according to Appendix C. Using the $DSM_{2m}$ and the $DSM_{50cm}$, we compute a national $SVF^{F}$ and a "true" $SVF^{T}$ in Geneva, respectively. The MBE between the two variables is 7.16%. This error of $SVF^{F}$ with respect to $SVF^{T}$ follows no identifiable pattern given the available features. We hence apply bias-correction (MBE+) to adjust the $SVF^{F}$ at national scale based on data from the high-resolution $DSM_{50cm}$, similar to $S_{sh}$ in Section 3.5.

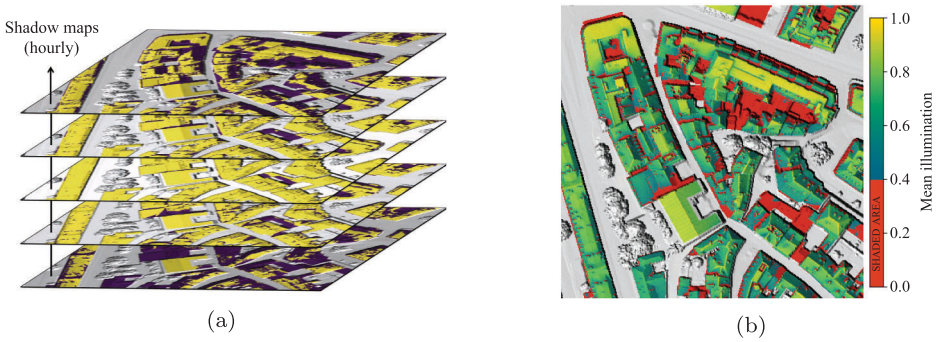



**Fig. 5.** (a) Binary shadow maps used to derive $S_{sh}$, for 5 example time steps (0: cast shadow, 1: sun exposure). (b) Mean illumination map used to compute $C_{sh}$ (red areas).

### 3.7. Flat rooftops

For a more realistic representation of PV panels on flat roofs, we assume PV panels to be installed in individual south-facing rows instead of placing them in consecutive rows along the roof aspect. Some spacing is left between the rows to reduce mutual shading effects between them (see Fig. 3). The tilt angle and row spacing is chosen as a trade-off between the total PV output and the capacity factor, which is a proxy of the economic feasibility. We have simulated different configurations and found a technically optimal trade-off for a tilt angle of 30° and a spacing of one panel height at the Swiss latitude.

The installation strategy on flat roofs impacts the PV potential in multiple ways: (i) tilted radiation increases due to the panel tilt of 30°, (ii) $C_{pv}$ decreases due to the spacing of the rows, (iii) mutual shading between adjacent rows of panels increases the hourly shaded fraction and (iv) the sky view factor is reduced. The $S_{sh}$ and the *SVF* of flat roofs are hence multiplied with an additional hourly shading factor and sky view factor, respectively, which we have simulated using a geometric model.

### 3.8. Uncertainty

The technical potential depends on multiple variables, as shown in Fig. 1. The best estimate for each variable is given by their first and second order moments, which may be used to combine various uncertainties using statistical methods. Uncertainties arise from different sources and are unknown in some cases. In this work, we consider only those uncertainties for which information can be extracted from the statistical analysis of our data. Further potential sources of uncertainty will be discussed in Section 5.4.

#### 3.8.1. Uncertainty for ML methods

To estimate the uncertainties for the output variables of the ML models, namely $G_h$, $G_B$, $\rho$, $C_{sh}$ and $C_{pv}$, we follow a two-stage approach. It distinguishes between the uncertainty arising from the modelling process (model uncertainty, $\sigma_M$) and the uncertainty related to the data noise (data uncertainty, $\sigma_D$) [60]. The data uncertainty may also represent errors introduced by previous processing steps, for example using GIS.

The model uncertainty is estimated as the standard deviation of the predictions from each ensemble member of the RF or ELM-E [55,60]. To quantify the data uncertainty, a second ML model is trained on the remaining residuals of the out-of-bag training data [54], which are derived from the squared difference between the targets and predictions [55,60]. This second ML model is used to predict $\sigma_D$ for each predicted output. Further information regarding our implementation of this method is provided in [61]. The total uncertainty of a variable estimated using ML is then the squared sum of its model uncertainty and its data uncertainty.

#### 3.8.2. Uncertainty for GIS methods

The uncertainty for the GIS-derived and bias corrected quantities ($S_{sh}$ and *SVF*) is estimated from the residuals between the values computed using the national $\mathrm{DSM}_{2m}$ ($S_{sh}^F$, $SVF^F$) and the "true" values extracted from the $\mathrm{DSM}_{50cm}$ ($S_{sh}^T$, $SVF^T$) in Geneva. The error is treated as a random variable, whose first and second order moments are computed as the mean and the variance of these residuals. The mean is used for the bias correction, as explained in Sections 3.5 and 3.6, while the variance represents the uncertainty.

#### 3.8.3. Uncertainty propagation

The propagation of uncertainty is performed by treating Eqs. (1)–(4) as functions of randomly distributed variables with statistical errors. To compute the variances of their output variables from a combination of the means, variances and covariances of their inputs (summarized in Table 4), we make the following assumptions: First, statistical independence is assumed between the solar radiation ($G_{h,B,D}$) and the DSM-derived variables ($S_{sh}$, *SVF*). This is valid as the uncertainty of $G_{h,B,D}$ is dominated by the meteorological variability, while that of $S_{sh}(t)$ and *SVF* is related to errors in the GIS methods. These are independent and uncorrelated processes. By contrast, $G_h$ and $G_B$, as well as $S_{sh}$ and *SVF*, exhibit a mutual correlation, as Table 9 shows. Therefore, their covariances must be considered in the uncertainty propagation. Second, we assume that $G_t$ and $A_{PV}$, and $C_{pv}$ and $C_{sh}$, are independent as their correlation coefficients are negligible (see Table 9). Third, we neglect the uncertainties of $\rho$ and of the temperature data, as these have a low impact on the final results. Fourth, we do not account for the uncertainty related to the physical models of $G_{Dt}$, $\eta_{PV}$ and *PF* (Section 3.1), due to a lack of data on their performance and the expected errors. This limitation will be discussed in Section 5.4.

