DEVELOPMENT OF MACHINE LEARNING-BASED APPROACH FOR SOLAR POTENTIAL ESTIMATION

XUAN LIAO

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

The Department of Land Surveying and Geo-Informatics

Development of machine learning-based approach for solar potential estimation

XUAN LIAO

A thesis submitted in partial fulfillment of

the requirements for the degree of Doctor of Philosophy

July 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

____Xuan LIAO_____(Name of student)

Abstract

The global pursuit of carbon neutrality aims to mitigate greenhouse gas emissions and establish a sustainable future. Solar energy is one of the most promising approaches, as it produces minimal greenhouse gas emissions. Installing solar photovoltaic panels on rooftops maximizes solar irradiation reception while reducing energy transmission losses and costs. Given the high installation costs of solar photovoltaic (PV) panels, accurately estimating solar potential to determine optimal installation locations is crucial to ensure economic benefits exceed installation costs, making the investment viable.

Accurate solar potential estimation faces several challenges: i) Various natural (such as clouds and weather) and artificial factors (building, morphological features) affect solar irradiation, making quantification and establishing non-linear relationships challenging. ii) Current algorithms struggle with the spatio-temporal characteristics of solar irradiation, which is crucial for effective estimation. iii) Large-scale solar potential estimation involves processing vast data, posing computational challenges.

This study employs a hierarchical assessment framework based on machine learning to estimate solar potential, including physical and geographical potential. The major achievements of this thesis are:

(1) Four machine learning models (Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Regression, Multilayer Perceptron) were used to estimate land surface solar irradiation in Australia, China, and Japan using meteorological data, Himawari-8 satellite cloud and aerosol products, and solar observation data. GBM showed the highest accuracy, suggesting its effectiveness for large regions and applicability globally with similar datasets. This method generated accurate and continuous solar maps to display solar resource distributions at large-scale regions.

(2) To address geographic heterogeneity in estimating land surface solar irradiation, the

Dual-gate Temporal Fusion Transformer (DGTFT) was proposed. Applied to datasets from Australia, China, and Japan, the proposed network outperformed traditional machine learning methods, with a minimum Coefficient of determination (R^2) increase of 23.88%, Mean Absolute Error (MAE) decrease of 43.18%, and Normalized Root Mean Square Error (nRMSE) decrease of 62.79%. These results suggest that the proposed network not only improves estimation performance but also provides interpretable results for understanding the network mechanism.

(3) This study proposes a parametric-based data and model dual-driven method to estimate annual rooftop solar irradiation at a fine spatial resolution. Three machine learning methods (RF, GBM, and AdaBoost) were cross-compared based on R^2, MAE, and computation time. In a Hong Kong case study, RF outperformed GBM and AdaBoost, with R^2 =0.77 and MAE=22.83 kWh/m^2/year. Training and prediction time for rooftop solar irradiation was within 13 hours, achieving a 99.32% reduction compared to the physical-based hemispherical viewshed algorithm, indicating the proposed method's accuracy and speed for large datasets.

(4) The DGTFT model was employed to estimate hourly rooftop solar irradiation, capturing spatio-temporal distribution variations. The proposed method achieved highly accurate results, with R^2=0.90, MAE=26.90 MJ/m^2, RMSE=32.39 MJ/m^2, and was 56 times faster than the model-driven method. These results demonstrate the high spatio-temporal resolution rooftop solar maps' reliability for solar potential assessment.

This thesis offers promising approaches for estimating solar potential from physical to geographical potential at high spatio-temporal resolution, utilizing Geographic Information System (GIS) representation of multi-source data and exploring non-linear relationships using Geospatial Artificial Intelligence (GeoAI) methods. The findings provide a reliable reference for planning and installing solar PV systems.

Publications Arising from the Thesis

Liao, X., Wong, M. S., Zhu, R. "A Dual-gate Temporal Fusion Transformer for forecasting large-scale land surface solar irradiation". *Renewable and Sustainable Energy Reviews* (2024) (Under review)

Liao, X., Wong, M. S., Zhu, R. "A temporal fusion transformer-augmented GeoAI framework for estimating hourly land surface solar irradiation: A case study in Australia". *Applied Energy* (2024). (Minor review)

Liao, X., Zhu, R., Wong, M. S., Heo, J., Chan, P. W., & Kwok, C. Y. T. "Fast and accurate estimation of solar irradiation on building rooftops in Hong Kong: A machine learning-based parameterization approach". *Renewable Energy* (2023): 216.

Liao X., Zhu R., Wong MS. "Simplified estimation modeling of land surface solar irradiation: A comparative study in Australia and China". *Sustainable Energy Technologies and Assessments* (2022): 5.

Acknowledgements

I am profoundly thankful for the opportunity to pursue my Ph.D. degree at the Hong Kong Polytechnic University. I wish to convey my sincerest appreciation to those who supported and guided me during my doctoral journey.

Foremost, my heartfelt gratitude goes to my supervisor, Prof. Man Sing WONG, who afforded me the invaluable opportunity to pursue my Ph.D. His extensive expertise, unwavering support, tireless guidance, and insightful comments played a pivotal role in helping me navigate the challenges encountered in my research. Over the past three years, I have felt privileged to learn and collaborate with him, making my Ph.D. journey both enriching and unforgettable. I truly appreciate the dedicated supervision and support of Prof. Man Sing WONG, without which my Ph.D. research would not have been possible.

I would also like to express my gratitude to Dr. Rui Zhu for his invaluable assistance in my research study, constant encouragement, constructive feedback, and support with the publications stemming from this thesis. Additionally, I extend my thanks to Dr. Guoqiang Shi for supporting my participation in the remote sensing competition.

My heartfelt thanks go out to my wonderful colleagues and fellow group members in the Department of Land Surveying and Geo-Informatics for their friendly assistance. Special appreciation to Coco Yin Tung Kwok, Xinyu, Yu, Keru Lu, Fan Xu, Songyang Li, Shaolin Wu, Jing Li, Meilian Wang, Qian Peng, Xindi Liu, Jingjing Li, and other colleagues who have supported and encouraged me from the past to the present.

Lastly, but certainly not least, I want to express my deep appreciation to my family members for their unconditional support and understanding throughout my academic journey."

Table of Contents

List of Figures

List of Tables

List of abbreviations

GBM	Gradient Boosting Machine
RF	Random Forest
SVR	Support Vector Regression
MLP	Multilayer Perceptron
COT	Cloud optical thickness
AOT	Aerosol optical thickness
R^2	Coefficient of determination
nRMSE	Normalized Root Mean Square Error
nMBE	Normalized mean bias error
T	Consumption of time
GBRT	Gradient Boost Regression Tree
MAE	Mean Absolute Error
ELM	Extreme learning machine
PV	Solar photovoltaic
DNN	Deep neural network
DEM	Digital elevation models
CSI	Clear-sky solar irradiation
DGTFT	Dual-gate Temporal Fusion Transformer
SVR_poly	SVR used the polynomial model
SVR_rbf	SVR used the radial basis model
MARS	Multivariate Adaptive Regression Spline
CART	Classification and Regression Tree
M5	Piecewise Linear Functions of Regression Trees
GBMs	Gradient Boosting Machines
DSM	Digital Surface Model
AHI	Advanced Himawari Imager
MaxT	Maximum temperature
MinT	Minimum temperature
H	Average humidity
WS	Average wind speed
P	Average atmosphere pressure
BP	Back Propagation
TFT	Temporal Fusion Transformer
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
ARMAX	Autoregressive moving average model with exogenous variables
ARFIM	Autoregressive fractionally integrated moving average
LSTM	Long Short-term Memory
TCN	Temporal Convolutional Network
RNN	Recurrent Neural Network
GRN	Gated Residual Network
DGRN	Dual-gate Gated Residual Network

DGMCA	Dual-gate Multi-head Cross Attention
CA	Cross attention
SVF	Sky View Factor
MT	Morphological tessellation
MTC	Morphological tessellation cells
CARF	Classification and regression tree
VIF	Variance Inflation Factor
RF_{46}	The RF model using 46 parameters
RF_{7}	RF model using seven parameters
$AdBoost_{7}$	AdaBoost model utilizing seven parameters
GBM_{7}	GBM utilizing seven parameters
MRE	Mean Relative Error
D	Diffuse proportion
T	Transmittivity

Chapter 1 Introduction

1.1 Research Background

As society and the economy continue to develop, the demand for energy has been steadily increasing. However, the excessive exploitation of traditional energy sources, such as oil and coal, has resulted in the depletion of these energy reserves on our planet. Additionally, the emissions of pollutants during the utilization of conventional energy sources have inflicted severe damage to our environment, giving rise to pressing issues like climate change and air pollution. In response to these challenges, a total of 195 countries entered into the "Paris Agreement", in 2015, which aims to limit the global average temperature increase in this century to below 2 degrees Celsius and strive for carbon neutrality by 2050 (Holden et al., 2018). To achieve this target, developing and utilizing low-pollution, renewable energy sources have become a critical and promising solution in addressing energy shortages and environmental pollution, as well as ensuring the sustainability of urban development (Vaka et al., 2020).

Solar energy stands out as one of the most widely distributed energy sources, distinguished by its near ubiquity. Sunlight reaches nearly every corner of the globe, thereby making solar energy a global energy potential. Unlike other renewable sources such as wind and hydropower, solar energy is not bound by geographic location. Consequently, even remote regions have the potential to efficiently harness solar energy resources. Furthermore, solar energy generation does not produce direct greenhouse gas emissions, aligning with carbon neutrality initiatives and assisting in mitigating the impacts of climate change. Therefore, solar energy is widely recognized as the future cornerstone of renewable energy.

Despite the significant potential of solar energy development in reducing carbon emissions, the current efficiency of solar energy utilization remains a challenge. Globally, renewables (excluding hydroelectricity) represented 7.5% of primary energy consumption in 2022, with solar power alone contributing 28.9% to global renewable energy generation, as reported in the 2023 Statistical Review of World Energy (Statistical Review of World Energy, 2023). However,

even in developed cities where natural conditions and economic incentives favor solar energy, actual adoption rates remain low. Take Hong Kong, for example, which enjoys favorable climatic conditions with abundant sunlight throughout much of the year and ample rooftop space suitable for solar photovoltaic panels. Despite these advantages, Hong Kong's utilization of solar power lags behind. According to the Hong Kong Energy End-Use Data 2023 (Hong Kong Energy End-use Data, 2023), solar power accounted for only 15.1% of the city's renewable energy production in 2021, and renewable energy as a whole comprised just 1% of Hong Kong's total energy consumption. These disparities underscore the need for more effective policies and incentives to enhance solar energy adoption, both globally and locally, to achieve substantial reductions in carbon emissions and foster sustainable energy transitions.

To fully harness solar energy resources, it is imperative to conduct precise assessments of solar potential. Solar potential assessment entails the estimation of solar radiation in specific regions and the analysis of the spatial and temporal distribution of solar resources. This assessment provides valuable information for selecting suitable sites for solar photovoltaic (PV) installation and serves as the foundation for evaluating the potential of solar energy utilization.

In current research, solar potential is categorized into three levels (Izquierdo et al., 2008): physical potential, geographic potential, and technical potential. These three levels represent a progressive relationship in the journey from natural resources to usable energy. The physical potential represents the total amount of radiation reaching the Earth's surface, primarily influenced by solar radiation intensity and local weather conditions. The geographical potential quantifies the total solar radiation that can be received by building rooftops and façades. In urban areas, solar panels installed on building rooftops and façades are affected by surrounding obstructions, reflections, and the tilt of the installation location of solar PV panels, which leads to a decrease in receiving solar radiation. Technical potential involves the conversion of solar radiation into usable electricity or heat using rooftop solar panels or solar thermal collectors. This level is mainly influenced by the type and performance of the equipment and the coverage area of the installation. Technical potential belongs to the field of semiconductor materials and this field is out of the boundary of the research region of this study, so this study just focuses

on the physical potential and geographical potential.

1.1.1 Solar physical potential

Researchers have conducted extensive studies on solar irradiation estimation, achieving significant advancements in developing models over the past two decades (Zhang et al., 2017). Traditional methods for solar irradiation estimation can generally be categorized into three groups: empirical (Bailek et al., 2018; Benatiallah et al., 2019; Makade et al., 2019), physical (Ceballos et al., 2004; Cogliani et al., 2007; Yeom et al., 2016), and machine learning models (Voyant et al., 2017; Guermoui et al., 2020; Zhou et al., 2021).

Several researchers have utilized empirical models to estimate solar irradiation using data from meteorological stations. These models include cloudiness-based, sunshine-based, temperature-based, and models based on multiple meteorological parameters (Fariba et al., 2013). Specifically, meteorological parameters-based models employ multiple meteorological variables to estimate solar radiation, such as the Swartman and Ogunlade model (Swartman & Ogunlade, 1967), Sabbagh model (Sabbagh et al., 1977), Lewis model (Lewis et al., 1983), and Garg and Garg model (Garg & Garg, 1982). In sunshine-based models, Joes et al. (2016) compared the Ångström–Prescott model with ten modified versions based on sunshine duration to estimate daily global and monthly averaged solar irradiation in Alagoas State, Northeastern Brazil. While these models generally yield more accurate results, their applicability is limited to regions with available solar irradiation records.

Considering the limitations of the previously mentioned models, temperature-based models are proposed as a convenient and accessible alternative. Marius et al.（2013) introduced a temperature-based model for global solar irradiance and applied it to estimate daily irradiation values. These models operate on the premise that the difference between maximum and minimum temperatures influences the fraction of extraterrestrial radiation reaching the ground. However, other factors such as cloudiness, humidity, latitude, elevation, topography, and proximity to large bodies of water can also affect temperature differences (Allen et al., 1997).

Cloud coverage, in particular, significantly impacts solar irradiation estimation (Wong et al., 2016) and can be readily obtained from satellite or ground-based measurements. For instance, Nikitidou et al. (2019) developed a novel method for estimating surface irradiance under clear skies based on short-term cloudiness forecasting. Despite the convenience of empirical models for estimating solar irradiation, their applicability is often confined to small regions, making it challenging to adapt the same empirical model to different areas.

Additionally, numerous studies have focused on physical models for estimating solar irradiation. These models include radiation transmission and parameterized models such as the METSTA model (Maxwell et al., 1998), Bird model (Richard et al., 1987), Yang model (Yang et al., 2001), and Page model (Page et al., 1997). Some researchers have utilized data from both meteorological stations and satellites. For instance, Chen et al. (2014) used MODIS atmospheric products, including cloud fraction, cloud optical thickness (COT), precipitable water vapor, and aerosol optical thickness (AOT), to estimate the monthly mean global solar radiation over China. Zhang et al. (2015) proposed an integrated approach combining a digital elevation model with MODIS atmospheric water vapor and aerosol products to estimate shortwave solar radiation on clear-sky days. Similarly, Feng and Wang (2021) merged ground-based sunshine duration observations with cloud fraction and aerosol optical depth to produce high-resolution, long-term surface solar radiation data over China. While satellite images used in these models provide extensive and continuous spatial distribution information, they generally estimate solar radiation at low temporal resolution, limiting their ability to achieve near real-time monitoring.

Furthermore, numerous studies have focused on traditional time series methods for estimating solar radiation. These methods include autoregressive integrated moving average (ARIMA) (Shadab et al., 2020), autoregressive moving average (ARMA) (Ji and Chee, 2011), autoregressive moving average with exogenous variables (ARMAX) (Silva et al., 2022), and autoregressive fractionally integrated moving average (ARFIMA) (Ismail and Karim, 2020). Time series models predict solar radiation based on historical data. While these models have generally been successful in estimating solar radiation, they fail to account for the significant

impacts of meteorological and geographical changes on solar radiation.

Given the versatility of machine learning in achieving accurate predictions, numerous machine learning methods have been developed for estimating solar irradiation in recent years (Zhou et al., 2021). Machine learning models can be broadly categorized into three types: ANN-based (Behrang et al., 2010; Wang et al., 2015; Khosravi et al., 2018), Kernel-based (Prada et al., 2018; Rohani et al., 2018; Sun et al., 2018), and Tree-based (Sharafati et al., 2019; Wu et al., 2019; Yagli et al., 2019). Ghimire et al. (2019) utilized a convolutional network (CNN) to extract features related to future solar radiation changes and incorporated these into a Long Short-Term Memory network for half-hourly global radiation forecasting. Their results indicated that this deep learning hybrid model outperformed all other models and is suitable for monitoring solar-powered systems. Similarly, Ravinesh et al. (2019) proposed an extreme learning machine (ELM) model to predict long-term solar radiation over Australia using data from the Moderate Resolution Imaging Spectroradiometer and geo-temporal input variables such as periodicity, latitude, longitude, and elevation. The ELM model demonstrated superior prediction accuracy compared to other artificial intelligence algorithms like Random Forest (RF), Piecewise Linear Functions of Regression Trees (M5 Tree), and Multivariate Adaptive Regression Spline. In addition, Badia and Xavier (2014) developed an ANN-based model using daily weather forecasts to predict daily global solar radiation, showing satisfactory accuracy. Compared to physical and empirical models, machine learning models offer moderate accuracy and broader application for solar irradiation prediction, making them a popular choice. Ramedani et al. (2014) compared the performance of support vector regression (SVR) and fuzzy linear regression for global solar radiation prediction in Iran, with the SVR using both polynomial (SVR_poly) and radial basis function (SVR_rbf) kernels. Their findings showed that the SVR_rbf model performed better than fuzzy linear regression. Mawloud et al. (2018) employed the Gaussian Process Regression (GPR) algorithm with various combinations of test data to predict daily global solar radiation on a horizontal surface, finding that the model based on sunshine duration, relative humidity, and minimum air temperature outperformed other combinations. Lee et al. (2020) compared the prediction performance of ensemble methods (Boosted Trees, Bagged Trees, Random Forest, and Generalized Random Forest) with common

methods (Gaussian process regression and SVR) for estimating solar irradiation in the United States. Using hourly data on cloud cover, temperature, wet-point temperature, relative humidity, wind speed, and visibility, the results showed that ensemble learning models had superior performance. Park et al. (2020) introduced a multistep-ahead solar radiation forecasting model based on the light gradient boosting machine, which outperformed other tree-based and deep learning models.

