A Fast Machine Learning Model for Large-Scale Estimation of Annual Solar Irradiation on Rooftops

Alina Walch[1], Roberto Castello[1], Nahid Mohajeri[2] and Jean-Louis Scartezzini[1]

[1] Solar Energy and Building Physics Laboratory, Ecole Polytechnique Fédérale de Lausanne (Switzerland)

[2] Urban Development Programme, Department for Continuing Education, University of Oxford (United Kingdom)

Abstract

Rooftop-mounted solar photovoltaics have shown to be a promising technology to provide clean electricity in urban areas. Several large-scale studies have thus been conducted in different countries and cities worldwide to estimate their PV potential for the existing building stock using different methods. These methods, however, are time-consuming and computationally expensive. This paper provides a Machine Learning approach to estimate the annual solar irradiation on building roofs (in kWh/m^2) for large areas in a fast and computationally efficient manner by learning from existing datasets. The estimation is based on rooftop characteristics, input features extracted from digital surface models and annual horizontal irradiation. Five ML models are compared, with Random Forests exhibiting the highest model accuracy. In the presented case study, the model is trained using data of the Swiss Romandie area and is then applied to estimate annual rooftop solar irradiation in remaining Switzerland with an accuracy of 92%.

Keywords: Rooftop photovoltaics, annual solar irradiation, city-scale PV potential, Machine Learning

1. Introduction

Rooftop-mounted photovoltaics (PV) have shown to play a key role in the transition to renewable electricity generation in cities and urban areas. Hence, several studies which estimate annual PV potential on building roofs have been conducted city or country scale, for example in Germany (Mainzer et al., 2017; Ramirez Camargo et al., 2015), in Spain (Izquierdo et al., 2008; Ordóñez et al., 2010) and in Switzerland (Assouline et al., 2017; Buffat et al., 2018; Klauser, 2016). Various input parameters need to be determined to estimate PV potential at large scale. These include (i) global, direct and diffuse horizontal irradiation at hourly or daily resolution, (ii) roof slope and aspect, (iii) shading effects from trees and buildings, and (iv) available roof area for PV panel installation. The parameters are determined using geographic information systems, image processing or Machine Learning (ML) techniques. A physical model is then applied to compute the tilted irradiation, which is multiplied by the available roof area for PV installation and the system's performance to obtain the technical rooftop PV potential.

Existing studies of PV potential hence require the collection of various input datasets and the implementation of computationally intensive data processing methods to compute each parameter. We propose a model which can accurately predict long-term annual solar irradiation on building roofs from a reduced set of inputs and with a significantly smaller computational effort compared to existing studies. Recent suggested methods such as a simplified skyline-based method (Calcabrini et al. 2019) are insufficient for this task, as separate models needed to be fitted for each pair of roof slope and aspect. Instead, we train a Machine Learning model using data from an existing dataset of rooftop PV potential at national scale in Switzerland, in order to learn the relation between rooftop features, weather data and rooftop irradiation. The use of ML for a large-scale estimation of rooftop PV potential has been tested at the scale of communes (Assouline et al., 2017) and pixels of 200m × 200m resolution (Assouline et al., 2018), but it has not yet been applied to the estimation of solar irradiation for individual roof surfaces. ML methods to predict solar irradiation perform a short-term forecasting for individual roofs, but do not address a large-scale potential estimation. An overview of these methods is provided by Voyant et al. (2017).

doi:10.18086/swc.2019.45.12 Available at http://proceedings.ises.org

In this work, we compare the performance of five different ML models in estimating the annual solar irradiation on building roofs and we quantify the uncertainty for the predicted values. We further use Machine Learning to reduce the number of input features by determining those with the highest importance for the prediction of annual irradiation. Our model is trained and tested on data belonging to two distinct geographic areas in Switzerland. The testing procedure demonstrates that our model predicts tilted irradiation (in kWh/m^2) with an accuracy of 92.3%. The results suggest that the proposed model is suitable to estimate annual solar irradiation on building roofs in regions with similar geographic and meteorological conditions, for example in Germany, France or northern Italy.

2. Data and Methods

To estimate annual solar irradiation on building roofs (G_t, in kWh/m^2), we account for different types of input features, which represent the main input parameters used in existing studies. These are (i) horizontal irradiation, (ii) roof aspect, (iii) roof tilt, (iv) shading effects from buildings and trees, (v) sky visibility, (vi) horizon maps. We use a feature selection procedure to extract the most relevant features, and compare five Machine Learning methods with respect to their performance in predicting G_t. The complete methodology is summarized in Fig. 1.

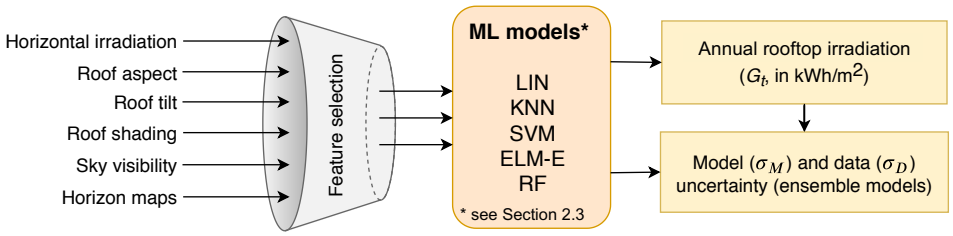

Fig. 1: Schematic for estimation of annual rooftop solar irradiation (G_t). Feature selection is applied to choose the most important features and to reduce the total number of features. Five Machine Learning algorithms are tested for the prediction of annual G_t. The model and data uncertainties are computed for the ensemble-based algorithms (ELM-E and RF).