**Table 9**
Linear correlation coefficients for pairs of potentially correlated random variables. [1]denotes a spatio-temporal mean, [2]indicates a mean across all time steps *t*.

| | $G_h$, $G_B$[1] | $S_{sh}$, $SVF$[2] | $G_t$, $A_{PV}$[2] | $C_{pv}$, $C_{sh}$ |
|---|---|---|---|---|
| Correlation | 0.87 | 0.36 | 0.04 | 0.02 |

Table 10 summarizes all variables for which uncertainties are considered, as well as the dimensions along which these are derived. Based on the above assumptions and given the uncertainties in Table 10, the variances of $G_D$ ($\sigma_{GD}^2$) and $G_t$ ($\sigma_{Gt}^2$) are derived from Eqs. (1) and (2) using error propagation theory [62,63], such that:

$$\sigma_{GD}^2 = \sigma_{Gh}^2 + \sigma_{GB}^2 - 2\ \mathrm{Cov}(G_h, G_B) \tag{5}$$

$$\sigma_{Gt}^2 = \sigma_B^2 + \sigma_D^2 + \sigma_R^2 + 2\ (\mathrm{Cov}(B, D) + \mathrm{Cov}(B, R) + \mathrm{Cov}(D, R)) \tag{6}$$

where the direct, diffuse and reflected components of $G_t$ are denoted as $B = (1 - S_{sh}) * G_{Bt}$, $D = SVF * G_{Dt}$ and $R = G_{Rt}$. The expressions for their variances ($\sigma_B^2$, $\sigma_D^2$, $\sigma_R^2$) and covariances are provided in Appendix D.

The variances of $A_{PV}$ ($\sigma_A^2$) and $E_{PV}$ ($\sigma_{PV}^2$) are derived from Eqs. (3) and (4), respectively:

$$\sigma_A^2 = A_t^2\ (\sigma_{Csh}^2\ C_{pv}^2 + \sigma_{Cpv}^2\ (1 - C_{sh})^2 + \sigma_{Csh}^2\ \sigma_{Cpv}^2) \tag{7}$$

$$\sigma_{PV}^2 = \eta_{PV}^2\ PF^2\ (\sigma_A^2\ G_t^2 + \sigma_{Gt}^2\ A_{PV}^2\ + \sigma_{Gt}^2\ \sigma_A^2) \tag{8}$$

## 4. Results

### 4.1. Physical potential estimation

The models of Section 3.3 are used to predict the MMH solar horizontal radiation and the monthly albedo for each pixel of the $(200 \times 200)\ \mathrm{m}^2$ output grid in Switzerland. The results, as well as the estimated model and data uncertainties (Section 3.8), are reported in Table 11 as monthly sums. The uncertainty of $\rho$ is not shown as it is neglected in the uncertainty propagation. For $G_h$ and $G_B$, the model uncertainty $\sigma_M$ is negligible compared to the data uncertainty $\sigma_D$ due to the large amount of training data. The model performs better for the prediction of $G_h$ than $G_B$. This may be explained by the large relative data uncertainty of $G_B$ (up to 32%), indicating that its hourly variability is higher than that of $G_h$. The results for $\rho$ are close to the standard

**Table 10**
Summary of the uncertainties for each variable in the RPV potential estimation. The dimension (dim) of the uncertainties refers to *R* as roof surfaces and *t* as time steps.

| | $G_{h,B,D}$ | $S_{sh}$ | $SVF$ | $C_{sh}$ | $C_{pv}$ | $G_t$ | $A_{PV}$ | $E_{PV}$ |
|---|---|---|---|---|---|---|---|---|
| Uncertainty | $\sigma_{Gh,B,D}$ | $\sigma_{Ssh}$ | $\sigma_{SVF}$ | $\sigma_{Csh}$ | $\sigma_{Cpv}$ | $\sigma_{Gt}$ | $\sigma_A$ | $\sigma_{PV}$ |
| Dim. of $\sigma$ | $R \times t$ | $t$ | 1 | $R$ | $R$ | $R \times t$ | $R$ | $R \times t$ |

**Table 11**
Results for the estimation of $G_h$, $G_B$ and $G_D$, showing the monthly predicted values (in kWh/m$^2$) and the model ($\sigma_M$) and data ($\sigma_D$) uncertainties (no distinction for $G_D$), as percentage of the monthly radiation.

| | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $G_h$ | 44.9 | 66.1 | 114 | 145 | 168 | 180 | 183 | 151 | 115 | 77.1 | 45.5 | 36.6 |
| $\sigma_M$(%) | 1.32 | 1.02 | 1.01 | 1.27 | 1.09 | 1.16 | 1.32 | 1.12 | 1.36 | 1.15 | 0.97 | 1.03 |
| $\sigma_D$(%) | 19.1 | 15.8 | 15.7 | 13.8 | 13.6 | 11.8 | 13.7 | 12.9 | 19.9 | 18.4 | 19.0 | 14.0 |
| $G_B$ | 22.1 | 35.2 | 62.3 | 79.6 | 87.6 | 101 | 107 | 89.7 | 68.2 | 43.1 | 23.3 | 17.8 |
| $\sigma_M$(%) | 2.28 | 1.60 | 1.55 | 1.91 | 1.91 | 2.10 | 2.23 | 1.93 | 2.04 | 1.83 | 1.82 | 2.18 |
| $\sigma_D$(%) | 30.1 | 23.1 | 26.8 | 26.2 | 29.2 | 23.2 | 26.0 | 22.5 | 32.9 | 26.3 | 31.8 | 26.2 |
| $G_D$ | 22.8 | 30.9 | 51.8 | 65.3 | 80.1 | 79.6 | 75.4 | 60.9 | 47.0 | 34.0 | 22.2 | 18.8 |
| $\sigma_{GD}$(%) | 19.0 | 17.0 | 17.3 | 16.1 | 15.8 | 14.6 | 18.3 | 16.7 | 24.7 | 20.9 | 19.3 | 13.5 |
| $\rho$ | 0.52 | 0.50 | 0.41 | 0.32 | 0.26 | 0.22 | 0.19 | 0.19 | 0.22 | 0.27 | 0.36 | 0.48 |