However, these traditional methods just investigate the non-linear relationship between many input parameters and solar radiation and ignore the time series characteristic of solar radiation, which may lead to suboptimal performance in solar radiation estimation. To solve this problem, some time series deep learning methods are proposed and show good performance in extracting time series features, such as Long Short-term Memory (LSTM), Temporal Convolutional Network (TCN), and Recurrent Neural Network (RNN). Alper et al. (Alper et al., 2023) used LSTM, Multilayer Perceptron (MLP), and adaptive neuro-fuzzy inference system with grid partition, and fuzzy c-means to predict the one-hour-ahead solar radiation in Tarsus. The results illustrate that the LSTM model in 1-h-ahead solar radiation forecasting yielded the highest accuracy performance. Similarly, Kong et al. (Kong et al., 2023) predicted solar radiation for space heating with a thermal storage system based on the Temporal convolutional network-attention model. The results show that the prediction performance of this model is superior to other algorithms, including the RNN, LSTM, and gated recurrent unit, with Root Mean Square Error (RMSE) = 45.07 W/m^2. However, these time series deep learning methods cannot consider the impact of the geographic variation in different regions on the distribution of solar radiation, and this geographic variation also plays a significant role in empirical methods (Ertekin and Yaldız, 1999). Additionally, a common limitation of both machine learning methods and time series deep learning is poor interpretability because of their black-box nature. Therefore, recent studies have focused on the interpretability of deep learning models because it helps to understand and trust the decisions made by these models. For example, Temporal Fusion Transformer (TFT) (Lim et al., 2021) is a novel attention-based deep learning method that shows good prediction performance using multiple time-series data with static data and can provide an interpretable model. The TFT method offers a robust solution for time-series

estimation by efficiently capturing complex temporal dependencies, combining the features of static data and time-varying data, and incorporating an interpretable explanation of temporal dynamics and high-performance forecasting over multiple horizons.

1.1.2 Solar geographic potential on building rooftop

1.1.2.1 Impact factors on rooftop solar irradiation

The spatio-temporal distribution of rooftop solar irradiation is influenced by various factors, including shading effects from buildings and mountains (Walch et al., 2020), rooftop slope and aspect (Mohajeri et al., 2018), and numerous urban morphological features (Sarralde et al., 2015), such as the Sky View Factor (SVF). In urban environments, the complexity of artificial and natural structures increases shaded areas on rooftops, thereby reducing their solar potential. Previous studies (Cheng et al., 2006; Robinson, 2006; Martins et al., 2014; Chatzipoulka et al., 2016; Mohajeri et al., 2016) have shown that these morphological features significantly affect the solar energy potential of buildings. For instance, Zhu et al. (2020) explored the relationship between solar capacity and urban morphology, demonstrating that urban morphological characteristics significantly influence solar capacity under varying weather conditions. Similarly, Poon et al. (2020) conducted a parametric study in Singapore to assess how urban morphological features correlate with annual average solar irradiation on rooftops and façades, finding that SVF is the most strongly correlated factor. Additionally, shading from nearby buildings and mountains directly and significantly reduces the received solar energy. Understanding the impact of these shades is crucial for accurately estimating and deploying photovoltaic (PV) arrays on rooftops. Li et al. (2015) highlighted the substantial influence of shading effects caused by building structures on installed power capacity and proposed a method for accurately computing shaded areas to estimate solar potential. Furthermore, the Digital Surface Model (DSM), along with rooftop slope and aspect, is crucial in calculating solar irradiation, as shown by Rich et al. (1994). In summary, urban morphological features, shading effects, DSM, and rooftop slope and aspect are key parameters in estimating rooftop solar irradiation. However, few studies have specifically investigated their impact on evaluating rooftop solar potential. Our study not only examines the correlation between these parameters

and rooftop solar irradiation but also develops an optimal machine learning model to explore their relationships with solar irradiation.

1.1.2.2 Morphological Tessellation

In urban morphology studies (Morganti et al., 2017; Yang et al., 2021; Chen et al., 2023), density is often defined by the ratio of footprint area to unbuilt space, with calculations typically based on a grid-defined boundary. Leng et al. (2020) utilized a 150m radius to examine urban morphological features such as building site cover, floor area ratio, building height, road network density, road height-width ratio, green space ratio, and total wall surface area. While this scale selection is grounded in the empirical evidence from previous studies (Wei et al., 2016; Javanroodi et al., 2018; Lima et al., 2018;), it lacks robust scientific backing and may not be universally applicable to other complex regions. To address this issue, Yong et al. (2017) analyzed the impact of spatial scale on estimation accuracy at resolutions of 100, 200, 300, 400, 500, and 600 meters. Their findings indicated that R square values improve as the spatial scale increases, suggesting that coarser scales yield better prediction accuracy. However, using a specific spatial scale to calculate density can result in average values that overlook site-specific and building-related morphological features.

To mitigate this limitation, Fleischmann et al. (2020) introduced the morphological tessellation (MT) method, which derives spatial units from building footprints for urban morphometric analysis. Their study applied the MT method to generate morphological tessellation cells across four different urban tissues (organic tissue of Niederdorf, compact tissue of Langstrasse, detached villas of Hottingen, and mixed post-war development of Friesenberg) and visually inspected these cells. The results suggested that the MT method could be effectively applied to similar urban tissues in other regions. Additionally, Boccalatte et al. (2022) employed the MT method to calculate morphological features and evaluate the impact of urban morphology on rooftop solar radiation in Geneva. Given the MT method's ability to assess the influence of each building on its surroundings and accurately calculate building-related density information, our study adopts this approach to determine morphological features related to building density.

1.1.2.3 Methods for estimating rooftop solar potential

Methods for estimating rooftop solar potential can be categorized into four types: sampling method, geostatistical method, physical modeling method, and machine learning method (Gassar et al., 2021). The sampling method involves calculating an estimate of the available rooftop areas for a selected region and then extrapolating this estimate to cover the entire area. For instance, Izquierdo et al. (2008; 2011) employed stratified statistical sampling to determine the technical potential for rooftop PV energy production in Spain. Their findings revealed that the total available rooftop area in Spain was approximately 571 km², with the potential to generate around 4% of the country's total electrical energy through PV systems. While sampling methods are useful for estimating available rooftop areas over large regions, they provide only rough approximations of rooftop solar potential and do not meet the requirements for highly accurate estimations of rooftop solar irradiation.

Geostatistical methods perform spatial statistical analysis to predict solar potentials through techniques such as spatial interpolation and statistical clustering. Fathizad et al. (2017) developed an air temperature-based model to estimate solar radiation and evaluated solar mapping performance using eight geostatistical methods: Inverse Distance Weighted, Global Polynomial Interpolation, Radial Basis Function, Local Polynomial Interpolation, Ordinary Kriging, Simple Kriging, Universal Kriging, and Empirical Bayesian Kriging. Their results indicated that the Radial Basis Function method was the most effective, with an R^2 of 0.904, Mean Absolute Error (MAE) of 3.02, and RMSE of 0.39%. Additionally, Mishra et al. (2020) utilized statistical clustering to determine the available rooftop areas for estimating solar potential in Uttarakhand, India. Their study revealed that 58% of rooftop areas receive solar radiation greater than 4 kWh/m²/day year-round, capable of generating 57% of the region's electrical energy consumption. While geostatistical methods focus on the total solar energy received and provide probabilistic estimations of solar potential, they, like sampling methods, are limited to rough evaluations and face challenges in offering accurate, high spatio-temporal resolution estimates for individual buildings.

Geographic Information System (GIS)-based physical modeling methods are regarded as

optimal for estimating rooftop solar irradiation due to their high accuracy and potential for automated application across various areas. For example, Saadaoui et al. (2019) assessed the solar PV potential on flat roofs in BenGuerir, Morocco, using GIS and photogrammetry, finding that more than 345 GWh of electricity could be generated annually. Similarly, Hong et al. (2017) developed a method for estimating hierarchical rooftop solar PV potential—encompassing physical, geographic, and technical aspects—using Hillshade analysis in Seoul. Their results showed that the Gangnam district had physical, geographic, and technical potentials of 9, 287, 982 MWh, 4, 964, 118 m², and 1, 130, 371 MWh, respectively. While these methods yield accurate results and consider multiple factors affecting rooftop solar irradiation, they require significant computation time for reliable estimations. For instance, Tabik et al. (2012) used a Gradient Ascent algorithm on a GPU-CPU heterogeneous system to compute maximum irradiation, taking 2.477 seconds to calculate a Digital Elevation Model (DEM) of 500 points. Consequently, these methods are more suited for micro-scale to medium-scale regions. Despite their accuracy and wide applicability, the significant execution time needed limits their use in large-scale regions.

In recent years, machine learning methods have gained popularity in studies related to solar potential estimation due to their advantages of fast computation, scalability, and ability to deliver high-accuracy and reliable results. For instance, Liao et al. (2022) proposed a method to estimate continuous land surface solar irradiation in Australia and China using four machine learning techniques: Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Regression, and Multilayer Perceptron (MLP). Their findings highlighted GBM as the most accurate model, achieving a coefficient of determination (R^2) > 0.7, with computation times under 10 seconds for processing extensive datasets. Similarly, Assouline et al. (2018) integrated solar models in GIS with the Random Forest algorithm to estimate rooftop solar potential at a resolution of 200×200 m² pixels in Switzerland. They estimated the total PV electricity production from building rooftops to be 16.29 TWh/year, capable of meeting 25.3% of the country's annual demand. Wang et al. (2018) introduced a PV power prediction model based on Gradient Boost Decision Trees, demonstrating good model interpretability, prediction accuracy, and stable error performance. Additionally, Babbar et al. (2021) employed Adaboost,

a hybrid of linear and non-linear machine learning models, for long-term solar power generation prediction, achieving superior performance with a MAPE of 8.88%. Compared to traditional methods, machine learning approaches excel in handling large datasets swiftly while maintaining high prediction accuracy. Thus, machine learning methods are increasingly recognized as suitable alternatives for assessing and estimating rooftop solar potential.

1.2 Research Gaps and Motivation

While numerous methods have been proposed for the estimation of solar potential, there remain several limitations in these approaches. The research motivations for this study are as follows:

1) Research gaps for estimation of solar physical potential:

- Although empirical methods for estimating solar irradiation have certain merits, they still have a weak capability to deal with a large geographical extent, such as an estimation covering the whole of Australia, China, and Japan. Physical methods generally combine with satellite images to estimate large-scale solar irradiation, while these images have a relatively low temporal resolution. In this regard, these methods are hard to meet the high accuracy requirement on solar irradiation estimation.

- Although traditional machine learning methods provide good estimation results, they are limited in their ability to investigate the impact of geographic variability on solar irradiation. GeoAI models have been proven to outperform traditional non-spatial machine learning models in several energy-related tasks. This is because GeoAI methods can overcome the limitations of geographic heterogeneity by integrating spatial characteristics with temporal characteristics. However, model interpretability remains a major challenge in GeoAI study due to its black-box nature. Additionally, current interpretable AI models have not adequately considered the impact of geographic heterogeneity on solar radiation. At present, few algorithms can simultaneously address the geographic heterogeneity of spatio-temporal data estimation and the interpretability issues in machine learning.

2) Research gaps for estimation of solar geographic potential:

- Morphological features, shade effects, DSM, and rooftop slope and aspect are significant parameters for estimating rooftop solar irradiation. However, quantifying these parameters poses a significant challenge. Furthermore, there are limited studies that simultaneously consider the multiple influences of building and terrain shadows, urban morphological parameters, meteorological conditions, and digital elevation models (DEM).

- Despite the physical methods used for estimating building rooftop solar irradiation show good performance with high accuracy, these methods have the limitations of low computational efficiency.

- Machine learning models used for estimating rooftop solar irradiation on buildings offer the significant advantage of rapid computation. However, these methods necessitate a substantial amount of ground truth data for modeling. The measurement of ground truth data for rooftop solar irradiation in large-scale regions still presents challenges due to the high cost of installing measurement equipment and limited access to rooftops.

- Machine learning methods used for estimating high temporal resolution rooftop solar irradiation also exhibit drawbacks, including limited interpretability and a reduced capacity to handle both static and dynamic data.

1.3 Research Objectives

This study proposes a hierarchical assessment framework for the machine learning-based estimation of physical solar potential and geographical solar potential. The specific research objectives of our study are detailed as follows:

1) A machine learning-based study of the estimation of land surface solar irradiation at a large-scale level

- To explore the optimal method among the traditional machine learning methods, which was used for estimating hourly/daily land surface solar irradiation in Australia, China, and Japan, using multi-source data, including AOT from Himawari-8 meteorological satellite images, COT from Himawari-8 meteorological satellite images, clear-sky solar irradiation (CSI), and meteorological data (i.e., air temperatures, humidity, wind, and atmospheric pressure).

- To derive solar maps in Australia, China, and Japan at various time scales (i.e., maximum and minimum monthly solar maps, seasonal solar maps, and annual solar maps), and analyze the distribution of the solar potential in these countries.

- To propose a novel deep learning method, the dual-gate Temporal Fusion Transformer for estimating land surface solar irradiation. This method overcomes the limitation of interpretability of the general machine learning methods and enables the integration of static geographical features with time-varying features.

2) A machine learning-based study of the estimation of rooftop solar irradiation at a city level

- To quantify the relevant parameters for estimating rooftop solar irradiation at a fine resolution, including, hourly shadow from buildings and terrain from 7 am to 5 pm, and 46 urban morphological parameters detailed in Appendix 1.

- To develop a fast and highly accurate estimation method for rooftop solar irradiation. This section contains the use of machine learning methods to estimate annual rooftop solar irradiation and the application of the proposed dual-gate Temporal Fusion Transformer to estimate hourly rooftop solar irradiation in Hong Kong.

1.4 Thesis Outline

This thesis employs the hierarchical framework to investigate the physical solar potential and geographical solar potential based on machine learning methods using multi-source data. This thesis was organized as follows, and the technical route map is shown in Figure 1.

Chapter 1 first introduces the importance of accurately estimating solar energy potential for achieving carbon neutrality. It then compares and analyzes the strengths and weaknesses of existing methods for estimating physical and geographical solar potential, considering both macro and micro perspectives. From a methodological viewpoint, it thoroughly reviews the current state of research on solar energy potential estimation techniques. Following analysis and discussion, a machine learning-based approach is selected as the technical route for conducting the specific research. The chapter also identifies gaps in existing studies and outlines the research objectives.

Chapter 2 proposes a simple and effective method for the estimation of land surface solar irradiation based on machine learning models using meteorological data, Himawari-8 satellite cloud and aerosol products, and solar observation data in Australia, China, and Japan. The estimation performance of solar irradiation based on four machine learning models, i.e., Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Regression, and Multilayer Perceptron (MLP), were compared in terms of accuracy and computational efficiency. The estimation of monthly, seasonal, and annual solar irradiation at nationwide levels was generated.

Chapter 3 proposes the Dual-gate Temporal Fusion Transformer (DGTFT), a novel interpretable deep learning network, to improve hourly land surface solar irradiation estimation. The ablation experiments were conducted to select the optimal network structure. Applied to datasets from Australia, China, and Japan, accurately estimated annual land surface solar irradiation maps were generated.

After investigating the distribution of physical solar potential at the large-scale level in Chapter 2 and Chapter 3, Chapter 4 and Chapter 5 give insights into the rooftop solar potential at the city level. Given the abundant solar resources identified in Hong Kong from the findings in Chapters 2 and 3, this city has been selected as the study case.

Chapter 4 proposes a parametric-based method to estimate annual rooftop solar irradiation at a fine spatial resolution. This chapter quantifies the parameters of the shadow from buildings and terrain and 46 urban morphological parameters, selects the significant parameters using the RF method for machine learning training, and uses the optimal machine learning methods among RF, Gradient Boost Regression Tree (GBRT), and AdaBoost to generate the annual rooftop solar irradiation map in Hong Kong.

Furthermore, Chapter 5 continues the exploration of rooftop solar potential by using the proposed DGTFT method to estimate hourly rooftop solar irradiation based on the data from Chapter 4.

Finally, Chapter 6 concludes with the major findings of the study, limitations, and scope for future research in solar potential estimation.

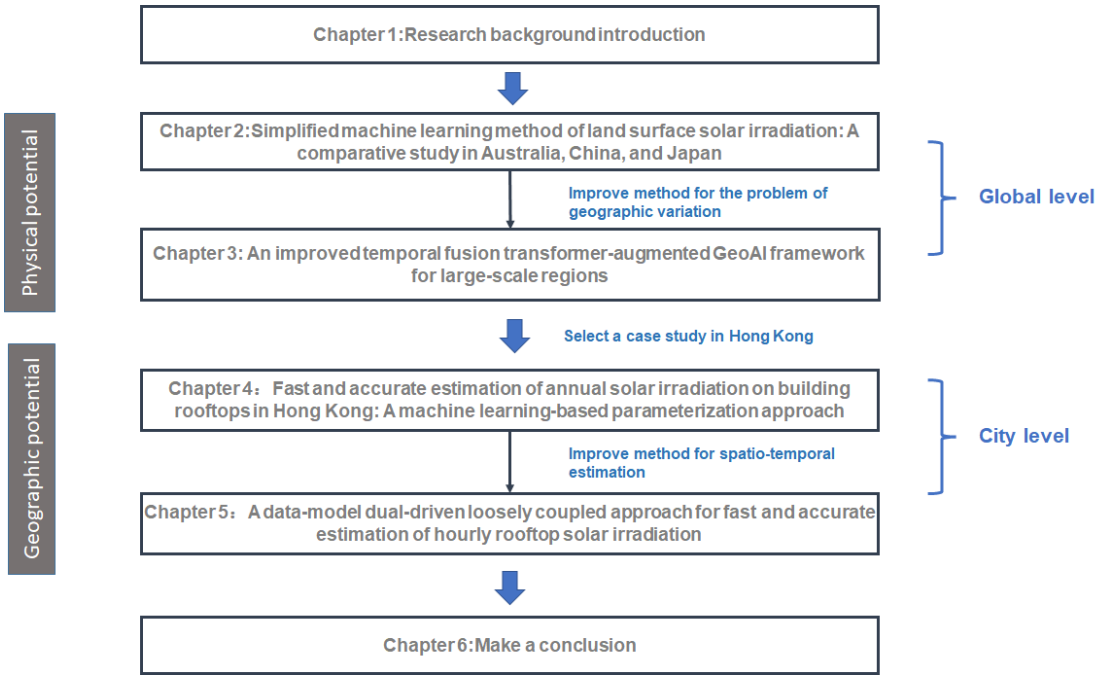


Figure 1 Technical route map

Chapter 2 Simplified estimation modeling of land surface solar irradiation

Solar irradiation maps are essential geospatial datasets utilized in various research fields. Accurately estimating the continuous distribution of solar irradiation over large areas is challenging with traditional interpolation or extrapolation methods that rely on a limited number of observation stations. To address this issue, the author proposed a method leveraging four machine learning models, Gradient Boosting Machine (GBM), Random Forest (RF), Support Vector Regression (SVR), and Multilayer Perceptron (MLP), to estimate spatially continuous land surface solar irradiation. Clear-sky solar irradiation data, determined by time and location, along with cloud optical thickness (COT) and aerosol optical thickness (AOT) retrieved from Himawari-8 meteorological satellite images, were used in conjunction with observation station data for training and evaluation. Additionally, air temperatures, humidity, wind, and atmospheric pressure were quantified and integrated into the models to account for weather effects on land surface solar irradiation. This comparative study collected six years of historical data to estimate solar distribution at a 5-km spatial resolution in Australia, China, and Japan. Based on metrics such as the coefficient of determination (R^2), normalized Root Mean Square Error (nRMSE), normalized mean bias error (nMBE), and time consumption (t), the results indicated that GBM achieved the highest accuracy with R^2 across all stations, followed by RF, SVR, and MLP. This suggests that the proposed method can provide accurate and reliable land surface solar irradiation estimates, compared to theoretical values unaffected by atmospheric obstacles. The annual solar distribution maps produced demonstrate that the proposed method is both simple and effective for large geographical regions and can be applied globally with similar datasets.