2.1. Dataset description

The target variable estimated in this work is the annual rooftop solar irradiation (G_t). We use values for G_t from a publicly available dataset of PV potentials for 9.6M rooftops in Switzerland (Klauser, 2016) as labels for training the ML models. The method applied by Klauser (2016) to obtain G_t is summarized in the Appendix. The dataset also contains information on roof slope and roof aspect, which are input features to the ML models. We further include the global and direct horizontal irradiation components (GHI and DNI) in the set of features. They are derived from satellite data provided by the Swiss meteorological office MeteoSwiss (Stöckli, 2013) for the years of 2004-2014, by linearly interpolating the satellite pixels to the coordinates of each roof surface and averaging the results for all years. The diffuse horizontal irradiation is omitted, as it is the difference between GHI and DNI.

The shading effects from surrounding buildings and trees and the sky visibility, which is quantified as the sky view factor (SVF), are typically derived from the horizon height of a roof. The horizon height can be interpreted as the minimum sun altitude to illuminate a given point (in degrees), as shown in Fig. 2a. It is typically computed for each azimuth angle in bins of 5° or 10° using a Digital Surface Model (DSM). The shading effects are then computed for each time step (e.g. hourly) as the portion of roof surface with a horizon height above the current sun altitude, while the SVF is obtained by integrating the horizon height for all azimuth angles. We use a DSM of Switzerland in (2x2)m^2 resolution, provided by the Swiss topographical office (Swisstopo, 2005), to compute the average horizon height for each roof. In the feature set, we include the roof shading, obtained by averaging hourly shading effects, the SVF and 9 of the DSM-derived horizons for azimuths of 60°-300° in bins of 30° (see Fig. 2a). Table 1 summarizes all 15 features considered in the feature selection procedure.

Tab. 1: All available features for the estimation of rooftop solar irradiation. Italic features may not be readily available or are computationally demanding to compute.

Roof features	Horizontal irradiation	Shade and skyview
Aspect angle	Annual global irradiation (GHI)	*Roof shading*
Tilt angle	*Annual direct irradiation (DNI)*	*Skyview factor (SVF)*
		9 × horizon height (30° angle bins)

Tab. 2: Characteristics of six cities selected for testing of the presented method. A_{roof} denotes the total roof surface.

City	Inhabitants	Area (km²)	# buildings	A_{roof} (km²)	A_{roof} / inhab.	Altitude (m)
Zürich	409,241	91.9	46,888	13.77	33.65 m²	400
Basel	171,513	23.9	11,122	6.27	36.55	260
Luzern	81,401	37.4	5,940	3.25	39.95	435
Lugano	63,494	86.2	10,052	3.10	48.85	270
Davos	10,937	284.0	2,582	1.21	1110.49	1,560
Interlaken	5,592	4.3	1,025	0.48	85.06	565

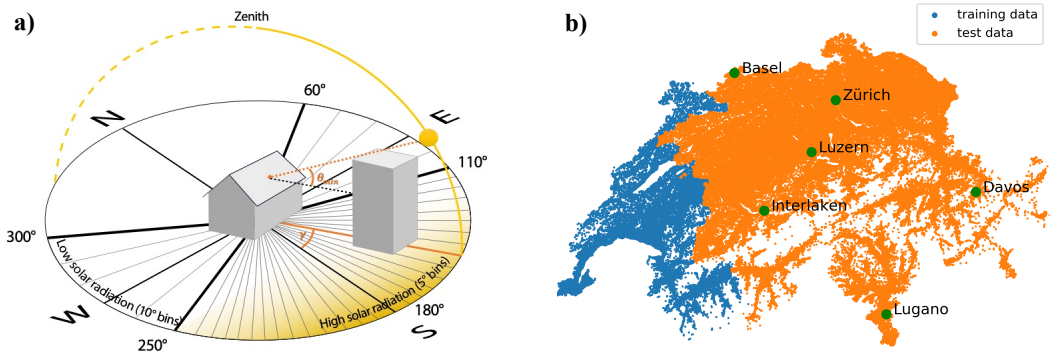

Fig. 2: a) Schematic representation of the calculation of horizon heights, computed for each DSM pixel and averaged per rooftop; b) Geographic location of training and testing data (Swiss Romandie area and the remaining Switzerland).

The data set is standardized and split into training and test data, covering separate geographic areas of Switzerland (see Fig. 2b). We use the Swiss Romandie area for training and the remaining country for testing. This boundary is chosen as it represents a cultural and language border which divides all three geographic terrains of Switzerland, namely the Alps, the Plateau (where most buildings are situated) and the Jura mountains. We randomly select 100,000 samples for training and 1,000,000 samples for testing with no noticeable reduction in the model performance. We choose six cities inside the test area, shown in Fig. 2b, for which we will predict the aggregated annual PV potential in order to test the performance of the proposed method at city level. The cities are selected to cover different population sizes, geographic areas and altitudes. The details for each city are shown in Table 2.

2.2. Feature selection

The reduction of the number of features to a small set of easily available variables is one of the key aspects of this work. To obtain such a reduced feature set, we use a recursive selection procedure that can be applied to any ML algorithm. The metric used for the selection is the mean-squared error (MSE), which is obtained using a k-fold cross-validation (CV). For this, the training data is randomly split into k subsets (folds), from which k ML models are trained. Each model uses $(k - 1)$ folds for training and the last fold for validation, i.e. to compute the MSE between the target (annual G_i) and the predicted value. In this work, we use a 5-fold cross-validation ($k = 5$).