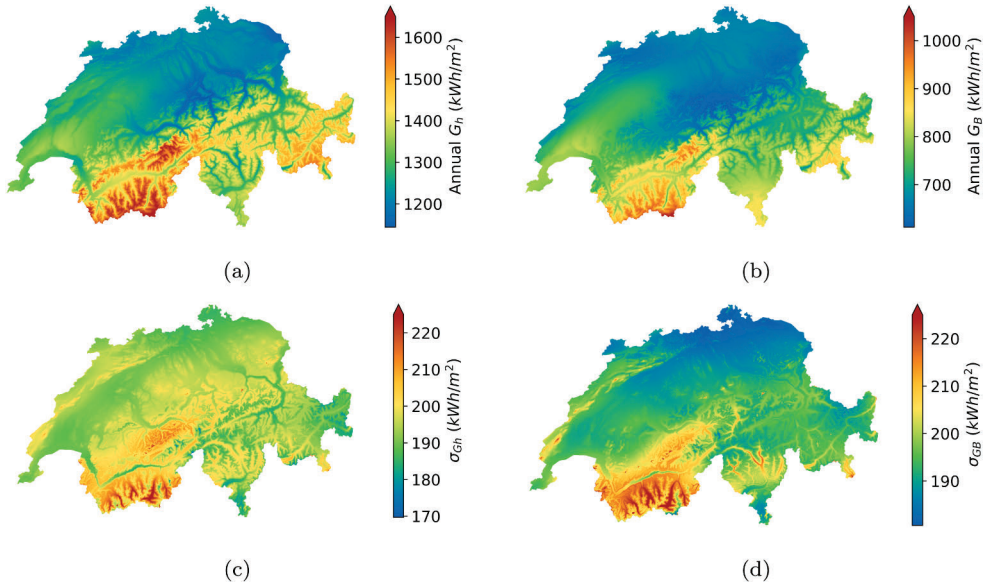



**Fig. 7.** Spatial distribution of annual predicted $G_h$ (a) and $G_B$ (b), as well as the respective total uncertainties $\sigma_{Gh}$ (c) and $\sigma_{GB}$ (d).

value of 0.2 in the summer and reach up to 0.5 on average in the winter due to snow coverage at high altitudes in the Swiss alps.

Fig. 7a and b show the spatial distribution of $G_h$ and $G_B$ for Switzerland, while Fig. 7c and d show $\sigma_{Gh}$ and $\sigma_{GB}$, which are dominated by the data uncertainty. The highest solar radiation and the largest uncertainty is found at high altitudes in the south of the country where weather extremes are frequent. The majority of buildings are however located at lower altitudes in the Swiss plateau, which spans from the north-west to the north-east of the country. The uncertainty tends to be low here, so the weather patterns are more predictable. In the plateau, also the values for $\rho$ (not shown in Fig. 7) lie below the averages quoted in Table 11 with a low uncertainty, which shows that $\rho$ mostly impacts the potential at high altitudes.

### 4.2. Geographic potential estimation

The available area for PV panel installation ($A_{PV}$) is obtained from the *shaded area coefficient* ($C_{pv}$) and the *panelled area coefficient* ($C_{sh}$) using Eq. (3). Its uncertainty $\sigma_A$ is derived from $C_{pv}$, $C_{sh}$ and their total uncertainties ($\sigma_{Cpv}$, $\sigma_{Csh}$) using Eq. (7).

Fig. 8 shows the final values for $C_{pv}$ (a) and $\sigma_{Cpv}$ (b) as a function of roof tilt and roof area, which is shown on a logarithmic scale. Large areas with a low tilt have the highest $C_{pv}$ (70–80%) and a small uncertainty. This is due to the high number of installed panels, which reduces the relative effects of the roof shape and the presence of superstructures on $C_{pv}$. Large roofs with a steep tilt have a high uncertainty, as they are rare. Due to the panel placement strategy for flat roofs (see Section 3.7), these tend to have a lower $C_{pv}$. The highest uncertainty appears for medium-sized roofs (10–100 m$^2$) with $C_{pv}$ in the range of 0.3–0.6. In this range, exact roof shapes as well as the favourable or unfavourable location of superstructures may change the number of installed panels considerably. Very small areas (<10 m$^2$) have nearly zero available area and a low uncertainty, as the roof shape is frequently unsuitable for the installation of the rectangular panels.

The predicted values for $(1 - C_{sh})$, grouped by roof aspect and tilt angles, are shown in Fig. 9a, with the related $\sigma_{Csh}$ shown in Fig. 9b. As expected, steep north-facing surfaces have the highest proportion of strongly shaded roof surface, i.e. the lowest values for $(1 - C_{sh})$, while this value is highest for steep south-facing surfaces. Interestingly, flat surfaces (in the centre of Fig. 9a) are significantly more shaded than roofs with a shallow tilt. This may be due to obstructing objects on flat surfaces or a stronger shading from surrounding buildings. The uncertainty is proportional to the shaded area coefficient and is hence highest for steep north-facing roofs.

To compute $G_t$ from Eq. (2) and its uncertainty $\sigma_{Gt}$ from Eq. (6), we combine the tilted radiation $G_{Bt,Dt,Rt}$ (Section 3.1) with the $S_{sh}$ (Section 3.5) and the *SVF* (Section 3.6). Fig. 10a shows the bias-corrected unshaded fraction $(1 - S_{sh})$ for roofs of different tilts and aspects for four example hours in June. The patterns of $S_{sh}$ follow the trajectory of the sun, with shading on west-facing surfaces in the morning hours and