2.1 Study area and data

This section outlines the study areas and the datasets utilized as input and output parameters for the machine learning models designated for estimating solar irradiation. Leveraging the advantages of satellite images, such as continuity, extensive coverage, and public availability, this study employed the geostationary satellite Himawari-8 to gather AOT and COT data,

updated hourly. Given the strong correlation between meteorological data and solar irradiation (Rabehi et al., 2020), variables including maximum temperature, minimum temperature, average humidity, average wind speed, and average atmospheric pressure were incorporated as input parameters. To enhance estimation accuracy, solar irradiation under clear-sky conditions was also computed and used as an input parameter for training the machine learning models. Additionally, the study utilized a Digital Surface Model (DSM) with 1m resolution in Hong Kong and the Hong Kong polygon shapefile.

2.1.1 Study area

To comprehensively evaluate machine learning-based solar estimation, this study focused on three countries: Australia, China, and Japan, which span significant geographical extents in both the southern and northern hemispheres. These countries cover a broad range of latitudes and encompass diverse local climates (Table 1), providing an excellent opportunity to validate the robustness and generalization of the proposed method. The study utilized data from 27 stations that collected the necessary datasets over six continuous years, from 2015 to 2020. This included 13 stations in Australia (Figure 2), nine stations in China (Figure 3), and five stations in Japan (Figure 4). Table 1 presents the range of hourly observed solar irradiation in Australia and Japan, as well as the range of daily observed solar irradiation in China.

Table 1 Climates and ranges of observed solar irradiation of the 27 meteorological stations

Country	Station Name	Station ID	Climate	Range of observed solar irradiation (kWh/m^2)
Australia	Adelaide	S1	Mediterranean	0-1.38
	Alice Springs	S2	Subtropical hot desert	0-1.48
	Broome	S3	Hot semi-arid	0-1.44
	Cape Grim	S4	Temperate oceanic	0-1.31

	Cocos Island	S5	Tropical rainforest	0-1.37
	Darwin	S6	Tropical savanna	0-1.45
	Geraldton	S7	Mediterranean	0-1.44
	Kalgoorlie-Boulder	S8	Semi-arid	0-1.39
	Learmonth	S9	Hot semi-arid	0-1.36
	Melbourne	S10	Temperate oceanic	0-1.41
	Rockhampton	S11	Humid subtropical	0-1.51
	Townsville	S12	Tropical savanna	0-1.57
	Wagga	S13	Humid subtropical	0-1.43
	Beijing	S1	Humid continental	0-9.66
	Guangzhou	S2	Humid subtropical	0.24-7.81
	Harbin	S3	Humid continental	0.13-12.13
China	Kau Sai Chau	S4	Humid subtropical	0-1.09
	King’s Park	S5	Humid subtropical	0-1.08
	Shanghai	S6	Humid subtropical	0.16-8.65
	Urumqi	S7	Continental cold	0-11.75

			semi-arid	
	Wenjiang	S8	Humid subtropical	0.21-8.39
	Wuhan	S9	Humid subtropical	0.14-8.40
	Fukuoka	S1	Humid subtropical	0.00-1.09
	Ishigakijima	S2	Humid subtropical	0.00-1.14
Japan	Minamitorishima	S3	Tropical savanna	0.00-1.10
	Sapporo	S4	Humid continental	0.00-1.14
	Tsukuba	S5	Humid continental	0.00-1.12

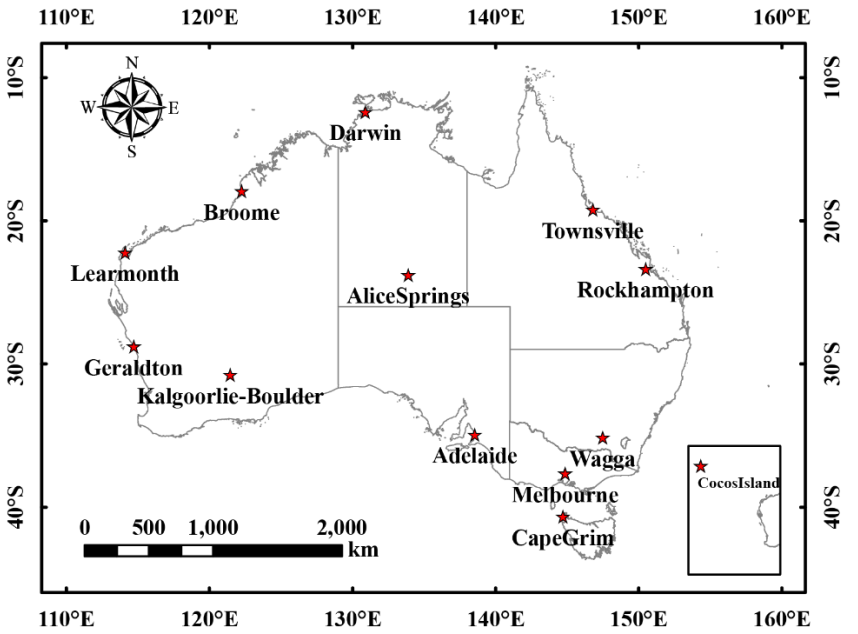


Figure 2 Distribution of the 13 stations in Australia

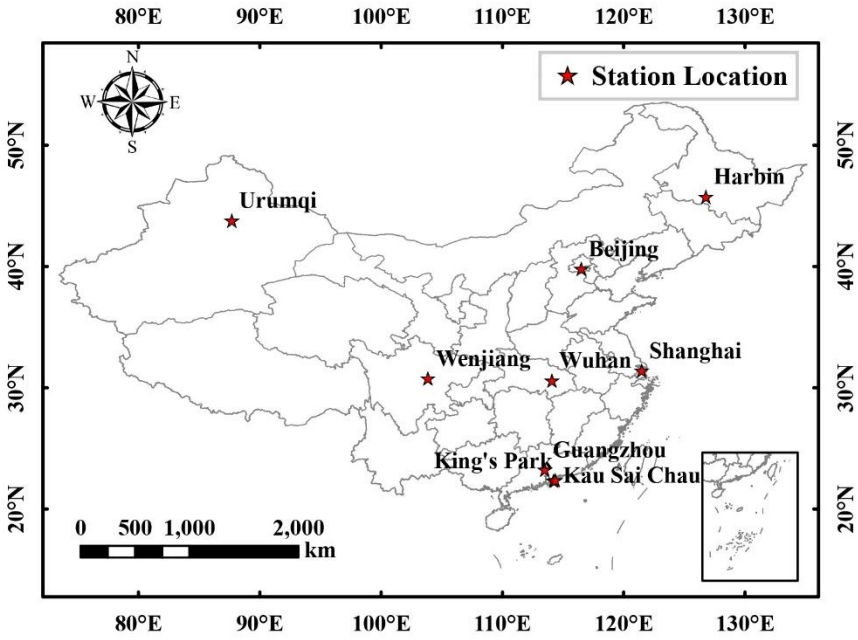


Figure 3 Distribution of the nine stations in China

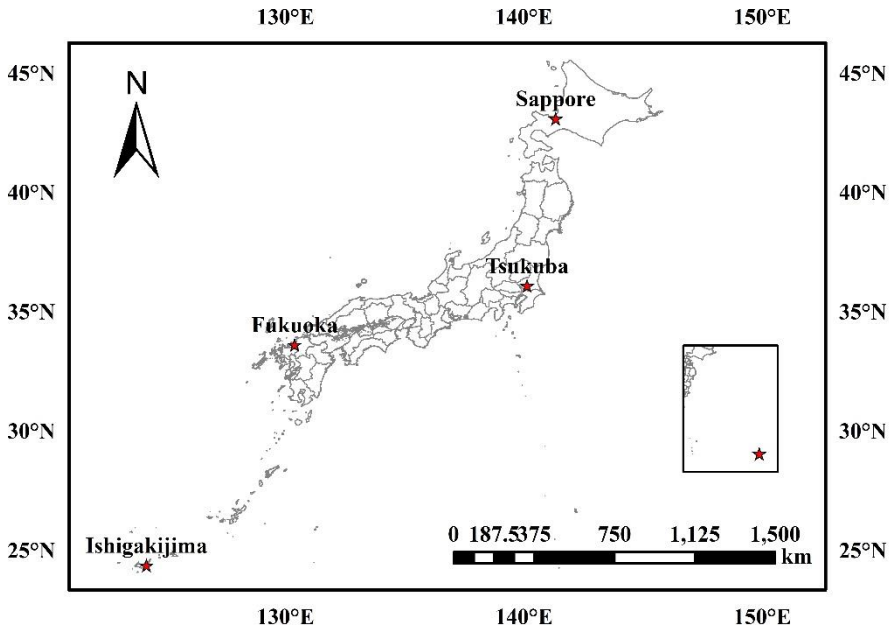


Figure 4 Distribution of the five stations in Japan

2.1.2 Data

2.1.2.1 Himawari-8 satellite products

Himawari-8, a geostationary weather satellite operated by the Japan Meteorological Agency (Japan Meteorological Agency, 2021), covers a vast geographical area from 60°S to 60°N and 80°E to 160°W, encompassing Oceania, Southeast Asia, and the Western Pacific. The Advanced Himawari Imager (AHI) on board Himawari-8 provides AOT and COT data. Satellite images in NetCDF format are freely accessible via the JAXA Himawari Monitor P-Tree System (Copernicus Global Land Service, 2021). This study utilized Himawari-8 level-2 AOT and COT data, which have a temporal resolution of 10 minutes and a spatial resolution of 5 km, spanning from 2015 to 2020. Huang et al. assessed the quality of Himawari-8 cloud products and reported high consistency due to active Radar-LiDAR observations (Huang et al., 2019). Additionally, Gao et al. confirmed that Himawari-8 provides reliable aerosol products for environmental research (Gao et al., 2021).

2.1.2.2 Calculated hourly clear-sky solar irradiation

Hourly clear-sky solar irradiation (CSI) for the 27 stations was computed using the Python library Pysolar. This library employs the Masters' algorithm (Masters, 2013) for calculating solar irradiation and the algorithm by Reda and Andreas (Reda and Andreas, 2004) for

determining solar position. These algorithms utilize parameters such as longitude, latitude, and specific time to compute the Sun's location in the sky, solar irradiation under clear-sky conditions, and the irradiation reaching both horizontal and inclined surfaces on the ground (Bishop et al., 1997; Gilbert et al., 2004). The resulting hourly CSI dataset included attributes like station name, time, and hourly clear-sky solar irradiation from 2015 to 2020 for Australia, China, and Japan. Given that the solar irradiation data from Chinese stations have a daily temporal resolution, the estimated hourly solar irradiation was aggregated daily to maintain consistency.

2.1.2.3 Observed land surface solar irradiation

Surface solar irradiation measured by these stations served as the reference data for evaluating machine learning-based solar irradiation estimation models. In Australia, solar irradiation was monitored at 13 meteorological stations (Figure 2 and Table 1), operated by the Australian Government Bureau of Meteorology (Australian Government Bureau of Meteorology, 2022). The original data was collected at one-minute intervals, but for consistency with other datasets, this study resampled the data to hourly intervals. In China, solar irradiation data were categorized into two types: daily updates (Figure 3 and Table 1), which represent the highest temporal resolution available from the China National Meteorological Information Center (China National Meteorological Information Center, 2021), and hourly updates from the Hong Kong Observatory (Hong Kong Observatory, 2021). For Japan, hourly solar irradiation data were obtained from five stations operated by the Japan Meteorological Agency: Sapporo station, Tsukuba station, Fukuoka station, Ishigakijima station, and Minamitorishima station. This agency provides direct solar irradiation data and diffuse solar irradiation data separately, without global solar irradiation data. Hence, the sum of direct solar irradiation and diffuse solar irradiation was used as the proxy for global solar irradiation in this study.

2.1.2.4 Meteorological data

Some researchers (Dahmani et al., 2016; Deo et al., 2016; Biazar et al., 2020; Rabehi et al., 2020; Zang et al., 2020) suggest that meteorological data are commonly used as the input parameters to estimate solar irradiation. Therefore, we employed meteorological data as the

input parameters, including the maximum temperature (MaxT), minimum temperature (MinT), average humidity (H), average wind speed (WS), and average atmosphere pressure (P). The hourly meteorological data in Australia, China, and Japan are purchased from the OpenWeather website (Openweather website, 2022).

2.2 Methodology

2.2.1 Construction of the datasets

The dataset in each station consists of meteorological data, AOT, COT, CSI, and the observed land surface solar irradiation from 2015 to 2020. The original AOT and COT data have a temporal resolution of 10 minutes, whereas solar irradiation data is updated daily in China and hourly in Australia and Japan. To obtain the same resolution for building the machine learning models, all data in each country are aggregated to the same temporal resolution, with the lowest resolution serving as the benchmark, i.e., daily in China and hourly in Australia and Japan.

2.2.2 Data pre-processing

Pre-processing operations have been conducted to train machine learning models. First, missing values and default values of all datasets have been checked and removed. In addition, due to the inconsistency of data sources between the two countries, solar irradiation was first transformed to the same unit (kWh/m^2). Note that the temporal resolution of solar irradiation in mainland China was daily updated while the data in Australia and Japan was hourly updated. Finally, in this study, the datasets were divided into training datasets and validation datasets by using K-fold cross validation (Rodriguez et al., 2009). Specifically, the original data set was randomly divided into K equal-sized sub-datasets. Of the K sub-datasets, a single sub-dataset was employed as the validation data to test the performance of machine learning, and the remaining K-1 sub-datasets were used as the training data. In this study, we set K equalling ten.

2.2.3 Constructing machine-learning based estimation models

In order to achieve accurate solar irradiation estimations with high computational efficiency, this study evaluated four different machine learning models to identify the most optimal one

for developing a reliable estimation model. Given the variability in estimation accuracy across different regions, a comprehensive comparison of the models was essential. All computations were conducted using the Python IDE, PyCharm (Pycharm, 2022). The Sklearn package (Sklearn, 2022) was utilized for training the machine learning models, while the Scipy package (Scipy, 2022) facilitated the calculation of estimation accuracy. To determine the best parameter values for the models, the GridSearchCV (GridSearchCV, 2022) function from the Sklearn package was employed.

2.2.3.1 Construction of the Support Vector Regression

To estimate land surface solar irradiation, the SVR (Awad et al., 2015) method was employed for regression. Our study followed these steps in the SVR process. First, meteorological data along with AOT, COT, and CSI data were used as independent variables to input into the model, while solar irradiation data from observation stations served as the label variables for the model's output. The regression model was then trained using a specific training function. The desired results were achieved by adjusting various kernel functions, gamma values, and the C parameter. The dataset was structured as $\{(X_i, Y_i), i=1,...,n\}$, where X_i represents the vector of meteorological, AOT, COT, and CSI data, Y_i corresponds to the solar irradiation data from the stations, and n is the total number of data points. In SVR, a linear function is defined as follows:

$$f(x) = \omega \cdot x + b \tag{2.2.3.1.1}$$

where ω is the weight vector and b is the constant. The coefficients ω and b are estimated by the minimization process:

$$y = min\frac{1}{2}\|\omega\|^2 + C\sum_{i=0}^{n}(\xi_i + \xi_i^*) \tag{2.2.3.1.2}$$

$$\begin{cases} y_i - \omega \cdot x_i - b \le \omega + \xi_i, i = 1,2,\dots,n \\ \omega \cdot x_i + b - y_i \le \omega + \xi_i^*, i = 1,2,\dots,n \\ \xi_i, \xi_i^* \ge 0, i = 1,2,\dots,n \end{cases} \tag{2.2.3.1.3}$$

where ξ and C are the prescribed parameters, and ξ_i and ξ_i^* are positive slack variables.

Lagrangian multipliers and Karush-Kuhn-Tucher (KKT) optimizing conditions are applied in the linear regression function as presented below:

$$f(x) = \sum_{i \in SVs}(a_i - a_i^*)(x_i, x) + b \quad (2.2.3.1.4)$$

where ai and a_i^* are Lagrangian multipliers.

2.2.3.2 Construction of the Random Forest

Random forest (Segal et al., 2004) is a flexible and easy ensemble learning method, which can usually obtain robust results for classification and regression tasks. Therefore, RF was employed to estimate the land surface solar irradiation. In this study, the input dataset was $\{X_i$, i=1,...,m$\}$ and the output dataset was $\{Y_i$, i=1,...,m$\}$, where X_i denotes the vector of meteorological data, AOT, COT and CSI data, Y_i is the solar irradiation of stations, and m denotes the number of datasets. On this basis, this study performed the RF regression model with the following three steps.

(1) Bootstrap sample method was employed to generate a training dataset by randomly drawing with replacement m samples, where m is the size of the original training dataset.

(2) A multitude of decision trees was constructed at training time and outputting the class that is the mode of mean prediction of the individual trees.

(3) After repeating step (2) for n times, we can obtain a number of n regression trees to generate the random forest. For any regression tree, the mean error of all the regression trees can be calculated for obtaining an unbiased estimation of the random forest. The calculation formula is as follows:

$$Y(x_i) = \frac{1}{n}\sum_{i=1}^{n} T_n(X_i), n = 1,2, \dots, n \quad (2.2.3.2)$$

where Tn denotes a regression tree, and *n* is the number of regression trees.

2.2.3.3 Construction of the Multilayer Perceptron

Artificial Neural Networks (ANNs), inspired by biological neural networks, can learn and generalize data relationships to predict trends. Among the most common ANN structures, the Multilayer Perceptron (MLP) (Murtagh et al., 1991) includes three layers: an input layer for structured meteorological data, CSI, AOT, and COT; a hidden layer using the Sigmoid function as the activation function; and an output layer providing estimated surface solar irradiation (see Figure 5). This study employed an MLP with a Back Propagation (BP) algorithm for training, involving forward data flow calculation and backward error propagation. The input layer processed the solar irradiation datasets, the hidden layer adjusted network weights for the regression model, and the output layer held the estimated irradiation data. If the output did not match the ground truth, weight adjustments were made based on an error function derived from the backward propagation algorithm. The neural network was continuously optimized by adjusting the weight parameters until the error fell below a predetermined threshold.