Starting from the complete feature set, we iteratively exclude one feature at a time and compute the MSE. The feature whose exclusion causes the lowest change in MSE (Δ_{MSE}) is permanently removed from the set of features. This procedure is repeated until only one feature remains. The Δ_{MSE} is recorded for each iteration and a threshold for Δ_{MSE} is chosen, which represents the maximum tolerable increase in MSE. The selected features are those which form the smallest feature set that keeps Δ_{MSE} below its threshold. As the number of cross-validations increases with N^2, where N is the number of features, this procedure may not be applicable for high-dimensional problems, but it is appropriate for the size of the feature set used in this study (up to 15 features). We further aim to obtain a feature set that contains only easily accessible features that can be obtained with a low computational effort. Three out of 15 features, highlighted in Table 1, are hence excluded before applying the recursive selection. These are (i) the DNI, which is frequently unavailable at measurement stations and not included in some future climate scenarios, (ii) the roof shading and (iii) the SVF, which both require the computation of the horizon height for all azimuth angles, which is the highest computational burden in large-scale PV potential studies.

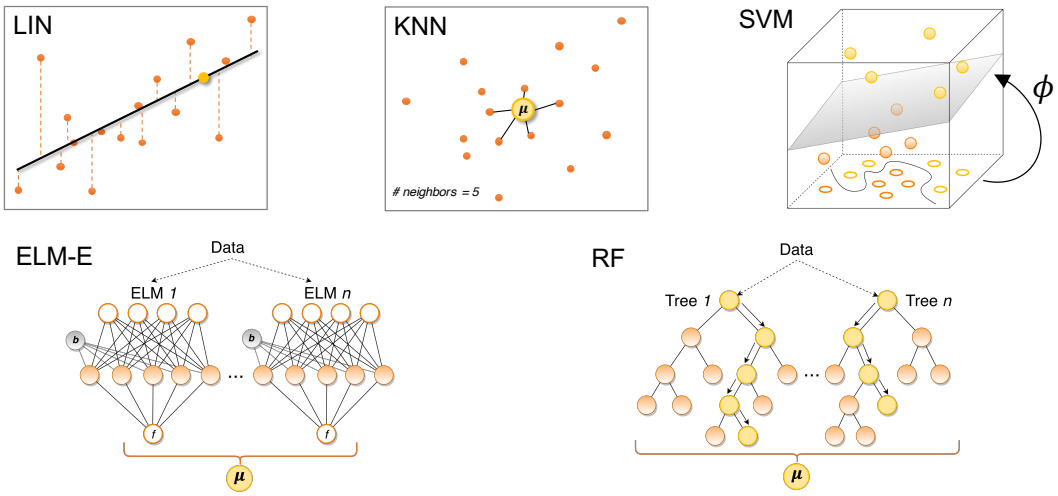

Fig. 3: Concepts of ML algorithms used in this work. The selected algorithms are Linear regression (LIN), K-Nearest Neighbor Regression (KNN), Support Vector Machine (SVM), Extreme Learning Machine Ensemble (ELM-E) and Random Forest (RF).

2.3. Machine Learning Algorithms

We compare five supervised regression models from different families of ML algorithms: (i) Linear Regression (LIN), (ii) K-Nearest Neighbors (KNN), (iii) Support Vector Machines (SVM), (iv) Extreme Learning Machine Ensembles (ELM-E), and (v) Random Forests (RF). In each case, the annual G_t is predicted from the selected features. The architecture of each model is determined through a tuning procedure in the training phase, where the parameters defining the model structure (hyper-parameters) are optimized in order to minimize the mean-squared error (MSE) between the prediction and the target. The ML algorithms and their hyper-parameters are summarized below and shown in Fig. 3. Table 3 gives the values of all hyper-parameters, which are tuned using 5-fold cross-validation. All algorithms except ELM-E are implemented using the *Scikit-learn* library for python (Pedregosa et al., 2011), while the ELM-E is based on the *HPELM* package for python (Akusok et al., 2015).

Linear Regression assumes that the target is a linear function of the inputs. The prediction is obtained from the linear combination of the features which minimizes the residual sum of squares between the target and predicted values. It is fast and requires no tuning of hyper-parameters but shows a low accuracy for non-linear problems.

K-nearest Neighbor Regression is an interpolation algorithm, which computes a prediction as the average of the targets of the k training samples whose features are closest to the given inputs. The training dataset hence works as a look-up table for the predictions, which makes it effective for low-dimensional problems but inefficient for large datasets. We use the Euclidean distance as a measure of "closeness" and tune the number of neighbors (k).

Extreme Learning Machine Ensemble is a collection of single-layer neural networks (ELMs), which were developed by (Huang et al., 2006). Each ELM consists of one hidden layer, which is optimized in a more efficient way than traditional neural networks. This results in a faster training time and a low number of hyper-parameters. The aggregation of n ELMs in an ensemble further increases the robustness of the model and reduces the risk of overfitting. We tune the size of the hidden layer (m_{ELM}) and the number of ELMs in the ensemble (n_{ELM}).

Support Vector Machine, introduced by (Cortes and Vapnik, 1995), is the most popular algorithm in the family of kernel methods. It exploits the *kernel trick*, which projects the features to a higher-dimensional space that allows for linear modelling. Its structure makes it particularly effective for high-dimensional problems, but it does not scale well with the number of samples. In this work, we use ε-Support Vector Regression with a radial basis function as kernel and tune the kernel coefficient (γ), the penalty parameter (C) and the error tolerance (ε).