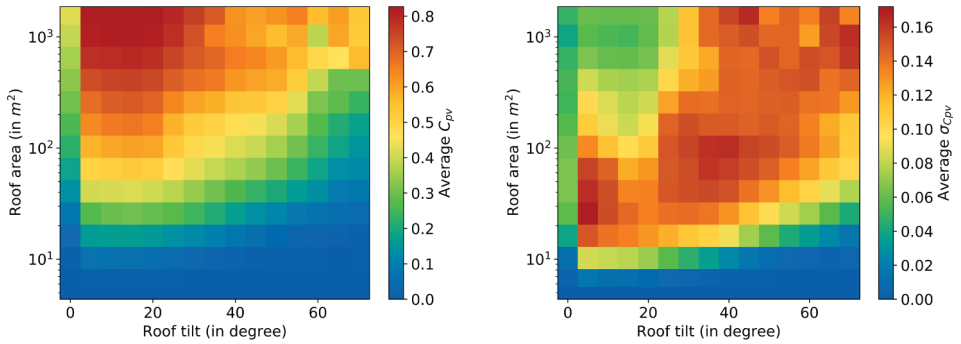



**Fig. 8.** (a) Panelled area coefficient $C_{pv}$ and (b) associated uncertainty $\sigma_{Cpv}$, for roofs with different tilt angle and roof area.

shaded east-facing roofs in the evening. Around solar noon (12 h–14 h), $S_{sh}$ is nearly zero for all roofs. This is expected as strongly shaded areas are already excluded (as part of $C_{sh}$). Fig. 10b shows the related uncertainty $\sigma_{Ssh}$ (blue line), which is strongly correlated with the zenith angle of the sun (orange line). As the uncertainty primarily arises from the discrepancy between the $DSM_{2m}$ and the $DSM_{50cm}$, these results indicate that the discretization error of the $DSM_{2m}$ has a higher impact on $S_{sh}$ at low sun altitudes. Consequently, the uncertainty is high in the morning and evening and low during midday in the summer months, when the solar radiation is highest. The average *SVF* on Swiss roofs is 0.69, with a $\sigma_{SVF}$ of 0.12 derived from the bias correction.

Fig. 11 shows the annual tilted irradiation on the building roofs ($G_t$) and its uncertainty ($\sigma_{Gt}$), again grouped by their tilt and aspect angles. The annual $G_t$ (a) is highest for south-facing roofs with a tilt angle of around 40°. This matches the latitude of Switzerland. As panels on flat roofs are oriented south and tilted at 30°, they receive near-maximum solar irradiation. Fig. 11b shows $\sigma_{Gt}$ as percentage of $G_t$. It is lowest for steep north-facing roofs, which receive low direct radiation and are hence less impacted by $\sigma_{GB}$. The highest relative uncertainty is found on steep roofs in the east and the west. These roofs receive their highest proportion of direct radiation at low sun positions, for which $\sigma_{Ssh}$ is high. Fig. 11c shows the hourly profiles for surfaces oriented north, east, west and south, averaged across all roofs of similar aspect. In summer, the peak of east-facing roofs in the morning is higher than that of the west-facing roofs in the afternoon. In winter, the opposite is the case. These results suggest that during summer the sky is clearer in the morning, while the weather conditions are generally better in the afternoon in winter.

### 4.3. Technical potential estimation

To obtain the technical PV potential ($E_{PV}$) and its uncertainty ($\sigma_{PV}$), Eqs. (4) and (8) are applied to $G_t$, $A_{PV}$, the module efficiency ($\eta_{PV}$) and the performance factor (*PF*), which are obtained as described in Section 3.1. Fig. 12 shows the annual $E_{PV}$ for Switzerland, grouped by its three main characteristics: tilt, aspect and roof area. For comparability, the values are normalized by the total roof area ($A_t$). Fig. 12a shows the potential with respect to roof tilt (x-axis) and roof area (y-axis), indicating the highest potential for large roofs (>500 $m^2$) with shallow tilt. The similarity to the pattern of $C_{pv}$ in Fig. 8a demonstrates the strong dependency of $E_{PV}$ on $C_{pv}$. As a consequence, the peak potential in Fig. 12b is no longer at 40° south, as it is the case for $G_t$ (see Fig. 11), but it appears instead at 10°–20°, which is the tilt angle of many large roofs. Flat roofs, on the other hand, show a relatively smaller potential, due the selected strategy for installing PV panels on flat roofs (see Section 3.7).

To obtain a realistic large-scale potential estimate, we exclude roofs with a small available area of $A_{PV} < 8\ m^2$ [27], which represents a minimal economic feasibility [14,26,27,64]. We further exclude all north-facing roofs with an aspect angle $|\gamma| > 90°$ from south [26,27], which aims at removing those roofs with a low PV potential. Other studies use a minimum annual $G_t$ of 1000 kWh/$m^2$ [14,20,64]. This threshold however is found to be very sensitive to small changes in the estimated potential. The aspect angle is hence preferred as selection criterion.

Applying these criteria to all 9.6 million roof surfaces in Switzerland, 2.7 million roofs on 2.3 million buildings remain suitable to install PV panels. The suitable roofs represent 56.4% of the total $A_{PV}$ of 267 ± 71 $km^2$, yielding a maximum technical PV potential of 24 ± 9 TWh. Fig. 13a shows the spatial distribution of the annual $E_{PV}$,

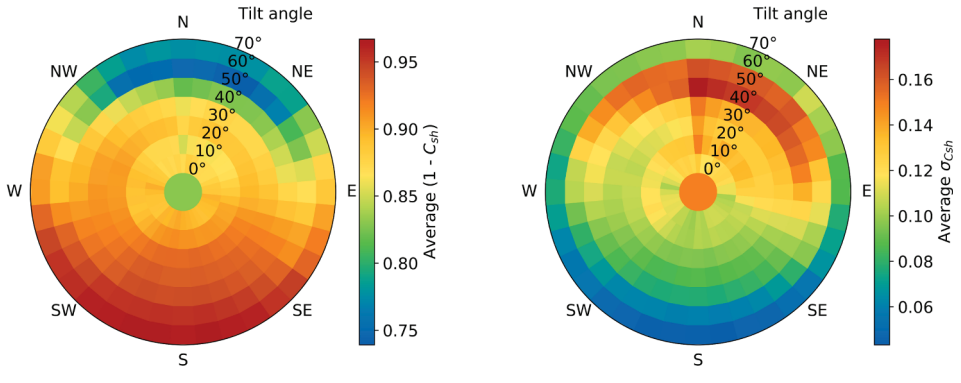



**Fig. 9.** (a) Proportion of roof area suitable for PV installation based on shading effects (1 − $C_{sh}$), (b) its uncertainty $\sigma_{Csh}$ per roof aspect and tilt.