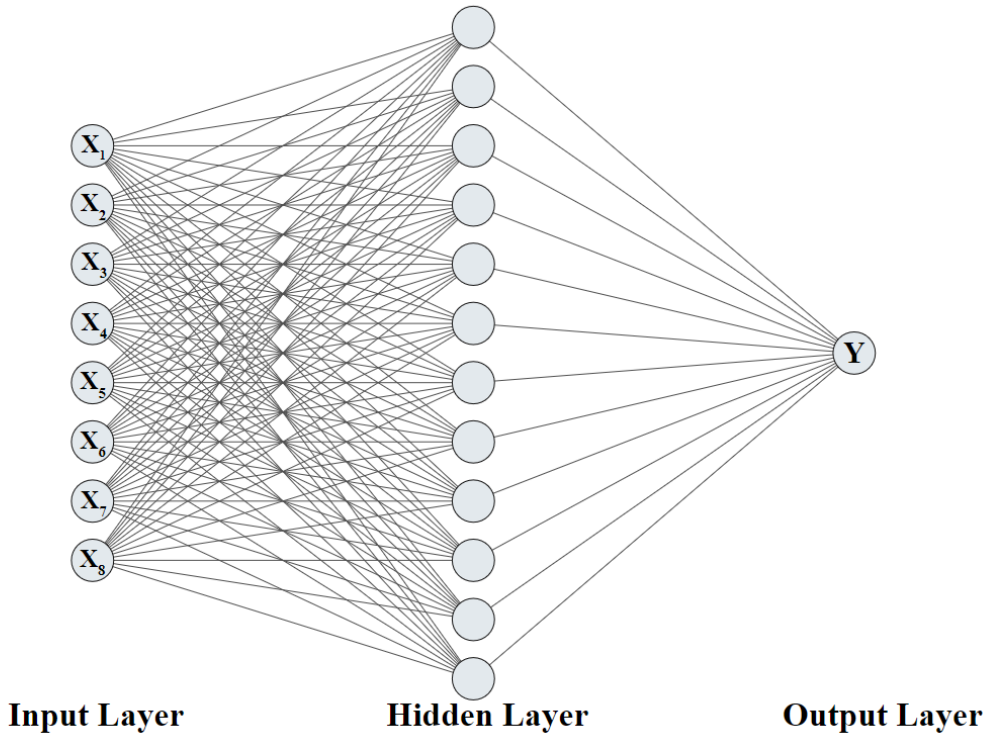


Figure 5 The architecture of MLP

2.2.3.4 Construction of the Gradient Boosting Machine

In this study, Gradient Boosting Machine (GBM) (Friedman et al., 2001) was employed to estimate land surface solar irradiation. GBM is a type of Boosting algorithm used for creating regression models. It builds an additive model by introducing a new decision tree at each iteration, thereby minimizing the deviation in the loss function. The steps for implementing the GBM model were as follows:

(1) Given a training dataset {(x_i, y_i), i=1,..., n} and the loss function L(y, $F(x)$), where xi was the vector of meteorological data, AOT, COT, and CSI data, y_i is corresponding solar irradiation of stations, and n denotes the number of datasets. The model was initialized using the fixed value γ:

$$F_0(x) = arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \qquad (2.2.3.4.1)$$

(2) Calculation pseudo-residuals rim, the formula is as follows:

$$r_{im} = \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right] F(x) = F_{m-1}(x), (i = 1,2,3, \dots, n) \quad (2.2.3.4.2)$$

(3) Calculation γ_m to solve the optimization problem:

$$\gamma_m = arg \min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i + \gamma h_m(x_i))) \quad (2.2.3.4.3)$$

where $h_m(x)$ denotes pseudo-residuals for the decision tree, the formula is as follows

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (2.2.3.4.4)$$

2.2.3.5 Estimation surface solar irradiation based on the optimal model

This study employed four evaluation indicators to evaluate the estimation accuracy of each model, namely a coefficient of determination (R^2), normalized Root Mean Square Error (nRMSE), normalized mean bias error (nMBE), and consumption of time (t). Specifically, nRMSE and nMBE were calculated as follows:

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \quad (2.2.3.5.1)$$

$$nMBE = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{y}_i - y_i}{y_i} \quad (2.2.3.5.2)$$

where n is the number of data, $\hat{y}_i$ denotes estimation value, and y_i is the actual value.

2.3 Results

The four machine learning models were used to evaluate the estimation accuracy at each station independently based on the four evaluation indicators. Through comprehensive comparison, the optimal machine learning model was selected for estimating surface solar irradiation in Australia, China, and Japan.

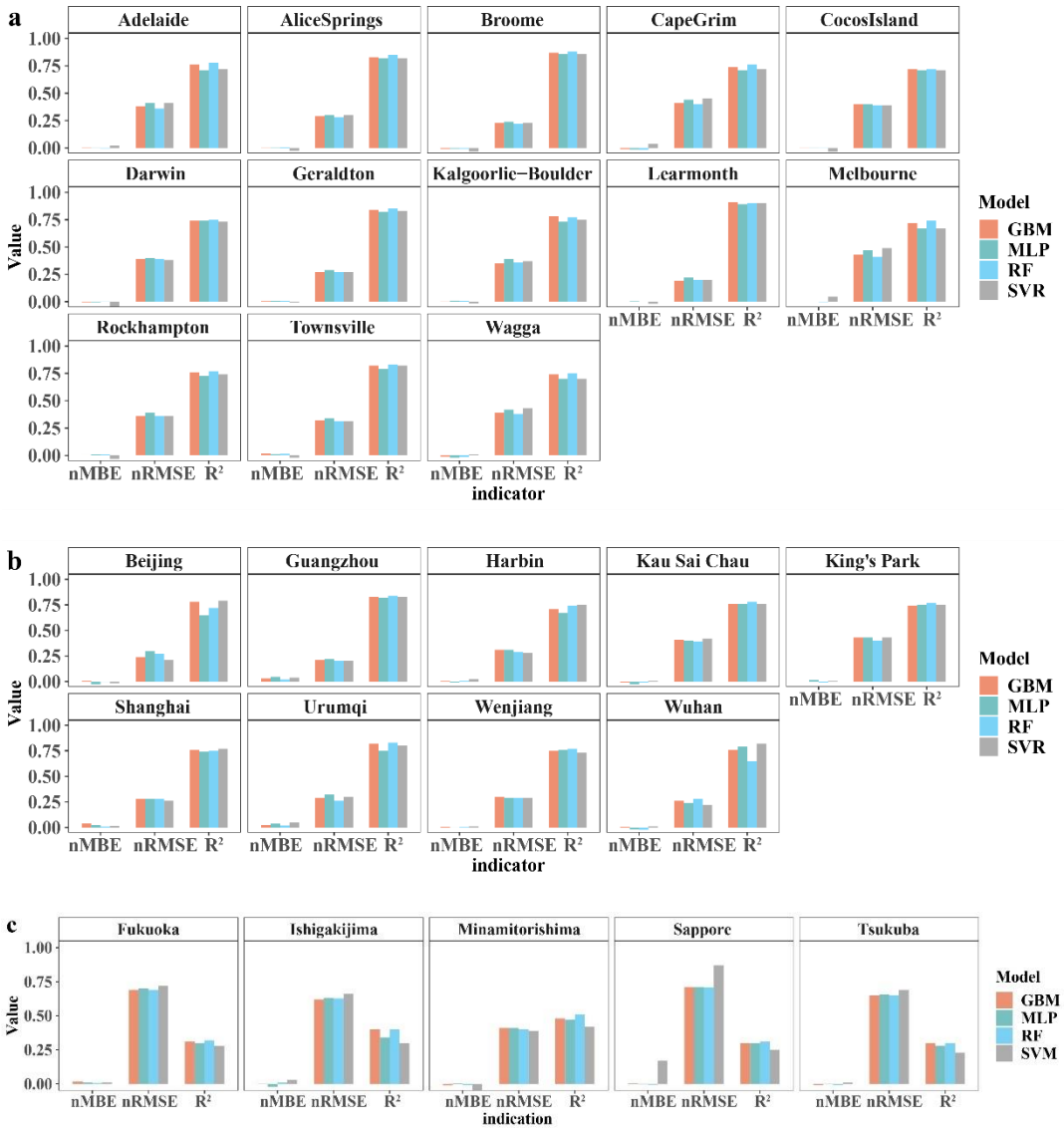
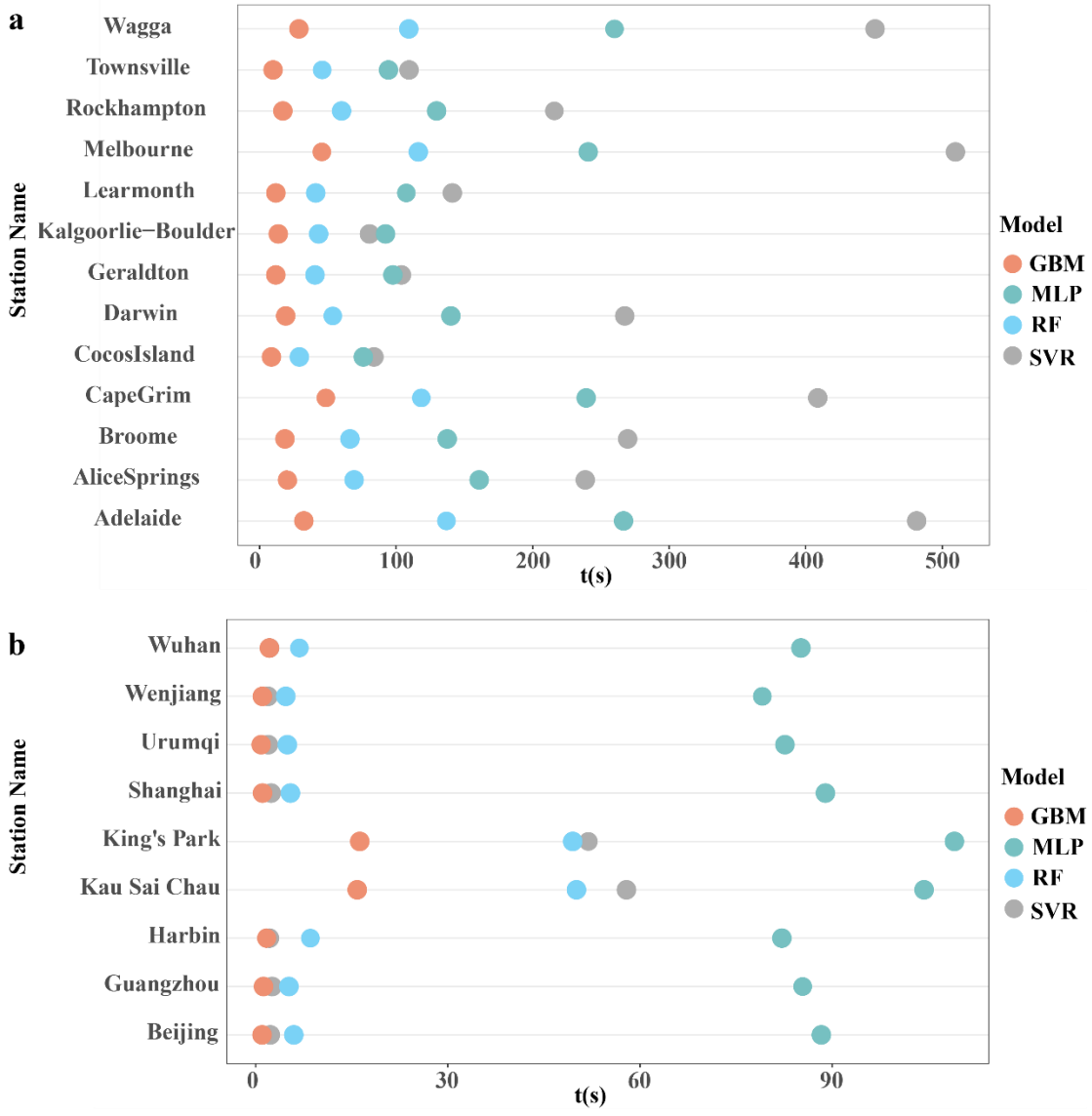


Figure 6 Estimation accuracy of the four machine learning models using R^2, nRMSE, and nMBE in all stations. (a) Results in Australia. (b) Results in China. (c) Results in Japan

2.3.1 Accuracy assessment of the models

Figure 6 systematically compares the estimated accuracy based on R^2, nRMSE, and nMBE in all 27 stations. Overall, it is found that the four models have similar estimation performance, and the performance in Australia and China is better than that in Japan. Specifically, all stations in Australia and China have $R^2 \geqq 0.7$ in both countries using the GBM model, and the proportions of the stations are about 38% for Australia and about 22% for China when $R^2 \geqq 0.8$. Besides, the nMBE values are significantly low in all stations, and the nRMSE values are between 0.2 and 0.4 only. The results suggest that the estimation models are reliable with high

estimation accuracy, which indicates that the proposed method can effectively estimate land surface solar irradiation over large regions. In contrast, the R^2 values of all stations in Japan using the GBM model are about 0.5, and corresponding nMBE values and nRMSE values are higher than those in Australia and China. This result suggests the estimation accuracy in Japan is lower than that in Australia and China. From the other perspective, Figure 7 summarizes the computation time of the four machine learning models in each station, which presents that the GBM model achieves the shortest time consumption. This suggests that GBM is outperformed for estimation accuracy and computational efficiency, especially for extensive computation when there are a large number of stations confined in a small area.



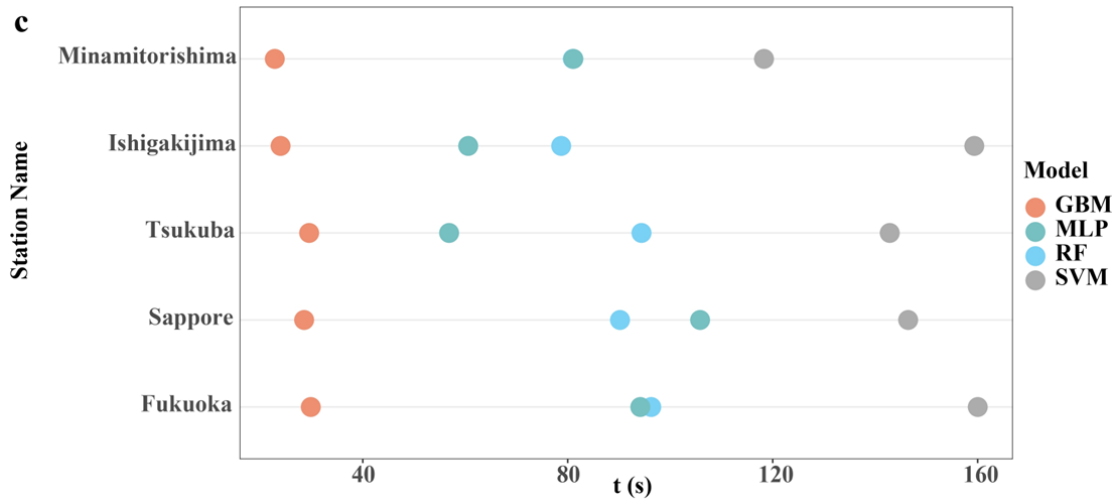
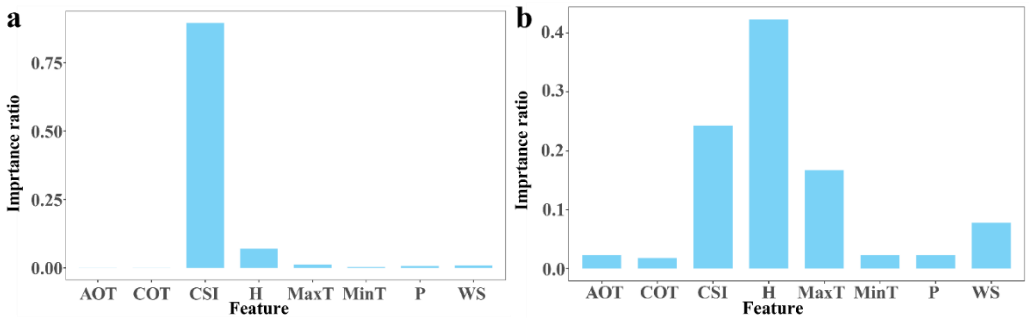


Figure 7 Computation time of the four machine learning in all stations. (a) Results in Australia. (b) Results in China. (c) Results in Japan

2.3.2 Feature importance analysis for the input parameters

Furthermore, the feature importance analysis is conducted to evaluate the impacts of each parameter on the estimation models (Figure 8). It shows that CSI is significantly larger than the second most important feature of H for estimating solar irradiation in Australia, leaving the rest features almost ignorable. This indicates that Australia has stable and solar favourable meteorological conditions, which thus have weak impacts on solar estimation. Likely, CSI is the most important factor in Japan, following by the feature of H. In contrast, the top three impact features are H, CSI, and MaxT in China, suggesting that the land surface solar irradiation is comprehensively affected by the meteorological features.



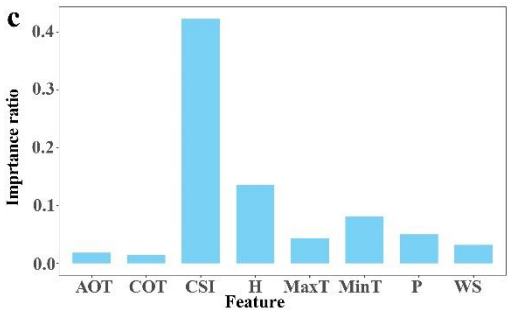
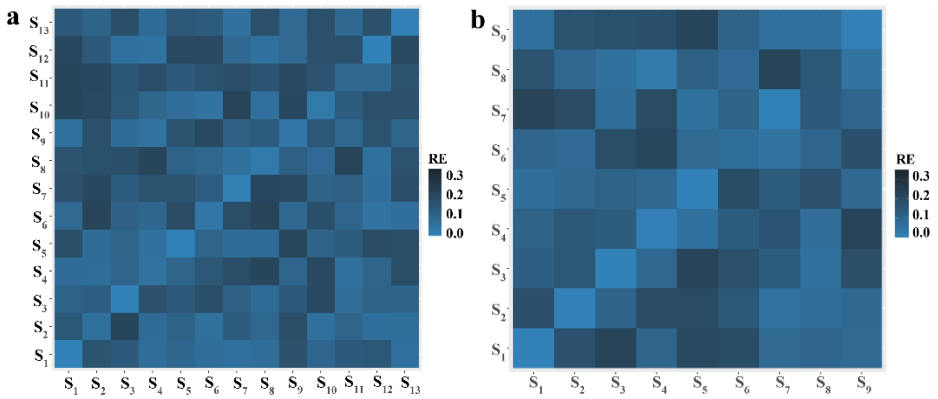


Figure 8 Importance ratios for the input features. (a) Australia. (b) China. (c) Japan

2.3.3 Generation of the land surface solar irradiation

To create seasonal and annual land surface solar irradiation maps at a 5-km spatial resolution in the three countries in 2020, the GBM model is used because it has achieved the highest estimation accuracy in both countries. The meteorological, COT, AOT, and CSI images are well prepared and used as the input parameters of the trained model. In addition, a set of meteorological images are obtained by using the Kriging interpolation method. Since the trained GBM model based on each observation station has relatively high accuracy as presented in Figure 6, this study used all the trained models to create the solar irradiation maps over the whole territory of Australia, China, and Japan.