Random Forest is another ensemble algorithm, which was proposed by (Breiman, 2001). It consists of regression trees, which pass a training sample along a set of nodes based on a threshold (defined during training) until a *leaf node* is reached. The prediction of each tree is obtained by averaging the target values in the respective leaves. It is a popular algorithm due to its good predictive power and high robustness. Its main hyper-parameters are the number of features considered for the optimization of each threshold (m_{ftrs}), the minimum number of samples in each leaf (m_{leaf}) and the number of trees in the ensemble (n_{est}).

Tab. 3: Hyperparameters for each Machine Learning algorithm after tuning using *k*-fold cross-validation (*k* = 5).

LIN	KNN	ELM-E	SVM	RF
-	$k = 17$	$m_{ELM} = 45$	$C = 5$	$m_{ftrs} = 3$
		$n_{est} = 25$	$\gamma = 200$	$m_{leaf} = 3$
			$\varepsilon = 0.1$	$n_{est} = 500$

2.4. Uncertainty Estimation for Ensemble Models

We implement a method to estimate uncertainties on the predicted values, which has been used successfully to estimate GHI at high spatio-temporal resolution (Walch et al., 2019). This method is applicable to models with an ensemble structure, i.e. ELM-E and RF. It allows to distinguish between the uncertainty arising from the modelling process (σ_M) and the uncertainty related to the data noise (σ_D). The model uncertainty is estimated as the standard deviation of the ensemble predictions, and the data uncertainty is derived from the modelling residuals, such that:

$$\hat{\sigma}_M^2(\boldsymbol{x_i}) = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_i^n - \hat{y}_i)^2\,, \quad i = 1, \dots, L \qquad \text{(eq. 1)}$$

$$\hat{\sigma}_D^2(\boldsymbol{x_i}) = \min\{(t_i - \hat{y}_i)^2 - \hat{\sigma}_M^2(\boldsymbol{x_i}), 0\}, \quad i = 1, \dots, L \qquad \text{(eq. 2)}$$

where $\boldsymbol{x_i}$ denotes input sample i, $\hat{y}_i^n$ is the prediction of each ensemble member n, $\hat{y}_i$ is the ensemble prediction (the mean of $\hat{y}_i^n$), t_i is the target for sample i, N is the ensemble size and L is the total number of samples. As t_i is not available for the prediction, a second ML model is trained with the residuals as targets, using the same hyper-parameters as the primary ML model. This second model is used to estimate σ_D during the test phase.

3. Results

3.1. Feature selection and importance analysis

As described in Section 2.2, we perform recursive feature selection for the reduced set of 12 features, which excludes the DNI, roof shading and SVF. We use the KNN as assessment model due to its high computational efficiency and low error. Figure 4a shows the results for the selection procedure as the increase in the cross-validation root-mean squared error (RMSE) for the selected features with respect to the RMSE obtained from the complete feature set (listed in Tab. 1). The curve falls steeply for a low number of features, where each additional feature brings a large improvement in performance, and flattens out when six or more features are used. We choose a threshold of 5kWh/m^2 as maximum acceptable increase in RMSE, corresponding to approximately 5% of the modelling RMSE (see Tab. 4). This gives a set of six remaining features, and six features are excluded. The removed features are shown in Fig. 4a in the order of exclusion, indicating that north-facing horizon heights have the smallest contribution in the modelling, followed by the near-south as well as the east and west horizon heights.

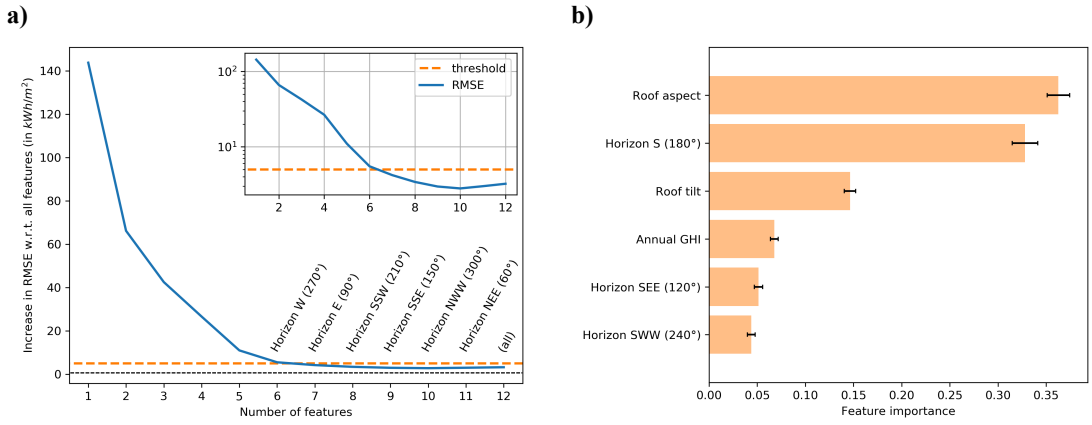

Fig. 4: a) Change in RMSE for recursively eliminating variables from the feature set, with indication of eliminated features (threshold = 5kWh/m^2); b) Feature importance score (obtained using RF) for all selected features used in the ML models.