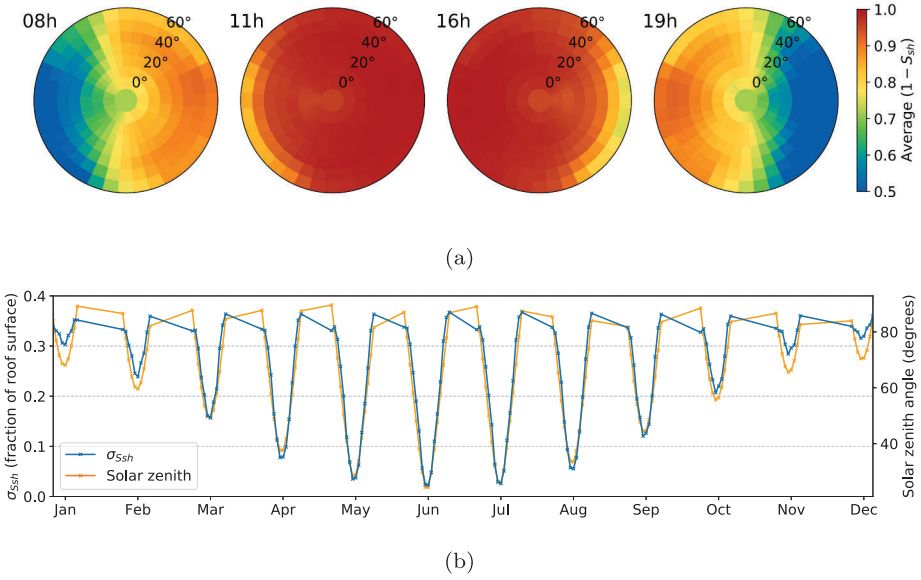



**Fig. 10.** (a) Hourly unshaded fraction of suitable roof area ($1 - S_{sh}$) for four example hours in June, averaged across roofs with equal aspect and tilt angles and (b) its uncertainty $\sigma_{Ssh}$ as a function of time. The yellow line shows the bias between the two DSMs, for which we correct the values of $1 - S_{sh}$.

aggregated to pixels of $(500 \times 500)$ m$^2$ for visualisation purposes. The potential is centered in the Swiss plateau, where most densely populated cities are located. The annual $E_{PV}$ per rooftop of one such pixel is shown in Fig. 13b, which demonstrates that large flat roofs have by far the highest potential to install PV panels. Fig. 13c shows the temporal variation of the PV potential and its total uncertainty, in comparison with the Swiss electricity demand of 2018 [65]. In absolute terms, the uncertainty is highest during the summer months, due to the high variability of the horizontal radiation in these months. The RPV potential exceeds the electricity demand even for the lower boundary of $E_{PV}$ ($E_{PV} - \sigma_{PV}$) from March until September, while a deficit is expected from November to January and during night hours.

### 4.4. Sensitivity analysis

We conduct a sensitivity analysis to quantify the impact of the individual parameters computed in this study on the technical PV potential. The analysis is performed in two stages: firstly, each parameter is independently varied in a fixed range of ± 50%. Secondly, all parameters with a quantifiable uncertainty are varied by ± $\sigma$. Due to the high correlation of the horizontal radiation components, $G_h$ and $G_B$ are varied simultaneously. A representative sample of 10,000 rooftops has been used for this analysis.

Fig. 14a shows the change in technical PV potential for varying each parameter by ± 50%. The horizontal radiation components ($G_h$), and the fractions of available roof area ($C_{sh}$, $C_{pv}$) are the most sensitive parameters and thus exhibit the steepest slope. This is expected, as these

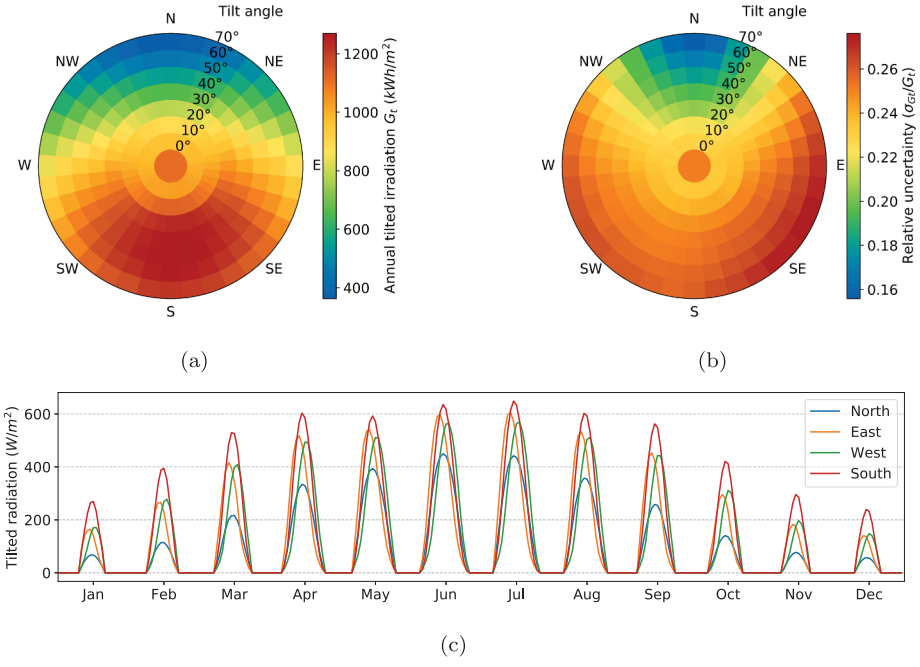
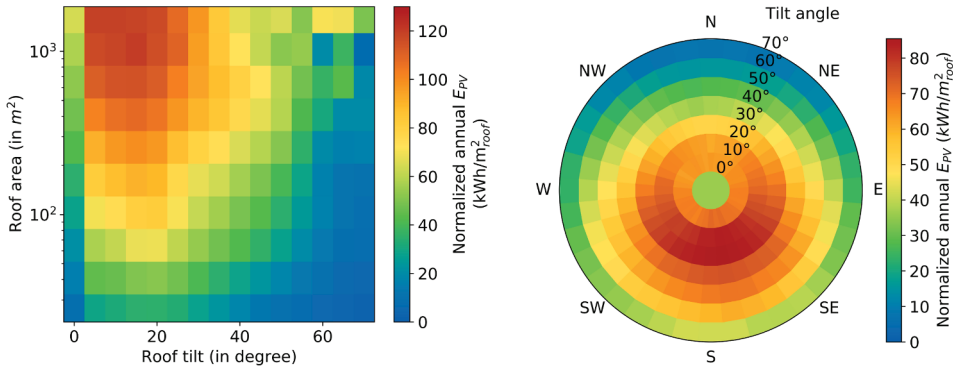