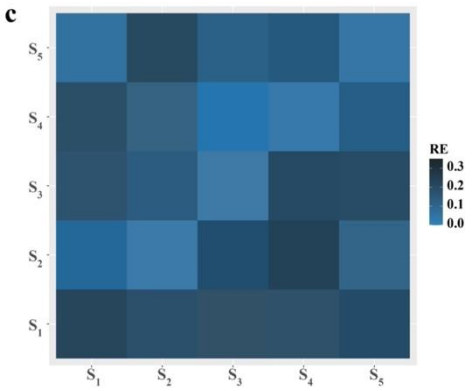


Figure 9 The relative errors between the estimated values and measured values in each station for each solar irradiation map. (a) There are 13 stations in Australia that correspond to a 13×13 matrix. (b) There are nine stations in China that correspond to a 9×9 matrix.(c) There are five stations in Japan that correspond to a 5×5 matrix

To systematically evaluate the accuracy of each created solar irradiation map, this study investigated the relative errors between the estimated values and correspondingly measured values located at all the stations in each solar irradiation map. Figure 9 shows that the relative errors in all stations are between 0.1 and 0.2, which suggests that the estimation results in all stations are accurate. Therefore, the mean values of all estimation maps were calculated and used as the final estimated solar irradiation map in the three countries. To avoid extremely big data computation, the solar irradiation on the middle day of each month is considered as the daily mean irradiation of that month, so that the monthly, seasonal, and annual solar irradiation can be accumulated over the corresponding time interval in each country.

2.3.3.1 Maximum and minimum monthly land surface solar irradiation

Figure 10, Figure 11, and Figure 12 show the maximum and minimum horizontal surface global solar irradiation in Australia, China, and Japan, respectively. Overall, the solar distribution in January is significantly higher than that in August in Australia, whereas the maximum solar distribution is in August and the minimum solar distribution is in January in China and Japan. In Australia, solar irradiation gradually increases from the northwest region to the southeast region in August (Figure 10(a)), with monthly values ranging from 171.78 to 76.08 kWh/m^2, while the irradiation in the central region is lower than in the other regions in January, (Figure 10(b)), with monthly

values ranging from 200.18 to 95.12 *kWh/m*^{2}. In China, solar irradiation in the southeast and central regions is lower than in other regions in January (from 69.88-147.98 *kWh/m*^{2}). In contrast, the irradiation is overall high in the whole country in August, with only part of the central region relatively low (from 97.56-223.89 *kWh/m*^{2}). In Japan, solar irradiation in the northeast is higher in other regions in January (from 86.98-189.10 *kWh/m*^{2}), and the irradiation is similar in the whole country in August (from 96.38-209.19 *kWh/m*^{2}).

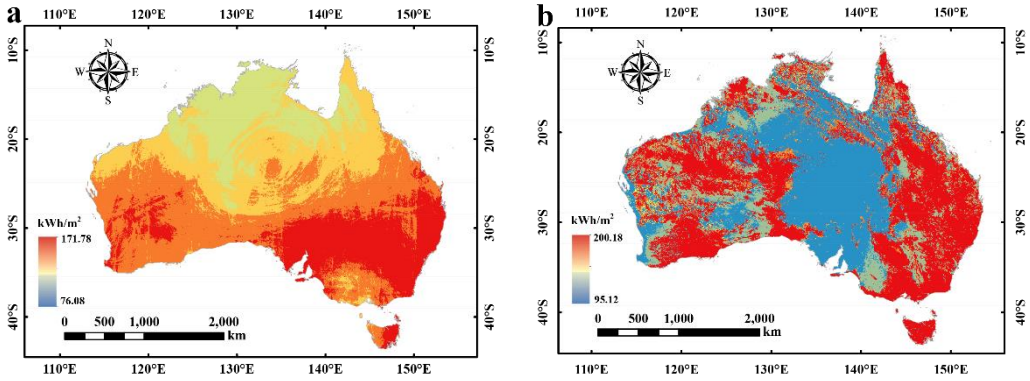


Figure 10 Distribution of the maximum and minimum horizontal surface global solar irradiation in Australia.(a) Distribution in August. (b) Distribution in January

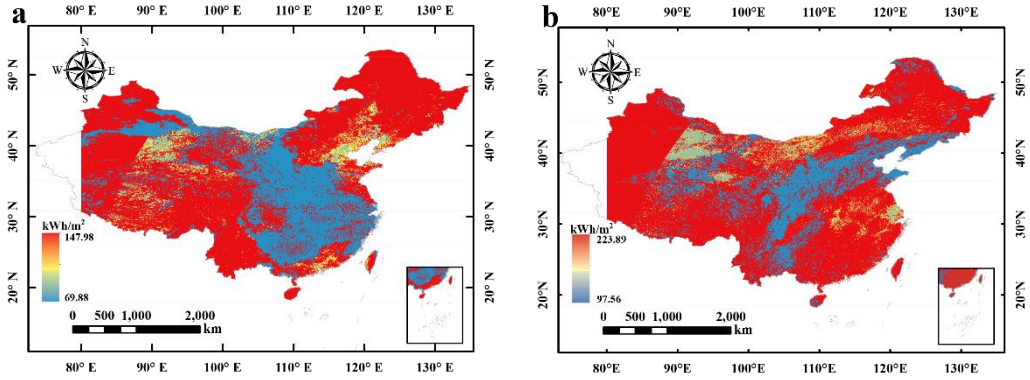


Figure 11 Distribution of the maximum and minimum horizontal surface global solar irradiation in China. (a) Distribution in January. (b) Distribution in August

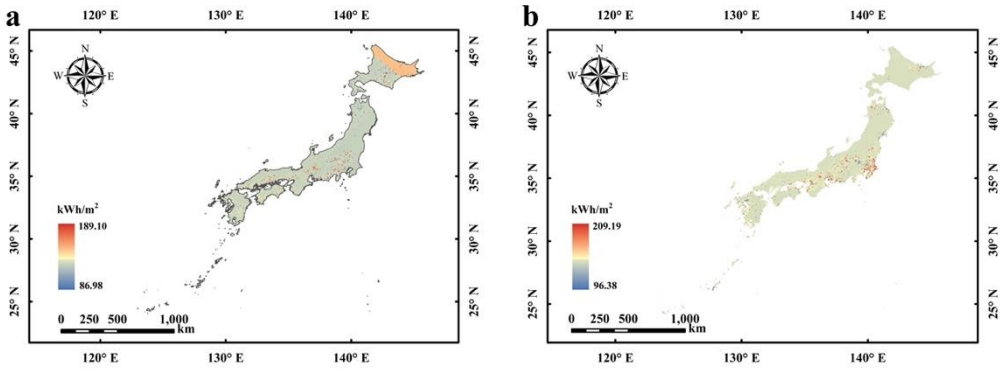
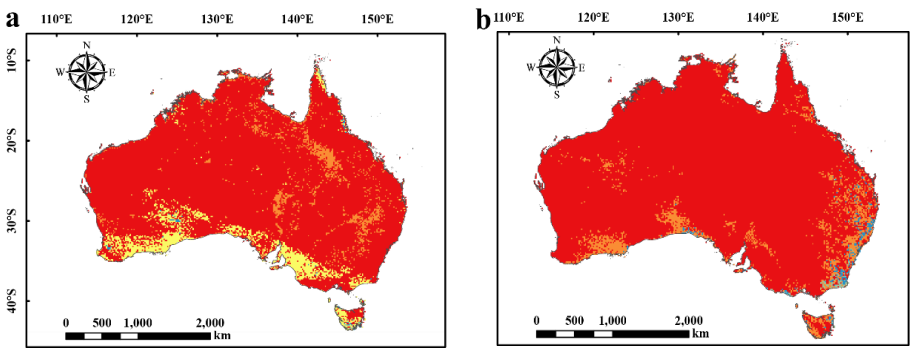


Figure 12 Distribution of the maximum and minimum horizontal surface global solar irradiation in Japan. (a) Distribution in January. (b) Distribution in August

2.3.3.2 Seasonal land surface solar irradiation

Furthermore, seasonal land surface solar irradiation maps were created for Australia (Figure 13), China (Figure 14), and Japan (Figure 15). Overall, the highest solar irradiation values are in summer in the three countries, followed by those in spring, autumn, and winter. The solar irradiation values in all seasons in Australia exhibit the narrow distribution, whereas those in China give the wide distribution. Figure 13 shows that Australia has an insignificant change in solar distribution during the four seasons, and most areas in Australia have a large amount of solar energy nearly 632 kWh/m^2. In China, solar irradiation in western and northeastern regions maintains a high level near 535 kWh/m^2 all year round, whereas, for southeastern regions in spring and summer, it is higher than that in autumn and winter. In Japan, solar irradiation in the southwest is higher than that in the northeast in Summer and Autumn, while the figure in the southwest is lower than that in the northeast in Spring and Winter.



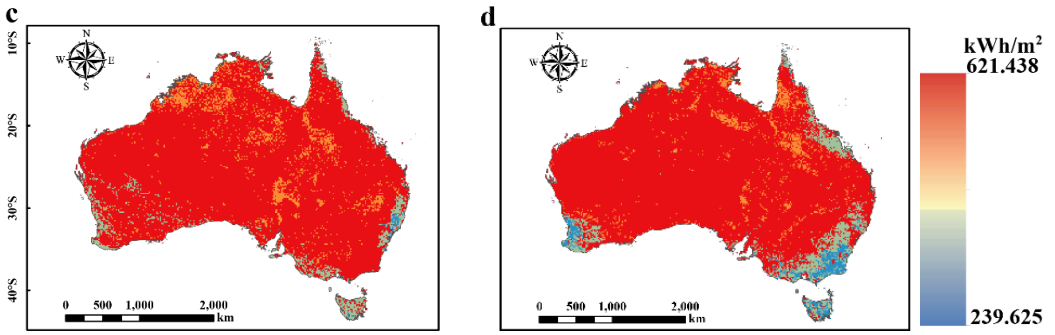


Figure 13 Seasonal distribution of land horizontal surface global solar irradiation in Australia. (a) The irradiation in spring (September to November). (b) The irradiation in summer (December to February). (c) The irradiation in autumn (March to May). (d) The irradiation in winter (June to August)

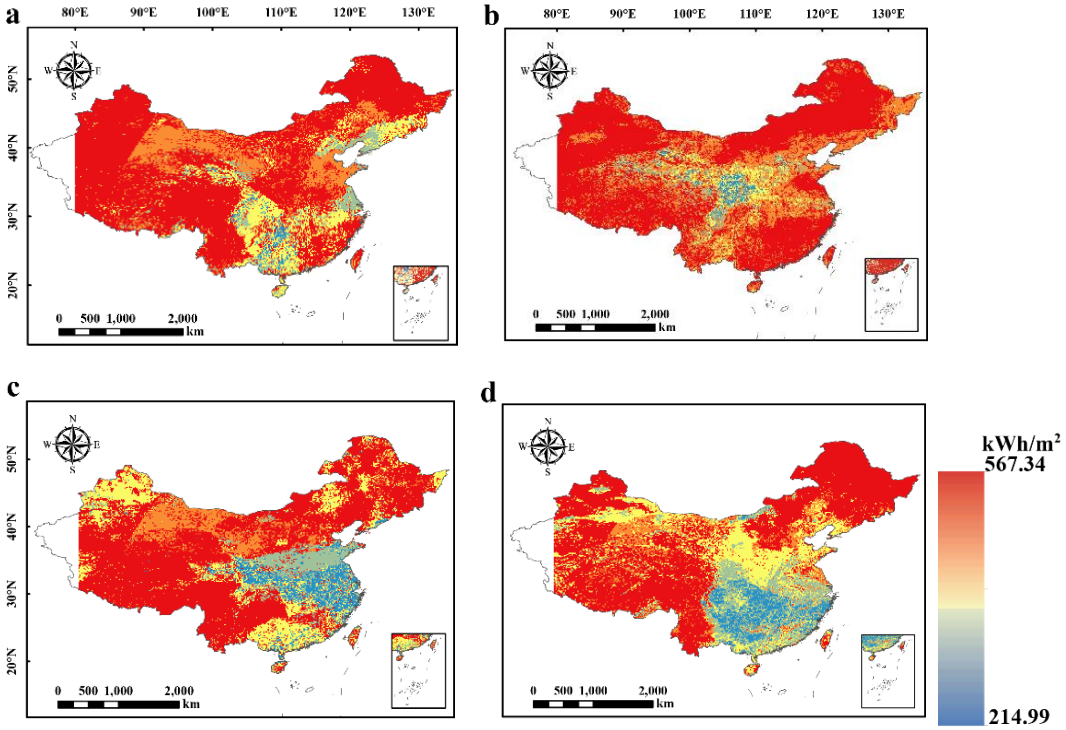
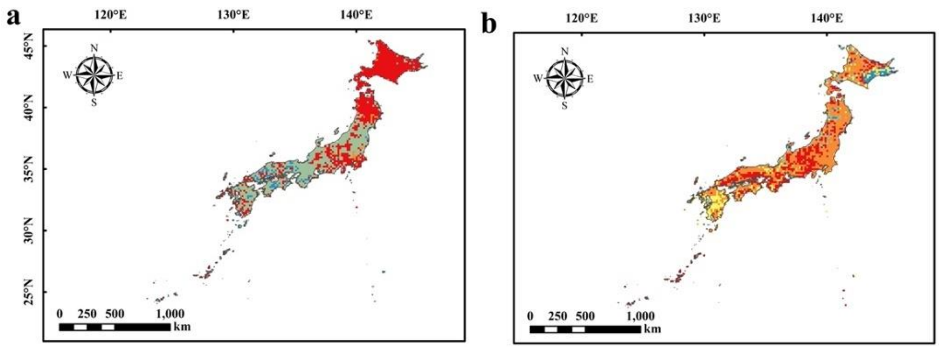


Figure 14 Seasonal distribution of land horizontal surface global solar irradiation in China. (a) The irradiation in Spring (March to May). (b) The irradiation in Summer (June to August). (c) The irradiation in Autumn (September to November). (d) The irradiation in Winter (December to February)



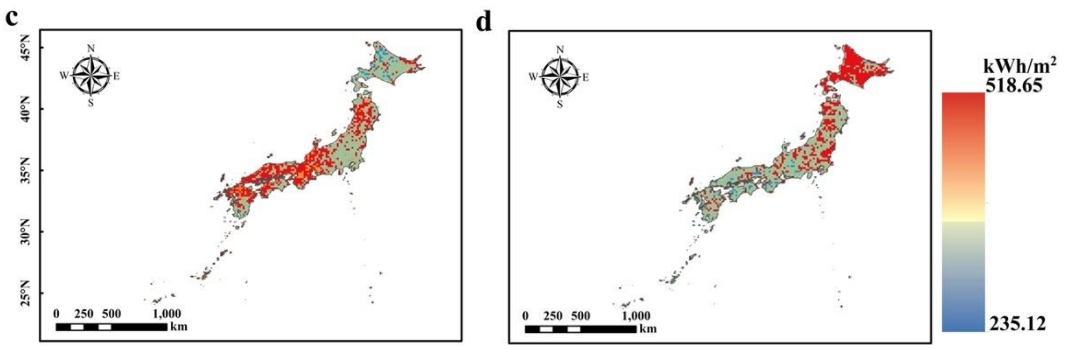


Figure 15 Seasonal distribution of land horizontal surface global solar irradiation in Japan. (a) The irradiation in Spring (March to May). (b) The irradiation in Summer (June to August). (c) The irradiation in Autumn (September to November). (d) The irradiation in Winter (December to February)

2.3.3.3 Annual land surface solar irradiation

Lastly, the annual land surface solar irradiation was estimated by accumulating four seasonal solar energy. Overall, the total irradiation in Australia (Figure 16 (a)) is higher than that in China (Figure 16 (b)), while the figure in Japan is the lowest (Figure 16 (c)). In detail, the vast majority of areas in Australia have abundant solar resources, suggesting that Australia is feasible to promote solar energy in most areas. In comparison, the distribution of the annual irradiation in China presents a gradual decrease from the northeast to the southwest. This indicates that southwest China has a relatively thick cloud cover that hinders the receiving of solar energy, meaning that latitude may not be a conclusive factor for using solar energy in large regions. In addition, the heterogeneous distribution of solar energy is apparent in central China, which indicates that our model is also sensitive to depicting regional differences in solar distribution. Compared to the annual irradiation in Australia and China, the solar irradiation value in the whole of Japan is relatively low, and this suggests the solar resource in Japan is worse than that of Australia and China. It is found that our results are consistent with the published maps created by Solargis, when comparing the quantitative ranges and the distribution patterns of the solar irradiation maps.

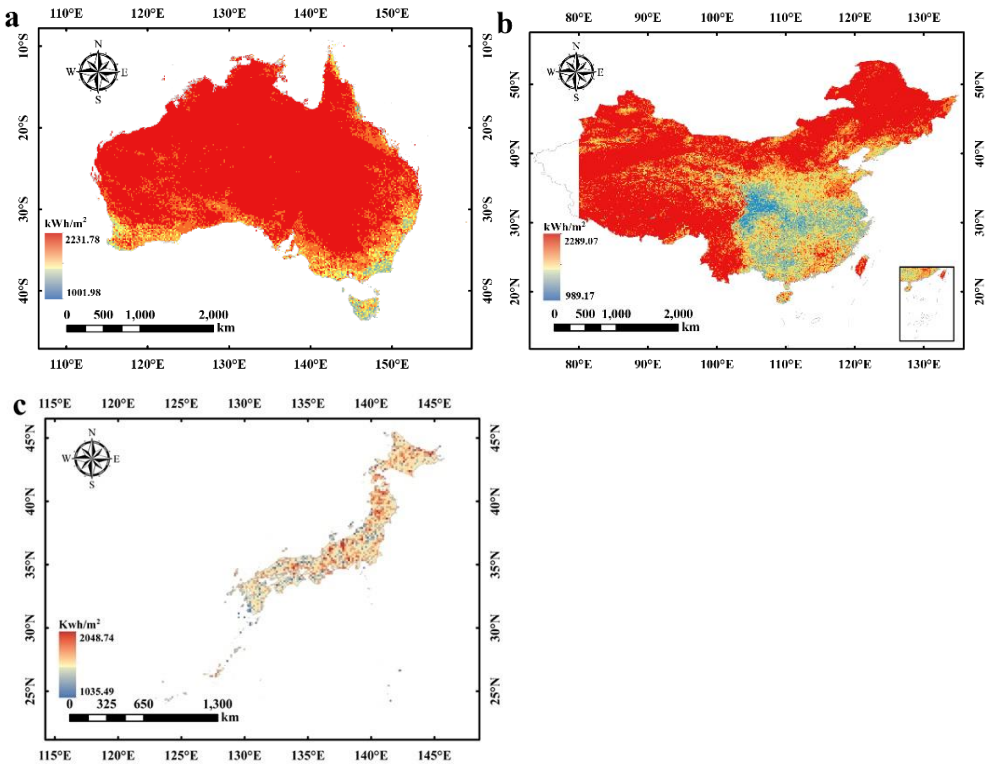


Figure 16 Annual horizontal surface global solar irradiation in three countries. (a) Distribution of annual irradiation in Australia. (b) Distribution of annual irradiation in China. (c) Distribution of annual irradiation in Japan

2.3.3.4 Analysis of annual land surface solar irradiation

The distribution of solar irradiation across different regions of China is primarily determined by geographical and climatic factors. The degrees of solar irradiation in the northwest and northeast of China are higher than those in the southeast due to unique geographical and climatic characteristics. The northwest region, including areas such as Xinjiang and the Tibetan Plateau, has high altitudes, low cloud cover, and low humidity, which result in lower atmospheric absorption and scattering of solar irradiation, allowing more solar energy to reach the surface. In comparison, the southeast region of China is affected by a humid subtropical monsoon climate characterized by high humidity, frequent rainfall, and extensive cloud cover, leading to significant absorption and scattering of solar irradiation and thus a reduction in the amount of solar energy reaching the ground. Furthermore, according to the 2020 China Climate Bulletin (China Meteorological News Press, 2021), regional annual precipitation was above average in Northeast China, the middle-lower reaches of the Yangtze River, North China, Southwest China, and Northwest China, but below average in South China. While the spatial distribution of annual precipitation appears to align with that of land surface solar irradiation,

this correlation is likely due to shared climatic drivers, such as cloud cover and atmospheric moisture content, rather than a direct causal relationship. Precipitation is one of several factors that influence solar irradiation availability, with cloud cover playing a more critical role in reducing solar energy reaching the Earth's surface.