The remaining features are shown in Fig. 4b, ranked by their importance score. This score is obtained from the built-in feature importance method of the Random Forest algorithm, which quantifies the contribution of each feature to increasing the model performance. The results confirm that all features are relevant to the modelling, with the roof aspect as well as the south-facing horizon height scoring highest, before the roof tilt and GHI. This analysis suggests that three horizon heights, at 120°, 180° and 240°, are sufficient for a good prediction of annual tilted rooftop irradiation.

3.2. Comparison of ML models

After choosing the final set of features (see Fig. 4b), we apply each of the five ML models to the test data. The models are previously tuned individually using 5-fold cross-validation, yielding the hyper-parameters shown in Tab. 3. The test errors and execution times for the algorithms are given in Tab. 4. Four error metrics are shown, namely RMSE, mean absolute error (MAE), mean bias error (MBE) and the R^2 coefficient of determination (R2), as well as the training and testing times (using a laptop with 4 cores and 8GB RAM). Linear regression is the fastest method but shows large errors. KNN and ELM-E show a similar performance and execution time, with particularly low training times. SVM has a similar error as KNN and ELM-E, but it is very slow in comparison. This is expected, as the SVM is designed for high-dimensional feature spaces rather than large datasets.

The RF shows the lowest errors with a reasonable execution time. Further tests have shown that this is due to the strong decrease in the error when increasing the size of the training set from 10k to 100k samples. This improvement, caused by a larger training sample, is much smaller for the other algorithms, which suggests that the RF is well suited for modelling large datasets. Using the RF model, the time for estimating the solar irradiation of 1M rooftops is 33.3s, which is considerably less than the hours needed to perform a detailed modelling. Figure 5a shows a density graph for RF of the targets against the predicted values and the fitted regression line. We see that the prediction tents to slightly underestimate G_t, as many values lie just above the diagonal. Figure 5b gives the predicted G_t for a random sample of 50 roofs, with their targets and uncertainties. The predicted values generally lie within one standard deviation of the total uncertainty (error bars). The MAE for the test sample (generalization error) amounts to 7.7% of the target potential, resulting in an average accuracy of 92.3%.

Tab. 4: Test errors and computational time for the five ML models, using 100k samples for training and 1M for testing.

	LIN	KNN	ELM-E	SVM	RF
RMSE (kWh/m^2)	157.03	115.48	117.55	113.92	94.69
MAE (% of target)	14.16	9.45	10.03	9.21	7.66
MBE (% of target)	-8.84	-6.38	-5.16	-5.88	-3.87
R2	0.55	0.76	0.75	0.76	0.84
Training time (s)	0.02	0.31	2.03	254.1	44.42
Prediction time (s)	0.16	29.94	39.44	1178.3	33.30

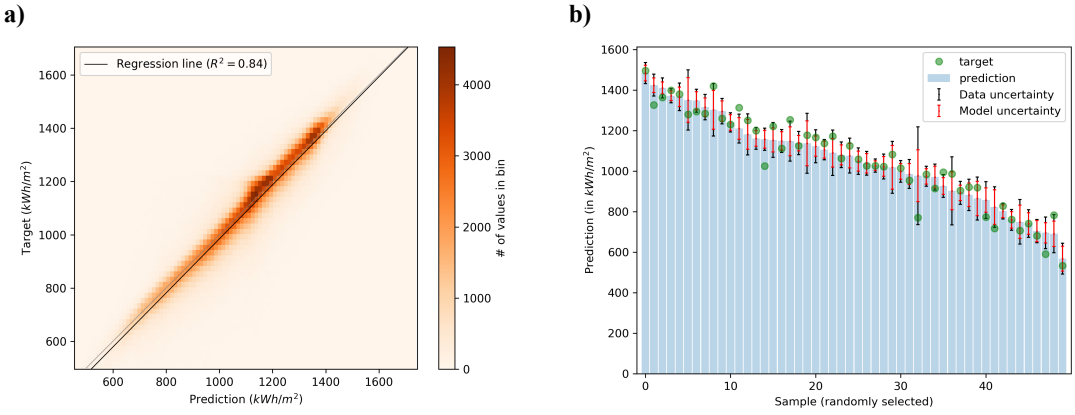

Fig. 5: a) Comparison of predicted and target irradiation for 1M roofs in the test area; b) Prediction (blue), target (green) and uncertainties (bars) for a random sample of the test data

3.3. Large-scale estimation of rooftop solar irradiation patterns

We use the test results for the RF model to analyze the solar irradiation patterns for roof surfaces in Switzerland with respect to their aspect and tilt angles. Figure 6a shows the predicted annual G_t, with flat roofs in the center. The largest irradiation is found for surfaces with a tilt angle of 30-50°, which is expected given Switzerland's latitude. Steep north-facing roofs have the lowest solar irradiation, of less than 700kWh/m^2 on average. Figure 6b shows the relative uncertainty, as percentage of predicted G_t. The relative uncertainty is lowest for the roofs with the highest solar irradiation, while it is high for roofs with low G_t. This may be due to a low number of steep roofs and strong shading effects on these. We also observe a relatively high uncertainty for (nearly) flat roofs, which may be due to unpredictable shading effects on these roofs, for example caused by objects installed on the surfaces.