**Fig. 11.** (a) Annual tilted irradiation ($G_t$, in kWh/m$^2$) and (b) relative annual uncertainty ($\sigma_{Gt}/G_t$) per roof aspect and tilt angle, (c) MMH profiles of $G_t$ for north, east, west and south-facing roofs.

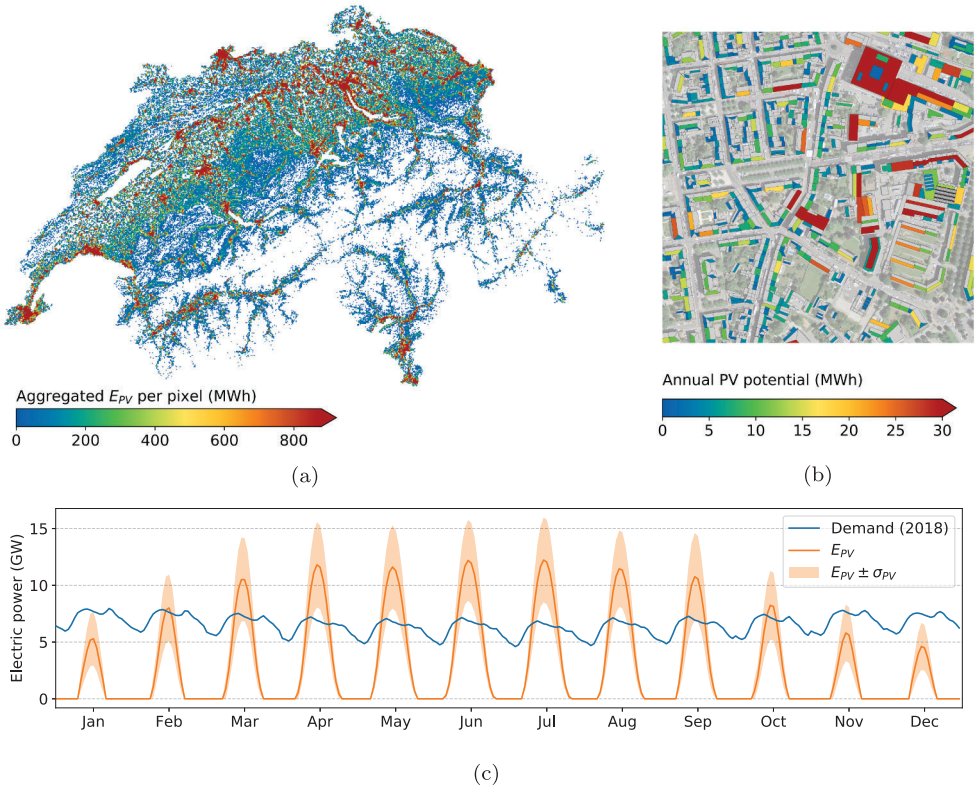



(a) (b)

**Fig. 12.** Annual technical RPV potential for all roof surfaces ($E_{PV}$, in kWh/m$^2$), (a) grouped by roof tilt and area, (b) grouped by roof tilt and aspect.

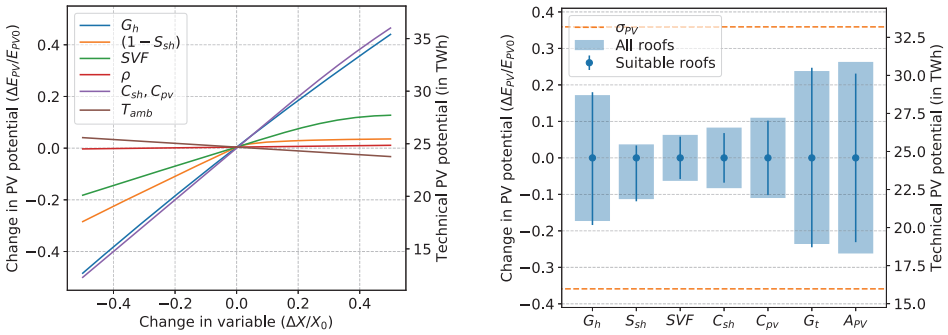
variables are directly correlated with the factors in the computation of $E_{PV}$ (see Eq. (4)). As the panel efficiency reduces slightly for high $G_h$, its sensitivity curve lies just below $C_{sh}$, $C_{pv}$. Changes in $S_{sh}$ and *SVF* have a smaller impact on the RPV potential, as they are multiplied with $G_{Bt}$ and $G_{Dt}$, respectively. The curves flatten out for large positive changes, as both factors are saturating at a value of 1. The albedo ($\rho$) has a very low sensitivity, as there are few steep roofs that are impacted by a change in $\rho$. The ambient temperature ($T_{amb}$) is the only curve with a negative slope, as high temperatures decrease $\eta_{PV}$. Its sensitivity is rather low, as $T_{amb}$ only indirectly impacts the PV potential as part of the physical model of $\eta_{PV}$.