Australia's abundant solar energy resources are largely due to its unique geographical location and climatic conditions. Situated in the mid-latitudes of the Southern Hemisphere, Australia has high solar angles throughout the year, which creates ideal conditions for strong solar radiation. This is particularly evident in the arid and semi-arid regions of central and western Australia, where extremely low humidity and minimal cloud cover allow maximum solar radiation to reach the surface. Furthermore, low levels of air pollution, sparse vegetation, and predominantly clear skies throughout much of the year enhance solar irradiance in Australia's interior, making it one of the world's most promising regions for solar energy exploitation.

Despite being located at the same latitude as Northeast China and Shandong, Japan has significantly lower solar energy resources due to its geographical location and maritime climate. As an island nation, Japan is heavily influenced by a maritime climate, characterized by high cloud cover and humidity throughout the year, particularly during the rainy and typhoon seasons, which limits the intensity of solar radiation reaching the ground. Additionally, Japan's mountainous terrain contributes to cloud formation, further reducing solar exposure compared to the relatively flat terrain of Northeast China and Shandong. Consequently, even at the same latitude, Japan's solar energy potential is markedly lower than that of China.

2.4 Conclusion

This study developed a method by integrating machine learning models and remote sensing technologies to estimate land surface solar irradiation at fine temporal resolutions (i.e., hourly to daily) over large geographical areas. Even though the study areas of Australia, China, and Japan are three big countries that contain a variety of climate zones, the trained models based on only a few stations still achieved high prediction accuracy with R^2>0.7 for all the stations. By comparing the generated maps with the published maps in terms of the spatio-temporal

distributions and the quantitative ranges, it is found that our results are broadly in line with the published maps. This suggests that the established models are accurate and reliable, and the proposed method can be used to estimate land surface solar irradiation in large-scale regions. In addition, the high availability of Himawari-8 satellite products with free licensed characteristics makes it possible to be widely used for an accurate estimation of solar irradiation over large regions, which is especially important for nations that aim to promote using solar energy.

This study used 27 datasets to train the machine learning models independently, which thus created a well-trained model for each of the 27 solar observation stations. As all the trained models obtained high estimation accuracy, all the models were used to create solar irradiation maps to make full use of the currently available datasets. However, as the solar observation stations have sparse distribution in each country, it is difficult to validate the prediction accuracy of each pixel value in the finally created solar irradiation maps. Alternatively, the observed solar irradiation data with determined geo-locations can be used as real samples to systematically investigate the final prediction accuracy.

The Kriging interpolation method was used to generate the spatially continuous meteorological images, which were used as the input parameters for estimating solar irradiation. Although the analysis shows that the overall interpolation accuracy is significantly high, it is hard to make sure that the whole areas maintain the same high accuracy. Nevertheless, the comparison of the published maps and the relative error matrices helps confirm that this method is feasible and the results are reliable. Meanwhile, this study conducted the importance analysis for the input parameters and it was found that the impacts of these parameters on solar estimation are different between the two countries. While in the same country, the impacts of the parameters are consistent for different models. This implies the effectiveness of the selected parameters for the solar estimation. It is worth mentioning that meteorological conditions can affect land surface solar irradiation to some extent, in which the humidity makes a great contribution.

The average values of a set of the estimated solar irradiation maps in the same spatial and

temporal domains are used to create the final solar irradiation map because of two reasons. First, the estimation accuracies (R^2) of all the models are basically consistent in a small range between 0.7 and 0.9. Second, the relative error matrices (Figure 2.3.3) between the estimated values and measured values are between 0.1 and 0.2 only. This demonstrates that the difference between each estimation solar irradiation map is rather small. Therefore, the estimated solar irradiation maps can make an equal contribution to creating the final solar irradiation map.

Chapter 3 A Dual-gate Temporal Fusion Transformer for estimating large-scale land surface solar irradiation

While the traditional machine learning methods used in Chapter 2 have significantly improved the rapid and accurate estimation of solar irradiation, they face challenges in handling geographical heterogeneity and providing interpretable results. To address these challenges, this chapter proposes the Dual-gate Temporal Fusion Transformer (DGTFT), a novel interpretable deep learning network, to improve hourly land surface solar irradiation estimation. Integrating the Temporal Fusion Transformer (TFT) with the Dual-gate Gated Residual Network (DGRN) and Dual-gate Multi-head Cross Attention (DGMCA), the optimal network achieved R^2=0.93, MAE=0.022 (kWh/m^2), and nRMSE=0.048 through ablation experiments. Applied to datasets from Australia, China, and Japan, the proposed network outperformed traditional machine learning methods with a minimum R^2 increase of 23.88%, MAE decrease of 43.18%, and nRMSE decrease of 62.79%. Accurately estimating land surface solar irradiation, providing interpretable results, and generating continuous irradiation maps for large-scale areas, the proposed network aids in quantifying solar potential and offers scientific guidance for the photovoltaic industry's development.

3.1 Methodology

3.1.1 Research framework

Figure 17 describes the research framework of this study. Firstly, we cleaned the collected multi-source data. Then, the geographical spatio-temporal dataset was constructed in GIS. Next, a novel interpretable deep learning network with improved structures was proposed to improve the estimation capability of spatio-temporal land surface solar irradiation, and the optimal network was determined based on a series of ablation experiments. After that, to evaluate the capabilities of transfer learning and the effectiveness of the proposed networks, the optimal network was trained using the hourly dataset in Australia, and the well-trained network was applied to the hourly dataset in Japan and the daily dataset in China. Additionally, the interpretability of the models applied in three countries was offered. Finally, the annual

continuous land surface solar irradiation maps in three countries were generated using the proposed network.

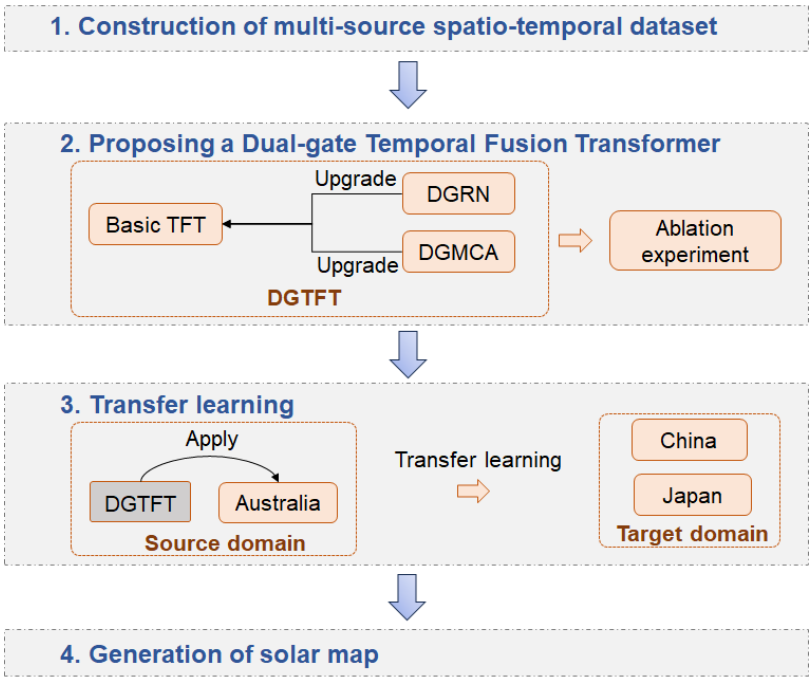


Figure 17 The research framework

3.1.2 Construction of spatio-temporal dataset

3.1.2.1 Spatio-temporal data

The data used in this Chapter is the same as that in Chapter 2. These data can be divided into two categories, spatial data and temporal data. As shown in Figure 1, the temporal data consists of MIs, solar irradiation from stations, CSI, COT, and AOT, and the spatial data consists of geographical coordinates, station names, and climate categories. This study considers each station as a point geographic object. The spatio-temporal attributions to the geographic object were assigned. The process of the GIS representation is shown in Figure 18. Specifically, the temporal data and the spatial attribution information (i.e., climate category and station name) were assigned to the corresponding geographic spatial locations.

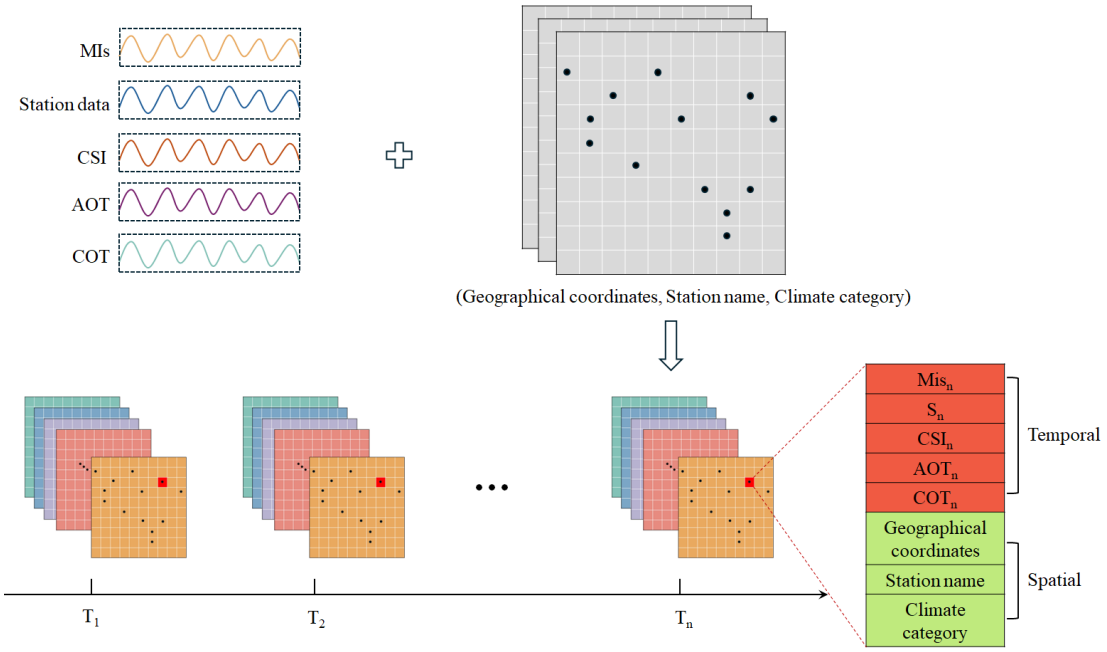


Figure 18 The process of the GIS representation for constructing the spatio-temporal dataset

3.1.2.2 GeoAI dataset

Firstly, the MissForest method (Arriagada et al., 2021) was employed to fill in the gaps in the dataset, which is a machine learning-based method for the simulation of the missing data. The missing values accounted for 0.02%, 0.001%, and 0.01% of the datasets in Australia, China, and Japan, respectively. Since the TFT model offers different network layers for training static and time-varying inputs, extracting and integrating various feature information all data types, spatial and temporal input variables were labeled in the dataset, and the measured hourly land surface solar irradiation observed from the stations was labeled as the training target. The static variables include the geographic coordinates of the 13 meteorological stations, the associated climate categories, and the station Name ID which is used for grouping the dataset to identify each station. The time-varying variables include COT, AOT, CSI, and MIs. To facilitate model training and evaluation, the entire dataset was divided into three sub-datasets: a training dataset, a validation dataset, and a test dataset, constituting 80%, 10%, and 10% of the data, respectively.

3.1.3 Temporal Fusion Transformer

The Temporal Fusion Transformer (TFT) model (Lim et al, 2021) is a novel attention-based deep learning architecture specifically designed for handling multi-dimensional time series

data. Solar irradiation is a classic time series. Let I represents unique entities in land surface solar irradiation. Each entity i consists of static metadata s_i, time series inputs $X_{i,t}$, and solar targets $y_{i,t}$ at time step t, $t \in [0,\ T_i]$. Time series inputs $X_{i,t}$ can be classified into two categories, $X_{i,t} = [z_{i,t}^T,\ x_{i,t}^T]^T$. Past inputs $z_{i,t}$ denote that theses variables can only be measured at each step and are unknown beforehand, and know future inputs $x_{i,t}$ represent they can be predetermined and the value of these variables are known before time step t. The prediction function is defined as follows:

$$\hat{y}_i(t,\tau) = f(\tau, y_{i,t-k:t}, z_{i,t-k:t,} x_{i,t-k:t+\tau}, s_i) \tag{3.2.1}$$

Where $y_{i,t-k:t} = \{y_{i,t-k}, \dots, y_{i,t}\}$ denotes targets until the time t, $z_{i,t-k:t} = \{z_{i,t-k}, \dots, z_{i,t}\}$ denotes past inputs, $x_{i,t-k:t+\tau} = \{x_{i,t-k}, \dots, x_{i,t+\tau}\}$ denotes known future inputs across the full range, and τ represents the prediction time point.

The TFT model is designed for high forecasting performance of long-term land surface solar irradiation by using effective components, as shown in Figure 19. The TFT model consists of five major constituents, namely, gating mechanisms, variable selection networks, static covariate encoders, temporal processing, and prediction intervals. Each constituent is detailed below.

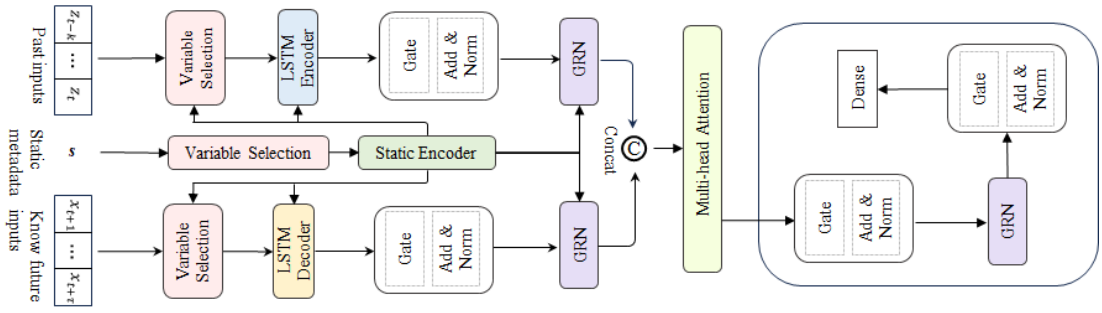


Figure 19 Temporal Fusion Transformer architecture

3.1.3.1 Gating mechanisms

Gating mechanisms can filter out unnecessary components of the architecture and can be

flexibly applied to non-linear processing only where needed. To achieve this aim, Gated Residual Network (GRN) is used as a building block of TFT. The architecture of the GRN is demonstrated in Figure 20. The GRN consists of two inputs, namely, a primary a and an optional context vector c. The GRN is described as follows:

$$GRN_{\omega}(a,c) = LayerNorm(a + GLU_{\omega}(\eta_1)) \tag{3.2.1.1}$$

$$\eta_1 = W_{1,\varpi}\eta_2 + b_{1,\varpi} \tag{3.2.1.2}$$

$$\eta_2 = ELU(W_{2,\omega}a + W_{3,\omega}c + b_{2,\omega}) \tag{3.2.1.3}$$

$$GLU_{\omega}(\gamma) = \sigma(W_{4,\varpi}\gamma + b_{4,\omega})\odot(W_{5,\omega}\gamma + b_{5,\omega}) \tag{3.2.1.4}$$

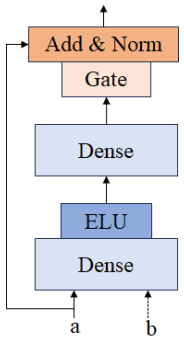


Figure 20 GRN architecture

Where LayerNorm is standard layer normalization (Ba et al., 2016) η_1 and η_2 are intermediate layers, ϖ is an index to represent weight sharing, $W_{(.)}$ and $b_{(.)}$ denote the weights and biases, ELU is the Exponential Linear Unit activation function (Clevert et al., 2015), $\odot$ is the element-wise Hadamard product, and $\sigma_{(.)}$ is the sigmoid activation function. GLU enables TFT to regulate the degree to which the GRN contributes to the original input, potentially bypassing the entire layer if required. This can occur when the GLU outputs are predominantly close to 0, effectively suppressing the nonlinear contribution.

3.1.3.2 Variable selection network

The variable selection network not only assesses the importance of each variable for selecting relevant input variables but also enables TFT to eliminate variables that demonstrate a detrimental effect on the prediction performance.

The static and time-series continuous variables are transformed into feature representations and dimensional vectors, respectively. Let $\xi_t^{(j)}$ represents the transformed jth variable at time t. At each time step, each $\xi_t^{(j)}$goes through its own GRN, as show in Eq. (3.2.1.2). $\tilde{\xi}_t^{(j)}$ is the corresponding processed feature vector for the jth variable. All past inputs are transformed into flattened vectors $[I]_t = [\xi_t^{(1)^T}, \dots, \xi_t^{(m_\chi)^T}]^T$. The variable selection weights $v_{\chi t}$ are calculated using Eq. (3.2.1.2.2), where c_s denote an external context vector. Finally, the processed features are weighted by their variable selection weights and combined as shown in Eq. (3.2.1.2.3).

$$\tilde{\xi}_t^{(j)} = GRN_{\tilde{\xi}(j)}(\xi_t^{(j)}) \tag{3.2.1.2.1}$$

$$v_{\chi t} = Softmax(GRN_{v_\chi}([I]_t, c_s)) \tag{3.2.1.2.2}$$

$$\tilde{\xi}_t = \sum_{j=1}^{m_\chi} v_{\chi t}^{(j)} \tilde{\xi}_t^{(j)} \tag{3.2.1.2.3}$$

3.1.3.3 Static covariate encoder

In contrast to other time-series deep learning methods, the static covariate encoder is designed to integrate static features extracted from static metadata into the TFT network. Individual GRN encoders are utilized to generate four different context vectors, namely, c_s, c_e, c_c, and c_h. These four context vectors are integrated with temporal features in the TFT network.

3.1.3.4 Interpretable multi-head attention module

TFT employs a self-attention mechanism that modifies from the multi-head attention mechanism in the standard transformer architectures proposed by Vaswani et al. (2017). This

modification enables TFT to capture long-term relationships across various time steps and enhances explainability. The Attention mechanism is described as follows:

$$Attention(Q, K, V) = A(Q, K)V \quad (3.2.1.4.1)$$

$$A(Q, K) = Softmax(QK^T/\sqrt{d_{attn}}) \quad (3.2.1.4.2)$$

Where Q is the "query", K is the "key", V is the "value", and $A(.)$ denotes a normalization function.