To assess the performance of the model on the large scale, we compute the PV potential for the six cities shown in Tab. 2. To obtain the PV potential, we follow the method suggested by (Portmann et al., 2016), which was developed for the dataset used here. The annual G_t is multiplied with the total roof surface, a roof utilization factor (which depends on the roof tilt and the building type), a module efficiency of 17% and a performance factor of 80%. Roofs with a surface area below 10m^2 and an irradiation below 1000kWh/m^2 are excluded. The aggregated PV potential for the selected cities is shown in Fig. 7a, while Fig. 7b shows the PV potential per inhabitant. The orange bars show the target value, the blue bars show the RF prediction and the error bars indicate the uncertainty. The largest cities clearly have the highest total PV potential, while the small cities (Davos and Interlaken) have the highest potential per inhabitant. It is interesting to see that Lugano has a noticeably higher potential than Luzern, despite a nearly identical roof surface, due to its location south of the Alps. Davos, with its high altitude and large roof area per person, has the highest potential per inhabitant. Comparing the targets and predictions, we see that the difference is small, and in all cases within the estimated uncertainty.

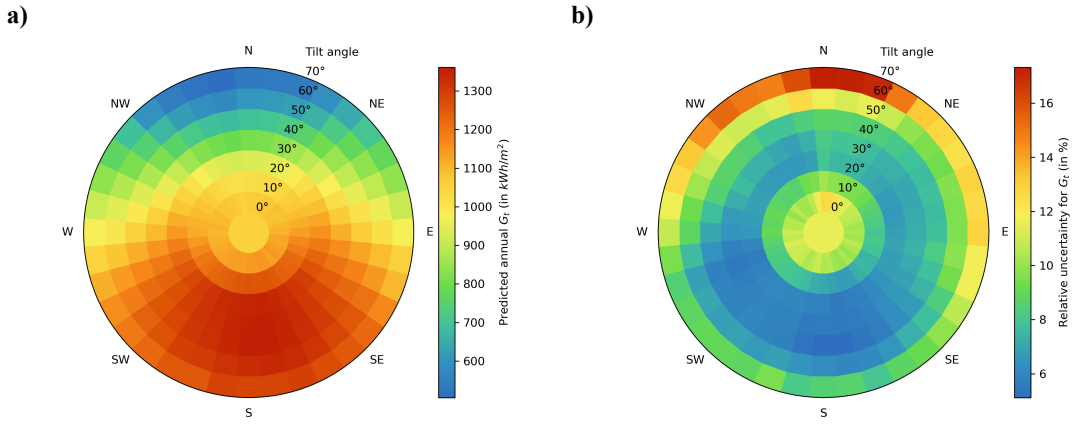

Fig. 6: a) Predicted pattern for annual solar irradiation (G_t) and b) estimated uncertainty (as percentage of predicted G_t) for building roofs with different aspect and tilt angles (grouped in bins of 10°)

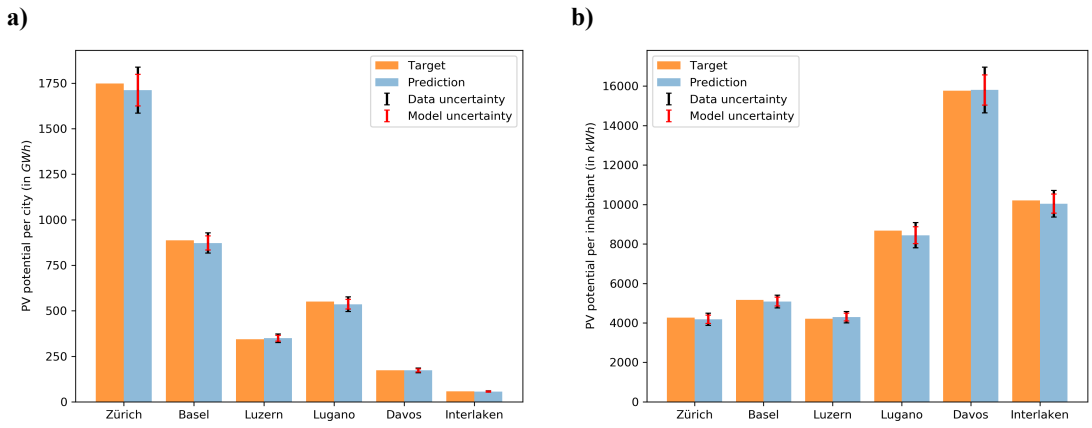

Fig. 7: a) Target and predicted PV potential for six Swiss cities of different sizes (see Tab. 2); b) PV potential per inhabitant for each city

4. Discussion

4.1. Limitations

The model proposed here reduces the set of features required for a large-scale estimation of annual solar irradiation to the GHI, the roof aspect, the roof tilt as well as three horizon heights, approximately towards south, east and west. While this largely reduces the computation required for such a study, a DSM and a reliable dataset with roof tilt and aspect angles are still necessary. The latter two are only available for cities with a 3D building model (LOD 2 or higher). If such a dataset is not available, a high-resolution DSM or a roof shape classification as proposed by (Mohajeri et al., 2018) may be used to derive tilt and aspect angles. As the model is trained in Switzerland, it is only applicable to locations with a similar latitude, as the latitude has a major impact on the peak irradiation and the corresponding tilt angle. To extend the area of applicability, results from a study in another geographic area can be added to the training dataset. Finally, we do not specifically address the available area for the installation of PV panels in this work. Studies show that the available area may be much smaller than the total roof surface (Assouline et al., 2017), which reduces the estimated large-scale PV potential.

4.2. Application and future work

The motivation behind this work is to facilitate the study of rooftop PV potentials in areas where no such study has yet been carried out, primarily in similar meteorological and geographic conditions as Switzerland. Applying our model to such a case study will be the principal objective of future work. The model can further be used to predict future PV potentials, accounting for climate as well as urbanization scenarios. The former leads to changes in GHI, while the latter reflects as changes in the number of roofs and the composition of tilt and aspect angles.