Fig. 14b shows the impact of varying each variable within their uncertainty ($\pm\ \sigma$). The upper and lower dashed lines represent the propagated $\sigma_{PV}$. The results suggests that in a year with low $G_h$, the electricity generation may be up to 18% lower than estimated, and similarly for the higher case. The $S_{sh}$ and *SVF* show the lowest sensitivity, which agrees with Fig. 14a. Comparing $C_{pv}$ and $C_{sh}$ shows that $\sigma_{Cpv}$, driven by the high uncertainty of medium-sized roofs, is overall larger than $\sigma_{Csh}$, with a potential impact of $\pm$ 11%. The bars in Fig. 14b indicate the potential change when all roofs are considered, while the lines show the potential change for the suitable roofs only (see Section 4.3). For $G_t$, the suitable south-facing roofs have a higher uncertainty than all roofs, as expected from Fig. 11b. The opposite is the case for $A_{PV}$, as the suitable roofs have a higher proportion of flat roofs with a low $\sigma_{Cpv}$.




(a) (b) (c)

**Fig. 13.** (a) Spatial distribution of annual $E_{PV}$, aggregated to pixels of (500 × 500) m$^2$ for visualization purposes, (b) annual $E_{PV}$ for the suitable roofs of a randomly selected (500 × 500) m$^2$ pixel in the city of Geneva, (c) monthly-mean-hourly profiles of $E_{PV}$, summed for all suitable roofs, the $\sigma_{PV}$ and the Swiss electricity demand of 2018.




(a) (b)

**Fig. 14.** Change in RPV potential for the variation of (a) each variable by ± 50% and (b) each uncertain variable by ± $\sigma$. Considered variables are the horizontal radiation ($G_h$), partly shading factor ($S_{sh}$), sky view factor ($SVF$), surface albedo ($\rho$), shaded and panelled area ratios ($C_{sh}$, $C_{pv}$) and ambient temperature ($T_{amb}$).

## 5. Discussion

### 5.1. Methodological contribution

Our methodology presents an end-to-end approach to estimate the RPV potential at monthly-mean-hourly temporal resolution for individual roof surfaces. It uses large building and environmental datasets and may be transferred to any region or country where such data is or will become available. The proposed method adapts best existing practices for a data-driven estimation of each parameter that impacts RPV potential. This includes (i) the spatio-temporal variation of the horizontal solar radiation, (ii) the effects of surrounding trees and buildings on roof shading and the sky view factor, (iii) the impact of roof geometry and superstructures on the available area for installing PV panels, and (iv) the temperature-dependence of the PV module efficiency. As a result, the potential computed here is more accurate and provides a higher spatio-temporal resolution than previous works.

The strength of our method lies in the combination of physical models with GIS and ML. The former two provide a detailed representation of the physical processes underlying the RPV assessment. The use of ML, applied previously only in [26,27], allows to include additional knowledge extracted from data with only partial spatial or temporal coverage, and hence improves the accuracy of the results. The ML approach also outlines a method to use sparsely available datasets for a PV assessment in other regions.

Furthermore, we propose a structured method to estimate and propagate uncertainty in large-scale PV potential studies. Using the ML and GIS methods, we are able to quantify uncertainties related to the data sources and individual processing steps. These values are then combined using statistical tools to estimate the total uncertainty. This allows decision-makers to better understand the potential contribution of RPV to future energy systems.

### 5.2. Practical contribution

The application of the methodology to Switzerland provides several practical contributions for an effective integration of RPV in the built environment. Firstly, our results form a dataset of hourly profiles of RPV potential and its uncertainty for 9.6 M rooftops in Switzerland, which will be publicly accessible. To the best of our knowledge, it is the first dataset of its type at this spatio-temporal resolution and scale. Our data can be used by the research community to study future energy systems with large shares of RPV, whereby the uncertainty permits the modelling of different scenarios.

Secondly, the large-scale estimate of 24 ± 9 TWh provides a realistic estimate of the maximum technical RPV potential for Switzerland. According to our data, electricity generation from PV may cover over 40% of Switzerland's annual demand of 57.6 TWh in 2018 [66]. However, our results also show that the potential contribution from RPV is insufficient during winter and night hours, while there may be a large surplus of RPV generation during peak hours in summer (see Fig. 13b).

Thirdly, the sensitivity analysis performed here identifies the parameters with the highest impact on the estimated RPV potential, as well as the main sources of uncertainty (see Fig. 14). The parameter with the highest impact is the horizontal radiation. Its uncertainty is mainly caused by the high intermittency of the solar resource and can not be significantly reduced, even if higher-quality data was available. The panelled area ratio represents the second largest source of uncertainty. It is due to inaccuracies in the input data and uncertainty in the modelling approach. These may be reduced by a more precise model, for example using image processing techniques [24], or a more detailed building model (i.e. LOD 4) that is available in the entire study region.

### 5.3. Comparison with existing studies

To set our large-scale estimate into context, we compare it to five national studies on RPV potential in Switzerland [20,22,26–28], including a national project known as *Sonnendach/toitsolaire* [22,64]. This allows to validate the magnitude of our results against existing work and to point out the improvements achieved through our methodology.

**Table 12**
Comparison of the results presented in 6 studies of technical RPV potential in Switzerland. To obtain comparable results, the entries labelled with ∗ are computed from values quoted in the respective publications as explained in [67].

| Study | $A_{PV}$ (km$^2$) | Suitable roofs (%) | $G_t$ (kWh/m$^2$) | $\eta_{sys}$ (%) | $E_{PV}$ (TWh) |
|---|---|---|---|---|---|
| IEA [28] | 251∗ | 55 | 1,088∗ | 10 | 15.04 |
| Assouline et al. [26] | 328 | 60.5∗ | 662∗ | 13.6 | 17.86 |
| Assouline et al. [27] | 252 | 60.5∗ | 786∗ | 13.6 | 16.29 |
| Sonnendach [22,64] | 439∗ | 71.6∗ | 1243∗ | 13.6 | 53.09∗ |
| Buffat et al. [20] | 485 | 70.1∗ | 1176∗ | 10.3 | 41.20∗ |
| Present study | 267 | 56.4 | 1186 | 13.8 | 24.58 |

Table 12 shows a quantitative comparison of all studies, which is further detailed in [67]. The metrics used for comparison are the total available area ($A_{PV}$), the percentage of total roof surface that is suitable for installing PV, the annual tilted irradiation ($G_t$), the system efficiency ($\eta_{sys}$), which combines module efficiency and performance factor, and the annual PV potential ($E_{PV}$).