In general, the multi-head attention mechanism is defined in Eq. (3.2.1.4.3), (3.2.1.4.4).

$$MultiHead(Q, K, V) = [H_1, \dots, H_{mH}]W_H \quad (3.2.1.4.3)$$

$$H_h = Attention(QW_Q^{(h)}, KW_K^{(h)}, V\,W_V^{(h)}) \quad (3.2.1.4.4)$$

Where $W_Q^{(h)}$, $W_K^{(h)}$, $W_V^{(h)}$ denote head-specific weights for queries, keys, and values, respectively. And W_H is the combination of outputs concatenated from all heads H_h.

The values learned in each head using multi-head attention are different, so attention weights would not represent the importance of specific features which enables the model to decrease explainability. Given this reason, the TFT model modifies multi-head attention to share values in each head and uses additive aggregation of all heads. The Interpretable multi-head attention is defined in Eq. (3.2.1.4.5), (3.2.1.4.6).

$$InterpretableMultiHead(Q, K, V) = \widetilde{H}W_H \quad (3.2.1.4.5)$$

$$\widetilde{H} = \tilde{A}(Q, K)VW_V) = 1/H \sum_{h=1}^{m_H} Attention\,(QW_Q^{(h)}, KW_K^{(h)}, V\,W_V^{(h)}) \quad (3.2.1.4.6)$$

3.1.3.5 Temporal fusion decoder

The four layers are employed in the temporal fusion decoder to learn temporal relationships in the dataset: i) Locality enhancement with sequence-to-sequence layer uses a sequence-to-sequence layer to capture local dependence; ii) Static enrichment layer employs GRN network to enhance temporal features with static data; iii) Temporal self-attention layer introduces interpretable multi-head attention to pick up long term dependencies and enhance explainability; iv) Position-wise feed-forward layer is used to process the outputs of the self-attention layer.

3.1.4 Dual-gate Temporal Fusion Transformer

In this study, a novel framework for estimating land surface solar irradiation is proposed named Dual-gate Temporal Fusion Transformer (DGTFT), which advances the backbone using the TFT (Lim et al, 2021) module. To greatly forecast time-series solar data, the author proposes: i) a novel Dual-gate Gated Residual Network (DGRN) that modifies from the GRN of the original TFT (Lim et al, 2021) for more accurate estimation performance. ii) a novel Dual-gate Multi-head Cross Attention (DGMCA) that integrates the interpretable Multi-head Attention that inherits the TFT (Lim et al, 2021) with Cross Attention (CA) (Chen et al., 2021) for effectively learning the spatio-temporal features from the dataset and greatly integration the static spatial features with the temporal features.

3.1.4.1 Model Overview

As shown in Figure 21, the proposed DGTFT is composed of a multi-data encoder and a temporal fusion decoder. There are three modules in the multi-data encoder, namely, a static encoder, a past-observed encoder, and a future-known decoder. The input data is classified into three categories (i.e., static metadata, past inputs, and known-future inputs) for feeding into the corresponding layers, and this aims to greatly distinct and extract useful static and time-varying features. In the static encoder, the static metadata is first embedded and fed into the variable selection, and then the output is transformed into four static context vectors for integrating with time-varying features. In the past-observed encoder and future-known decoder, the data processing is the same. Specifically, the inputs are also embedded and fed into the variable

selection, and then the LSTM module is employed for learning temporal features. The variable selection module and LSTM model inherit from the TFT (Lim et al, 2021), which are mentioned in Chapter 3.1.3.

After the multi-data encoder, the outputs are fed into the temporal fusion decoder. The temporal fusion decoder is composed of a DGRN, a DGMCA, and a position-wise feed-forward layer. The static context vectors are integrated with the outputs of the past-observed encoder and future-known decoder using the DGRN for the static enrichment, respectively, and then the two outputs of the DGRN are concatenated to be fed into the DGMCA for picking up long-range dependencies. Finally, non-linear processing in the position-wise feed-forward layer is applied to the outputs of the DGMCA.

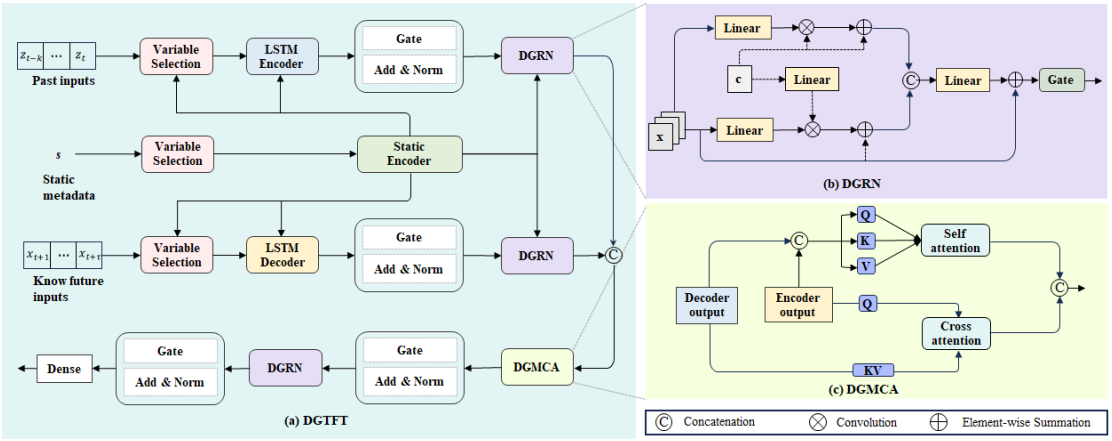


Figure 21 Dual-gate Temporal Fusion Transformer architecture

3.1.4.2 Dual-gate Gated Residual Network

GRN plays a crucial role in TFT (Lim et al, 2021) to flexibly provide non-linear processing, which is applied in data encoding, variable selection, and enhancing the temporal features with static data. Although the aim of the simple design of GRN is to enable the model flexible to give precise insights into the non-linear relationship between inputs and targets, the excessively simple structure of this design may not accurately describe the non-linear relationship. Therefore, the author proposes a novel Dual-gate Gate Residual Network (DGRN) to improve the non-linear processing ability of GRN.

In this section, we detailed the proposed DGRN, which is composed of two branches of non-

linear processing. Since the DGRN is applied in different modules for processing the single input X and dual-branch inputs (i.e., X and static context c_s), the DGRN contains two modules to greatly process the inputs, namely, a single-input module and a dual-input module. In the single-input module, to greatly construct the non-linear relationship, X is fed into two branches in parallel and each branch contains one Linear layer and a Tanh active function. Then, the outputs of the two branches are concatenated to be fed into one Linear layer and the Tanh active function. To avoid the degradation of the model, the residual connection is conducted, and the output is fed into the gate layer. In the dual-input module, the inputs contain X and static context c_s. X is also fed into two branches for processing one Linear layer and a Tanh active function. c_s is also fed into two branches for integrating with the features of X. After that, the outputs of both branches are fed into the layers that are the same as those in the single-input module.

3.1.4.3 Dual-gate Multi-head Cross Attention

In this section, the author details the proposed DGMCA, which is composed of a self attention and a cross attention. The output of a past-observed encoder and the output of a future-known decoder are fed into the DGMCA to learn long-term temporal dependency. To greatly learn the information of past time and the estimation information, we design a dual-gate structure using a self attention and a cross attention. Since a self attention module and a cross attention module are in parallel, the output of a past-observed encoder and the output of a future-known decoder are fed into two modules. Specifically, in the self attention module, the output of a past-observed encoder is first concatenated with the output of a future-known decoder, and then the concatenated output C_{ts} are transformed into the query, the key, and the value for performing the self attention. In the cross attention module, only the output of a future-known decoder serves as the query, and the output of a past-observed encoder are transformed as the key and the value. After this dual-gate attention structure, the output of the self attention module is concatenated with the output of the cross attention module. We detail a self attention and a cross attention next.

This module employs the Interpretable multi-head attention that is mentioned in Chapter

3.2.3.4. The self attention is performed using the concatenated output C_{ts} of a past-observed encoder and a future-known decoder. To enhance the forecasting performance, the query is transformed from intercepted C_{ts} related to the known-future time-series data, and the key and value are transformed from C_{ts}.

Cross attention (CA) is performed between the output of a past-observed encoder E_p and the output of a future-known decoder D_f. Mathematically, CA can be expressed as

$$q = D_f W_q,\ k = E_p W_k,\ v = E_p W_v \quad (3.2.2.3.1)$$

$$A = softmax(qk^T/\sqrt{C/h}) \quad (3.2.2.3.2)$$

$$CA = Av \quad (3.2.2.3.1)$$

Where W_q, W_k, W_v are learnable parameters, C and h are the embedding dimension and number of heads, and A denotes the attention map. It is noticed that the computation and memory complexity of generating A in cross attention are linear rather than quadratic as in all-attention because we only employ D_f in the query, and it leads to enhance efficiency of the entire process (Chen et al., 2021). Furthermore, as in self attention, multi-head mechanism is also used in CA.

3.1.5 Implementation details

The implementation of the TFT model involves the use of Python 3.8 along with TensorFlow 2.12.0, PyTorch-forecasting 0.10.3, and PyTorch-lightning 1.8.6. We employed the Python library "TimeSeriesDataset" to split the data. Early stopping was utilized to prevent overfitting. The computations were executed on a high-performance computer equipped with an intel (R) Core (TM) i7-6800K CPU, operating at 3.40 GHz, 6.0 TB RAM, and running on the Ubuntu 16.04 LTS system.

3.1.6 Evaluation metrics

To evaluate the estimation performance of the proposed network, the most frequently used evaluation indicators, i.e., the coefficient of the determination (R^2), the mean absolute error (MAE), and normalized Root Mean Square Error (nRMSE) were adopted, given as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (3.2.6.1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (3.2.6.2)$$

$$nRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \quad (3.2.6.3)$$

Where $\hat{y}_i$ and y_i are estimated and measured land surface solar irradiation values, respectively. $\bar{y}_i$ is the average value of measured land surface solar irradiation.

3.1.7 Generation annual land surface solar irradiation maps

After training the models using datasets in Australia, China, and Japan, these well-trained models were employed to generate annual land surface solar maps at a 5-km spatial resolution in three countries in 2020. The meteorological, COT, AOT, and CSI images are well prepared and used as the input parameters of the trained model. In addition, a set of meteorological images are obtained by using the Kriging interpolation method.

3.2 Results and discussion

3.2.1 Ablation study

To verify the effectiveness of each component in the proposed DGTFT, we conduct ablation studies on the dataset in Australia. We use the TFT as the backbone, and we substitute a novel DGRN and DGMCA for the original GRN and interpretable multi-head attention, respectively. The components being evaluated contain DGRN and DGMCA. We also further conduct ablation studies on the Linear layers in DGRN with different action functions (i.e., null, Tanh,

Sigmoid, and Softmax) to explore the optimal combination of the Linear layer and the active function. Three indicators are employed to evaluate the performance of different combinations, namely, R^2, MAE, and nRMSE. The results are shown in Table 2.

Overall, the combination of "Baseline+DGRN+ DGMCA" shows the best prediction performance based on three indicators, with R^2=0.9260, MAE=0.02198 (kWh/m^2), and nRMSE=0.04845, following by the "Baseline+DGMCA". Although the nRMSE value of this combination is slightly higher than that of the "Baseline+DGMCA", it outperforms other combination based on the values of R^2 and MAE. Therefore, "Baseline+DGRN+ DGMCA" shows the best performance for predicting the land surface solar irradiation based on the comprehensive evaluation of these three indicators. Furthermore, it outperforms the benchmark "Baseline" by 2%, 13%, and 7% for R^2, MAE, and nRMSE. These results suggest that the "Baseline+DGRN+ DGMCA" effectively improves the prediction capability for land surface solar irradiation.

3.2.1.1 Effect of DGMCA

Compared to the benchmark, the "Baseline+DGMCA" increases by 2% for R^2 and decreases by 12% and 8% for MAE and nRMSE. This suggests that the designed DGMCA module is able to learn better long-term temporal dependence and spatial features than the original TFT model.

3.2.1.2 Effect of DGRN

Compared to the benchmark, the "Baseline+DGRN" increases by 1% for R^2 and decreases by 5% and 4% for MAE and nRMSE, which indicates that the proposed DGRN module improves the prediction capability. Furthermore, we give insights into the effect of the combination of DGRN and DGMCA. The result of the "Baseline+DGRN+ DGMCA" is superior in R^2 and MAE than the "Baseline", "Baseline+DGRN", and "Baseline+DGMCA". Additionally, we also investigate the impact of the commonly used active functions on the prediction performance, including Tanh, Sigmoid, and Softmax. From the results, although the performance of these three combinations is better than that of the benchmark, their performance is worse than that of the "Baseline+DGRN+ DGMCA". This suggests that these active

functions are not suitable for adding the DGRN module.

Table 2 The performance of different components of our model on the test dataset of the Australia dataset

Architecture	R^2	MAE (kWh/m^2)	nRMSE
Baseline	0.9091	0.02535	0.05253
Baseline+DGRN	0.9150	0.02411	0.05054
Baseline+DGMCA	0.9257	0.02226	**0.04828**
Baseline+DGRN+ DGMCA	**0.9260**	**0.02198**	0.04845
Baseline+DGRN+ DGMCA+Tanh	0.9186	0.02318	0.05053
Baseline+DGRN+ DGMCA+sigmoid	0.9166	0.02270	0.05073
Baseline+DGRN+ DGMCA+softmax	0.9197	0.02326	0.05015

3.2.2 Evaluation of the performance of DGTFT

3.2.2.1 The performance of transfer learning

To evaluate the capability of transfer learning of the proposed DGTFT, we employ three datasets in Australia, China, and Japan to calculate the estimation accuracy based on R^2, MAE, and nRMSE. The results are shown in Table 3. Overall, the performance of the proposed DGTFT is all superior to other traditional machine learning methods, which suggests that the DGTFT can provide highly accurate and reliable prediction performance and has the excellent capability of transfer learning. The capability of integrating static spatial data with temporal data of the DGTFT may lead to highly accurate estimation performance. Tradition machine learning methods is difficult to use the static information to enhance the model learning ability. Distinct from the methods which train the individual model for each station in Chapter 2, we just train the one model for each dataset. Therefore, we can notice that machine learning methods are limited in processing spatio-temporal data, while the DGTFT shows the good capability to investigate this non-linear relationship integrated static spatial data with temporal data.

Furthermore, among the three datasets, the estimation results of the dataset in Australia are better than those of other two datasets, which indicates that the DGTFT model is slightly more adaptable to the Australian dataset than the other two datasets. it is noticed that the estimation results using our model are far better than other traditional machine learning methods using the

dataset in China. It is possible that this is because the size of the dataset in China is smaller than the other datasets. Since the temporal resolution of the dataset in China is daily and other datasets are hourly, the size of the dataset in China is obviously smaller when the time span of the study is consistent. These results indicate that traditional machine learning methods cannot work well in the small-size dataset, while the DGTFT is insensitive to the size of the dataset indicating that it has good robustness.

Table 3 The estimation performance of datasets in Australia, China, and Japan using the DGTFT

Model	Dataset in Australia			Dataset in China			Dataset in Japan		
	R^2	MAE (kWh/m^2)	nRMSE	R^2	MAE (kWh/m^2)	nRMSE	R^2	MAE(kWh/m^2)	nRMSE
RF	0.74	0.12	0.39	0.45	0.88	0.75	0.67	0.15	0.43
GBM	0.69	0.14	0.43	0.64	0.97	0.33	0.62	0.177	0.47
AdaBoost	0.57	0.18	0.49	0.28	1.64	0.76	0.47	0.22	0.55
MLP	0.69	0.14	0.43	0.19	1.15	0.92	0.63	0.16	0.46
Our method	**0.93**	**0.022**	**0.048**	**0.88**	**0.50**	**0.12**	**0.83**	**0.037**	**0.16**

3.2.2.2 Generation annual land surface solar irradiation maps

Figure 22- Figure 24 describe the distribution of annual land surface solar irradiation in Australia, China, and Japan, respectively. Overall, the land surface solar irradiance levels across Australia predominantly reside within higher ranges, contrasting with Japan where they mostly fall within lower ranges, with China positioned intermediate to the two. This underscores Australia's abundant solar energy resources. Specifically, in Australia, land surface solar irradiance levels are generally high, except for a small portion near the southern coastal areas where values are relatively lower. Across China, land surface solar irradiance diminishes gradually from the northwest to the southeast. Conversely, in Japan, solar irradiance levels predominantly register within lower ranges, with sporadic higher values scattered across its northern and central regions.

To evaluate the estimation accuracy of the generated maps, we calculated the annual cumulative absolute errors between estimated values and measured values in 27 stations in three countries. Figure 25 shows the result. Overall, the annual cumulative absolute error values across these 27 stations are relatively small, with 92.86% of stations exhibiting annual cumulative error values below 400 (kWh/m^2). Specifically, the annual cumulative error values at Australian sites are slightly lower compared to those in China and Japan. These findings suggest the high precision of our trained model in generating large-scale continuous solar irradiance distribution maps, thus affirming the strong generalization capability and broad applicability of the proposed neural network model.