While the estimation of annual irradiation gives a useful indication of the order of magnitude of the large-scale PV potential, a higher temporal resolution is required to assess its seasonal and intra-day variation. Studies in monthly (Assouline et al., 2018) or hourly (Buffat et al., 2018) resolution may be used for this task. Furthermore, transfer learning techniques may be applied to retrain our model on different latitudes, meteorological conditions and roof characteristics, by using large-scale PV potential studies from other regions or countries.

5. Conclusion

This study provides a Machine Learning algorithm to estimate annual solar irradiation on building rooftops for areas at similar latitudes as Switzerland. The proposed model is trained using data of existing PV potential studies and uses the following six input features: annual GHI, roof tilt, roof aspect and three mean horizon heights for each roof (120°, 180°, 240°). Comparing five ML algorithms, we found that Random Forests are most suited for this task due to their high accuracy, reasonable execution time and their ensemble structure, which in turn allows for a sound estimation of uncertainties. The obtained results show that the proposed model can effectively learn from existing large-scale datasets of PV potential and accurately estimate annual irradiation for individual rooftops and entire cities. The computational time for estimating solar irradiation at city or even country scale can thereby be reduced from hours to seconds.

6. Acknowledgments

This research is supported by the Swiss National Science Foundation (SNSF) under the National Research Program 75 (Big Data) Project 167285 and by the SNSF Mobility Fellowship P300P2 174514.

7. References

Akusok, A., Bjork, K.-M., Miche, Y., Lendasse, A., 2015. High-Performance Extreme Learning Machines: A Complete Toolbox for Big Data Applications. IEEE Access 3, 1011–1025. https://doi.org/10.1109/ACCESS.2015.2450498

Assouline, D., Mohajeri, N., Scartezzini, J.-L., 2018. Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests. Applied Energy 217, 189–211. https://doi.org/10.1016/j.apenergy.2018.02.118

Assouline, D., Mohajeri, N., Scartezzini, J.-L., 2017. Quantifying rooftop photovoltaic solar energy potential: A machine learning approach. Solar Energy 141, 278–296. https://doi.org/10.1016/j.solener.2016.11.045

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

Buffat, R., Grassi, S., Raubal, M., 2018. A scalable method for estimating rooftop solar irradiation potential over large regions. Applied Energy 216, 389–401. https://doi.org/10.1016/j.apenergy.2018.02.008

Calcabrini, A., Ziar, H., Isabella, O., Zeman, M., 2019. A simplified skyline-based method for estimating the annual solar energy potential in urban environments. Nature Energy 1. https://doi.org/10.1038/s41560-018-0318-6

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach Learn 20, 273–297. https://doi.org/10.1007/BF00994018

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: Theory and applications. Neurocomputing, Neural Networks 70, 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Izquierdo, S., Rodrigues, M., Fueyo, N., 2008. A method for estimating the geographical distribution of the available roof surface area for large-scale photovoltaic energy-potential evaluations. Solar Energy 82, 929–939. https://doi.org/10.1016/j.solener.2008.03.007

Klauser, D., 2016. Solarpotentialanalyse für Sonnendach.ch (Schlussbericht).

Loutzenhiser, P.G., Manz, H., Felsmann, C., Strachan, P.A., Frank, T., Maxwell, G.M., 2007. Empirical validation of models to compute solar irradiance on inclined surfaces for building energy simulation. Solar Energy 81, 254–267. https://doi.org/10.1016/j.solener.2006.03.009

Mainzer, K., Killinger, S., McKenna, R., Fichtner, W., 2017. Assessment of rooftop photovoltaic potentials at the urban level using publicly available geodata and image recognition techniques. Solar Energy 155, 561–573. https://doi.org/10.1016/j.solener.2017.06.065

Mohajeri, N., Assouline, D., Guiboud, B., Bill, A., Gudmundsson, A., Scartezzini, J.-L., 2018. A city-scale roof shape classification using machine learning for solar energy applications. Renewable Energy 121, 81–93. https://doi.org/10.1016/j.renene.2017.12.096

Ordóñez, J., Jadraque, E., Alegre, J., Martínez, G., 2010. Analysis of the photovoltaic solar energy capacity of residential rooftops in Andalusia (Spain). Renewable and Sustainable Energy Reviews 14, 2122–2130. https://doi.org/10.1016/j.rser.2010.01.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825−2830.

Perez, R., Ineichen, P., Seals, R., Michalsky, J., Stewart, R., 1990. Modeling daylight availability and irradiance components from direct and global irradiance. Solar Energy 44, 271–289. https://doi.org/10.1016/0038-092X(90)90055-H

Portmann, M., Galvagno-Erny, D., Lorenz, P., Schacher, D., 2016. Sonnendach.ch: Berechnung von Potenzialen in Gemeinden.

Ramirez Camargo, L., Zink, R., Dorner, W., Stoeglehner, G., 2015. Spatio-temporal modeling of roof-top photovoltaic panels for improved technical potential assessment and electricity peak load offsetting at the municipal scale. Computers, Environment and Urban Systems 52, 58–69. https://doi.org/10.1016/j.compenvurbsys.2015.03.002

Stöckli, R., 2013. Daily, monthly and yearly satellite-based global radiation (Documentation of MeteoSwiss Grid-Data Products). Federal Office of Meteorology and Climatology MeteoSwiss.

swisstopo, 2019. Produktinformation swissBUILDINGS3D 2.0. Federal Office of Topography.