Our results lie in the mid-range of the existing work. The estimate for $A_{PV}$ is relatively low in comparison with other studies. It is in the range of the estimates by Assouline et al. [26,27], which use similar data-driven methods for estimating $A_{PV}$ as applied here. This suggests that the ratio of available roof area for PV installation lies much below current expert recommendations, which are used in [64]. This may be due to the fact that current recommendations are focused on roofs with a high potential, which may not be representative for all roofs in the Swiss building stock. In [20], the available area for PV is not specifically addressed.

The annual tilted irradiation is relatively high and comparable to that estimated in [20,22]. Both studies apply best practices for the estimation of shading effects and use high-resolution satellite data. A validation of $G_t$ against measurement data is provided in [20], who find a negligible mean error in summer when production is highest, and a small overestimation in winter. This suggests that the annual irradiation is close to its real value, while it is likely underestimated in [26–28], possibly due to a different computation of the shading effects.

Several complementarities exist between this study and previous work, given by the use of similar datasets and methods. This leads to comparable aggregated values and allows for the validation of our results against existing approaches. The added value of our work is given firstly by the computation of the results in monthly-mean-hourly resolution, instead of yielding monthly or yearly values as used in [22,26,27]. Secondly, we contribute to the existing work by assessing the available area for PV installation for each roof surface, rather than using communes [26] or pixel sizes of $(200 \times 200)$ m$^2$ [27]. Thirdly, our study is the only one that quantifies the uncertainty on the final potential estimate. Uncertainties are not specifically addressed in [26–28] and qualitatively assessed in [20,22].

### 5.4. Limitations

As all large-scale potential studies, our study relies on various data sources and a combination of statistical and empirical modelling steps. In this work, emphasis has been placed on systematically identifying and combining uncertainties in order to obtain a global uncertainty estimate for the technical RPV potential. Furthermore, a sensitivity analysis has been conducted to assess the impact of the uncertainty of individual steps on the final potential.

Some assumptions were made that impact these results: (i) The input data is assumed to be coherent and without error. In reality, some features may be missing or incorrectly represented, and discrepancies between the datasets may arise from different data collection dates and methods of creation. (ii) The systematic error introduced by using a fixed number of azimuth angles for computing the shading effects (38 bins) and the SVF (32 bins) is neglected. This error is expected to be negligible as the number of bins used here is relatively high. (iii) The potential uncertainties arising from the physical models of tilted radiation and module efficiency are neglected, as no comprehensive quantification of these has been found in the literature. (iv) The PV panel performance parameters as well as the system losses are treated as assumptions rather than random variables, due to the large variety of available technologies and their fast development. (v) PV panels on flat roofs are all placed in a technically optimal fashion (see Section 3.7). A comparison with other installation practices, such as lower tilt angles or the alternation of east and west-facing rows, is beyond the scope of this work. To consider different configurations of PV on flat roofs and the uncertainty in technological developments, a scenario-based approach may be followed.

### 5.5. Application and future work

A monthly-mean-hourly technical RPV potential for Switzerland is useful for applications including policy making, urban planning and the design of future energy systems. Policy makers may use our results, aggregated at regional or national scale, in order to formulate effective policies to integrate RPV into the built environment. Urban planners may assess the potential self-consumption and the electricity demand which could be covered by installing PV on existing roofs. They may further estimate the expected PV yield for new roofs taking into account the roof size, tilt and orientation from our results. Energy system designers can use this work to simulate future electricity networks at local scale, which allows to assess the potential mismatch between supply and demand and the resulting storage needs.

The generated dataset may be further used to estimate the PV potential for future scenarios that account for urbanization, climate change and technological advancement of PV. This is possible as the existing building stock covers different types of roofs in various urban contexts and climatic conditions, which exist in the Swiss plateau, the Jura mountains and the Alps. A techno-economic potential may be formulated for these scenarios by combining the maximum technical potential with economic factors such as installation and operational cost.

Finally, a further validation of the estimated RPV potential provides a challenging subject of future work. We validate our results against those of previous work (Section 5.3), some of which have been compared to measurement data [20]. While this approach is feasible for validating the tilted radiation, a direct validation of the available area is difficult due to the lack of a ”ground truth”. A possible method to establish such a ground truth may be the application of image segmentation techniques, for example using Convolutional Neural Networks [68], to high-resolution aerial imagery. This would allow for a more accurate detection of obstructing objects on rooftops as well as the recognition of already installed PV panels.

## 6. Conclusion

In this work, we present a big data mining approach to estimate the PV potential on 9.6 million rooftops at monthly-mean-hourly temporal resolution and propose a quantification the uncertainty on the estimated potential. The developed Machine Learning methodology uses high-resolution building and environmental data to extract information from data which is only available in parts of the study area. The Machine Learning algorithms are further used to quantify the uncertainties related to the estimated parameters. These are combined with uncertainties arising from other modelling steps and propagated to obtain an uncertainty on the technical PV potential estimate.

The national-scale application of our method results in a total rooftop PV potential for Switzerland of $24 \pm 9$ TWh, which could meet more than 40% of the country’s electricity demand in 2018. This potential is in the mid-range of other large-scale estimates for Switzerland. The added value of our approach lies in the higher spatio-temporal resolution of the resulting datasets compared to existing studies, as well as in the uncertainty estimate for each spatio-temporal instance. This highlights areas and time spans with potentially large inaccuracies. Our results may be used to quantify the potential mismatch between PV production and supply, which is relevant for the design of future energy efficient districts and for the formulation of regional and national incentive policies for the diffusion of rooftop PV.

The work presented in this study provides an important contribution for the decarbonisation policies in Switzerland, as it enables the large-scale modelling of future electricity grids with high shares of rooftop PV using hourly data for individual buildings. The proposed PV assessment method using data mining and uncertainty propagation is transferable to any region or country with sufficient high-quality data, where it can contribute to the transition towards low-carbon energy systems.