Swisstopo, 2005. DOM - Die Geodaten der Schweiz des Bundesamtes für Landestopografie für den professionellen Einsatz. Swiss Federal Office of Topography.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., Fouilloy, A., 2017. Machine learning methods for solar radiation forecasting: A review. Renewable Energy 105, 569–582. https://doi.org/10.1016/j.renene.2016.12.095

Walch, A., Castello, R., Mohajeri, N., Guignard, F., Kanevski, M., Scartezzini, J.-L., 2019. Spatio-temporal modelling and uncertainty estimation of hourly global solar irradiance using Extreme Learning Machines. Energy Procedia, Innovative Solutions for Energy Transitions 158, 6378–6383. https://doi.org/10.1016/j.egypro.2019.01.219

Appendix: Annual solar irradiation on building roofs

This section summarizes the data and the method used in Klauser (2016) to compute the annual solar irradiation on building roofs in Switzerland, which is used as target in the Machine Learning model proposed in this work.

1. Data

Meteorological data
Four types of meteorological datasets, are used in the study. They are the global, direct and diffuse horizontal radiation (in W/m^2) and the surface albedo (in [0,1]). The data has an hourly temporal resolution for the years of 2004-2014 and a spatial resolution of 1.25 degree minutes (≈1.6km × 2.3km) covering all of Switzerland. It is provided by MeteoSwiss (Stöckli, 2013).

Roof surface polygons
A dataset of roof geometry polygons has been derived from the latest CityGML LOD2 building cadaster of Switzerland (swisstopo, 2019). It contains 9.6M roof surfaces for 3.7M buildings in the cadaster. The attributes include the roof tilt and roof aspect angle and its tilted area. This dataset is used in the present work as features for the ML model.

Digital Surface Models (DSM)
Four surface models are overlaid to create a combined DSM at a 0.5m pixel resolution. It consists of the rasterized geometries of the building cadaster, the DSM and digital terrain models of Switzerland (interpolated to 0.5m pixels), and a Radar Topography surface model (SRTM) for 100m pixels outside of Switzerland. Details and sources are provided in (Klauser, 2016).

2. Method

Horizon matrix
To compute the effects of roof shading and the skyview factor, a horizon matrix is computed for each roof. It represents the mean sky visibility of a roof for each zenith and azimuth angle, in a resolution of 1° (zenith) × 5° (azimuth). It is computed by overlaying three types of horizons maps: (i) a far horizon (25km distance) for each roof, based on the combined DSM aggregated to 100m, (ii) a medium distance horizon (1km) for the center of each roof, based on the DSM aggregated to 10m and (iii) a near horizon (100m) for each pixel of the DSM, which is averaged across each roof surface.

Shading effects and skyview factor (SVF)
The roof shading is computed for each hour of a year and represents the reduction in direct radiation received by a surface due to obstructing objects or landscape features. Its value equals the value of the horizon matrix corresponding to the solar position (zenith and azimuth angle) at a given hour. The SVF represents the mean visibility of the sky and is obtained by numerically integrating the horizon matrix.

Tilted radiation
A physical model is used to compute the tilted radiation (G_t) on the rooftops for each hour in the meteorological dataset. It is based on the global (G_h), direct (G_B), and diffuse (G_D) horizontal radiation components, which are linearly interpolated from the satellite grid to the roof coordinates. The annual solar irradiation for each building roof is obtained by summing the hourly tilted radiation for each year and averaging the results.

The tilted radiation consists of a direct (G_{Bt}), diffuse (G_{Dt}) and reflected (G_{Rt}) component such that:

$$G_t = G_{Bt} + G_{Dt} + G_{Dt} \qquad \text{(eq. 3)}$$

The basic form of the hourly geometric model for the direct tilted radiation is given by:

$$G_{Bt} = G_B * max\left(0, \frac{\cos\theta}{\cos\theta_Z}\right) \qquad \text{(eq. 4)}$$

where $\cos\theta = \sin\beta\sin\theta_Z\cos\gamma_S - \gamma + \cos\beta\cos\theta_Z$. The angles θ_Z and γ_S describe the sun zenith and azimuth angles, while the roof tilt and aspect are given by β and γ. In (Klauser, 2016), the direct tilted radiation is multiplied with the hourly roof shading.

The diffuse tilted radiation is computed using the Perez model (Perez et al., 1990), which is formulated as:

$$G_{Dt} = G_D * \left[(1 - F_1)\left(\frac{1+\cos\beta}{2}\right) + F_1\frac{a}{b} + F_2\sin\beta\right] \qquad \text{(eq. 5)}$$

where F_1 and F_2 are empirically fitted functions and a and b are geometric angles. The three addends in eq. 5 denote the isotropic diffuse radiation, the circumsolar radiation and the horizon brightness, respectively (see Loutzenhiser et al., (2007) for details). In (Klauser, 2016), the isotropic radiation is multiplied with the SVF, while the circumsolar radiation is multiplied with the hourly roof shading.

The reflected radiation is computed using the surface albedo ρ such that:

$$G_{Rt} = G_h * \rho\left(\frac{1-\cos\beta}{2}\right) \qquad \text{(eq. 6)}$$

In (Klauser, 2016), the reflected radiation is multiplied with (1 – SVF).

A more detailed description of the physical model is provided for example in (Loutzenhiser et al., 2007).