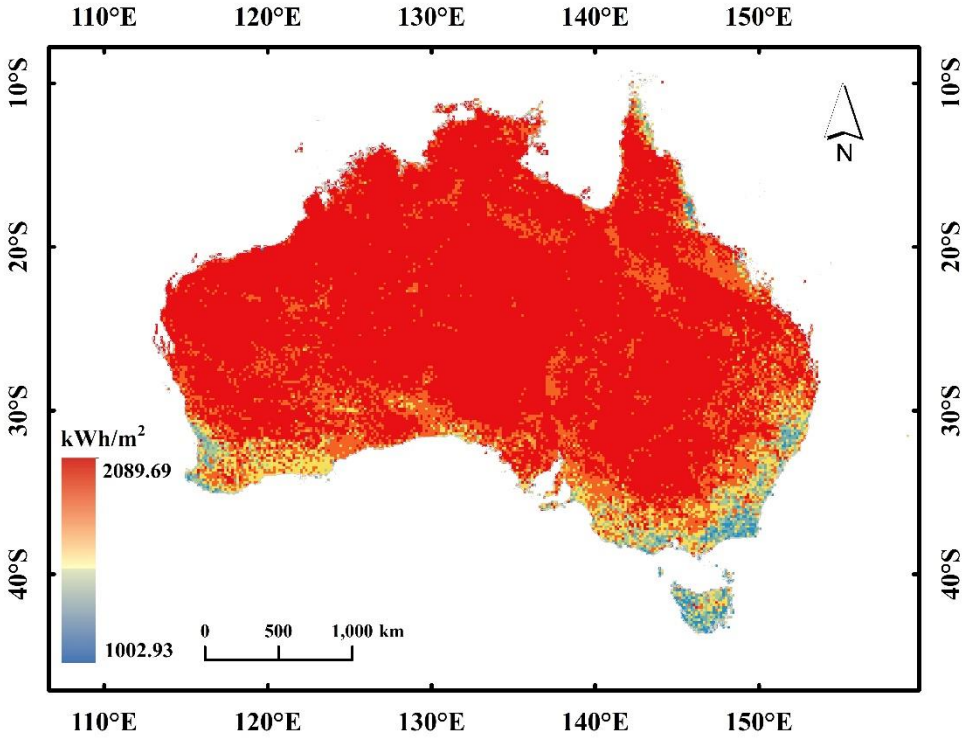


Figure 22 Distribution of annual land surface solar irradiation in Australia

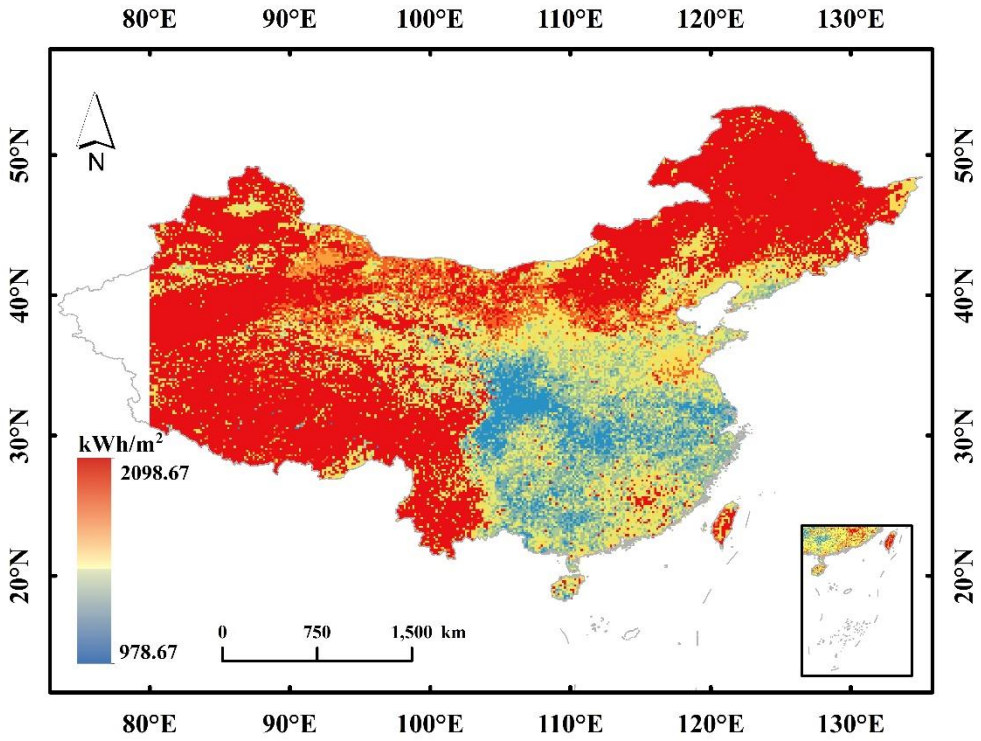


Figure 23 Distribution of annual land surface solar irradiation in China

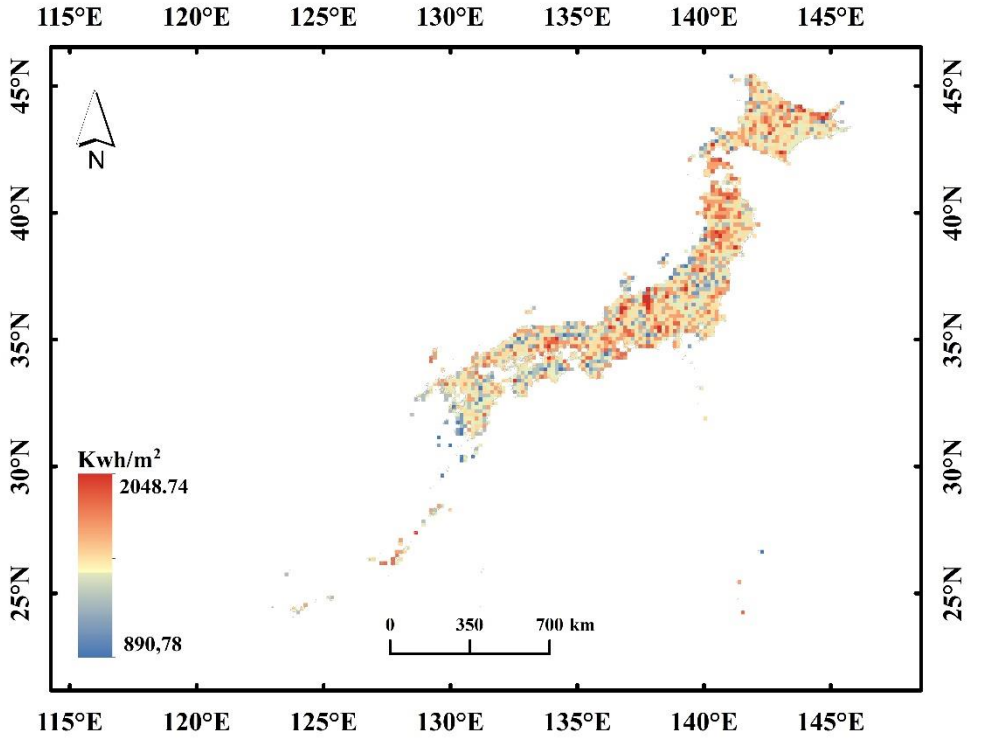


Figure 24 Distribution of annual land surface solar irradiation in Japan

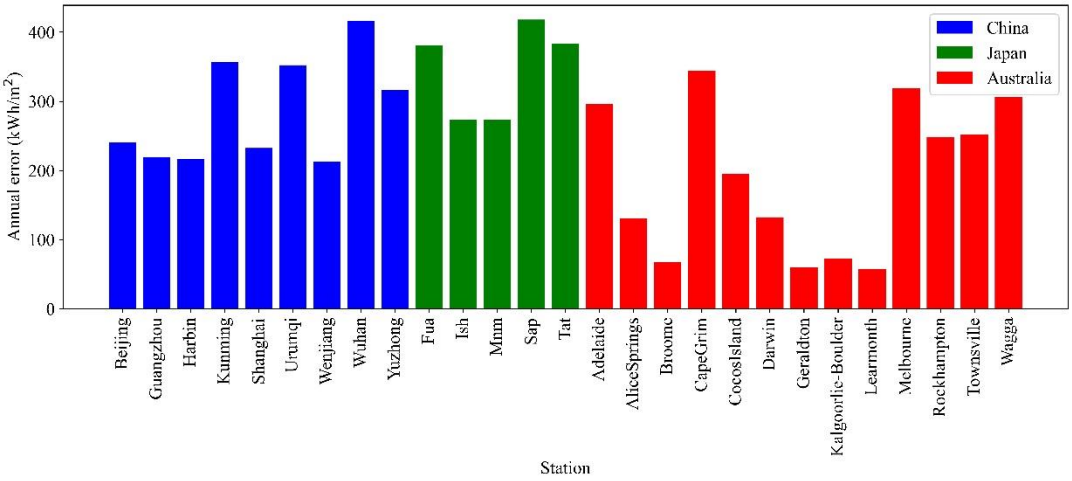


Figure 25 Annual absolute errors between estimated values and GroundTruth values of 27 stations in Australia, China, and Japan

3.2.3 Interpretability of DGTFT

The DGTFT enables its network structure interpretive by quantifying the importance of variables in the different layers, including past-observed encoder, future-known decoder, and static encoder. Figures 26- Figures 28 show the importance of variables in the past-observed encoder, future-known decoder, and static encoder of the models trained by datasets in Australia, China, and Japan.

In the Decoder network layer, for models trained on the Australian and Japanese datasets, CSI (Channel State Information) emerged as the variable contributing the most to network training, with importance indices of approximately 40% and 50%, respectively. Conversely, for the model trained on the Chinese dataset, the variables of highest importance were the maximum temperature and humidity, with importance indices exceeding 20%.

In the Encoder network layer, for models trained on the Australian and Japanese datasets, solar irradiation emerged as the variable contributing the most to network training, with importance indices of approximately 85% and 40%, respectively. However, for the model trained on the Chinese dataset, CSI was the most important variable, with an importance index of approximately 24%.

In the Static network layer, for the model trained on the Australian dataset, the most important variable was $solar_{scale}$, with an importance index of approximately 33%; for the model trained on the Chinese dataset, the most important variable was $Station_{ID}$, with an importance index of approximately 86%; and for the model trained on the Japanese dataset, the most important variables were Longitude and $solar_{center}$, with importance indices of approximately 23%. These findings elucidate the varying contributions of different variables across different network layers during model training, thereby enhancing the interpretability of deep learning networks.

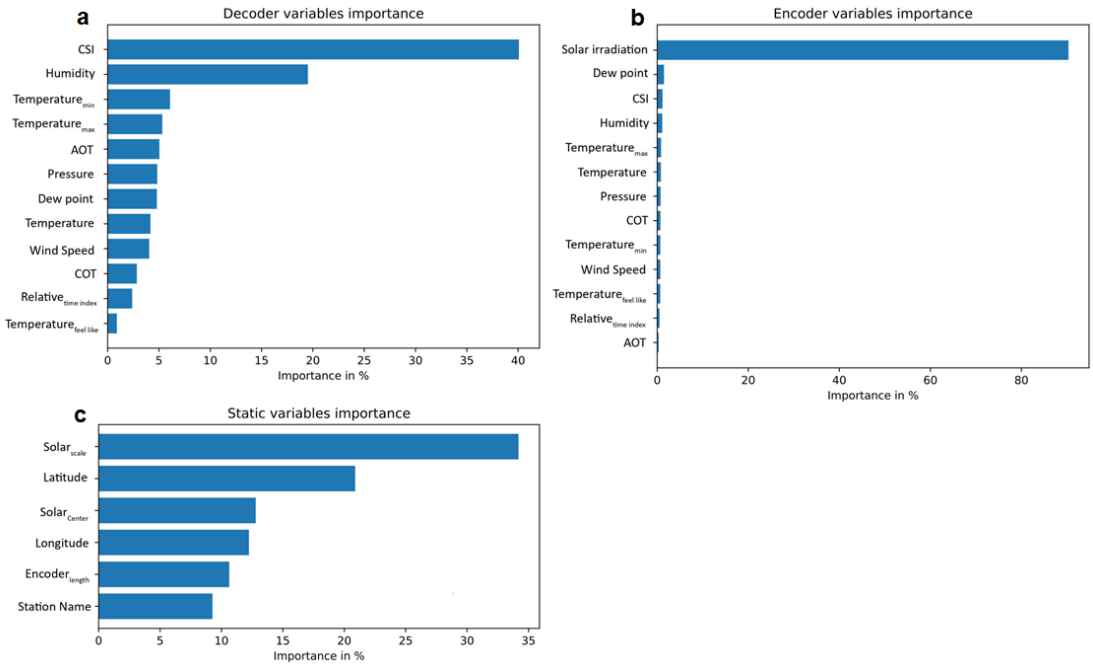


Figure 26 The importance of variables in the past-observed encoder, future-known decoder, and static encoder in Australia.

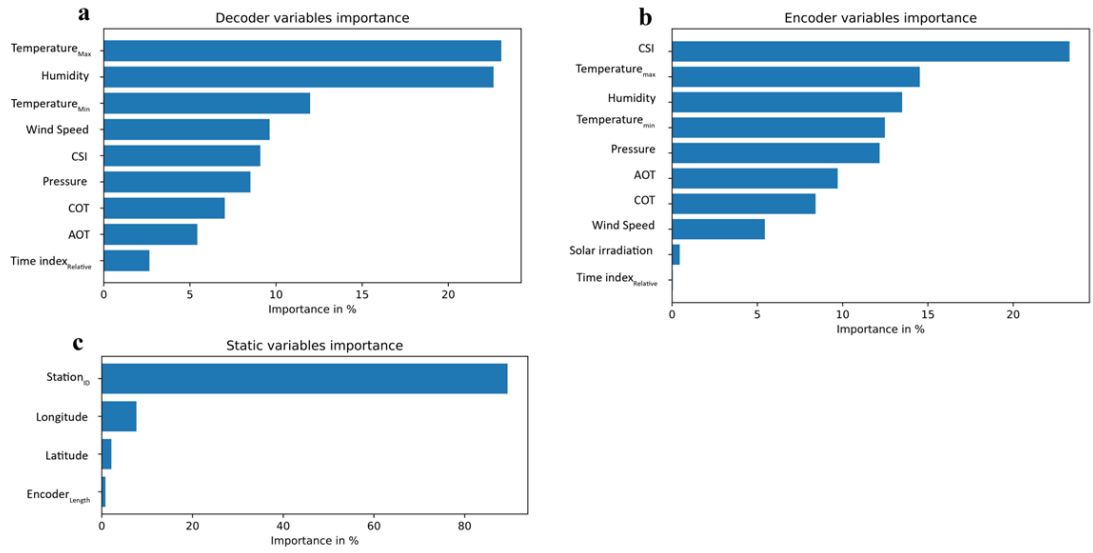


Figure 27 The importance of variables in the past-observed encoder, future-known decoder, and static encoder in China.

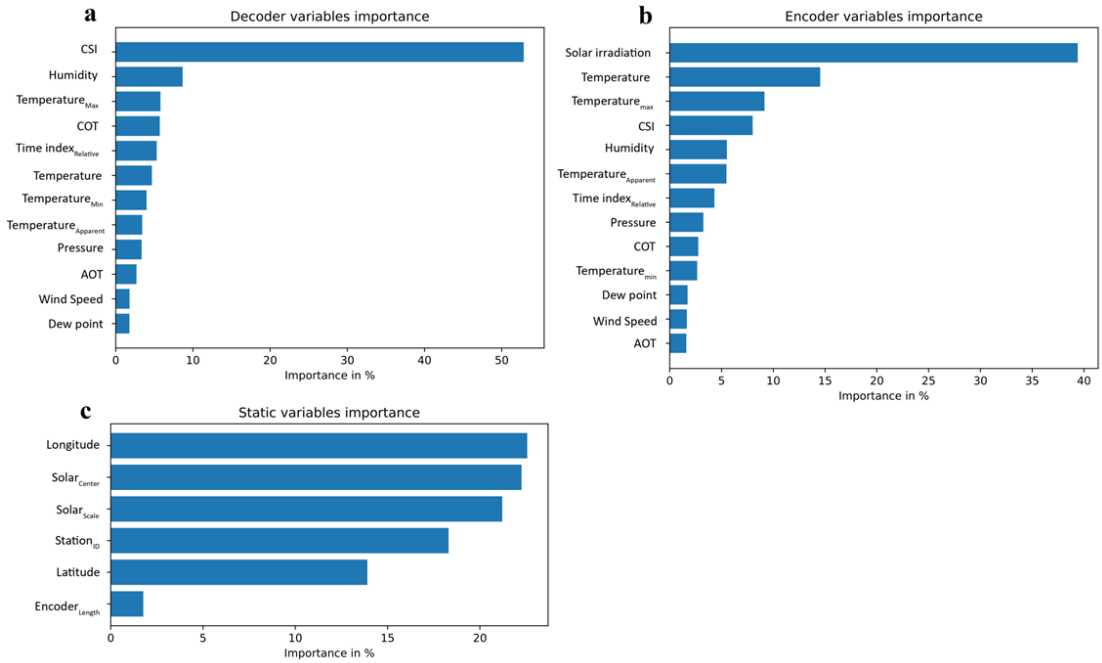


Figure 28 The importance of variables in the past-observed encoder, future-known decoder, and static encoder in Japan.

3.3 Conclusion

This Chapter proposes the state-of-the-art deep learning model DGTFT to explore the non-linear relationship between multi-source variables and land surface solar irradiation and provides the interpretive and high-accuracy method for estimating hourly/daily land surface solar irradiation in Australia, China, and Japan. Compared to other traditional machine learning models, the DGTFT model is the optimal method for estimating long-term time series hourly land surface solar irradiation, which has super high accuracy with R^2=0.92, R^2=0.82, R^2=0.83 using datasets in Australia, China, and Japan.

The DGTFT model shows the strong capability to use spatial and temporal characteristics from multi-source data. Compared to other commonly used methods, the DGTFT model not only can extract information from the static spatial data and integrate it with the time-varying data, which can significantly enhance the efficiency of the dataset. Solar irradiation has a strong spatial-temporal distribution, and this model can greatly consider the impact of static

geographic information on the estimation of the land surface solar irradiation for the issue of geographical heterogeneity.

Also, the DGTFT model provides a relatively transparent interpretable network. Specifically, the selection of suitable input variables is conducted in all key layers, namely the static layer, the Encoder layer, and the Decoder layer. The mechanism for this selection is based on the magnitude of their importance. The mechanism of self-selecting variables during network training can significantly reduce the complexity of the network architecture, thereby improving computational efficiency.

The DGTFT model used in this study can provide high-accuracy, interpretive, and reliable estimation maps for land surface solar irradiation, which can provide a reliable reference for the design of solar power generation systems.

Chapter 4 Fast and accurate estimation of solar irradiation on building rooftops in Hong Kong: A machine learning-based parameterization approach

A set of maps of monthly, seasonal, and annual land surface solar irradiation in Australia, China, and Japan were generated using the methods in Chapter 2 and Chapter 3, which display the distribution of the solar sources in these countries. To effectively harness the solar source, installing solar PV panels on building rooftops in regions with abundant solar potential is considered the most promising method. The aim of the estimation of the physical potential in Chapter 2 and Chapter 3 is to identify optimal regions with significant solar potential. After determining the optimal development regions, it is necessary to estimate more precise rooftop solar irradiation in the research region to provide a reliable and scientific guideline for developing distributed solar systems in the city. The findings from Chapters 2 and 3 indicate that Hong Kong has abundant solar resources, making it an ideal city for developing distributed rooftop solar systems. Therefore, Hong Kong is selected as the research region in this chapter for investigating rooftop solar potential.

In this Chapter, the author proposes a parametric-based method to estimate annual rooftop solar irradiation at a fine spatial resolution. Specifically, seven parameters (Digital Surface Model, Sky View Factor, shadow from buildings, shadow from terrain, building volume to façade ratio, slope, and aspect) are determined that have great importance in modeling rooftop solar irradiation. Three machine learning methods (RF, GBRT, AdaBoost) trained by the selected parameters are cross-compared based on R^2, MAE, and computation time. As a case study in Hong Kong, China, the RF outperformed GBRT and AdaBoost, with R^2=0.77 and MAE=22.83kWh/m^2/year. The time for training and prediction of rooftop solar irradiation is within 13 hours, achieving a 99.32% reduction in time compared to the physical-based hemispherical viewshed algorithm. These results suggest that the proposed method can provide an accurate and fast estimation of rooftop solar irradiation for large datasets. The results also indicate that the proposed method can provide a reliable and accurate reference for urban planners and the government to promote PV system installation on rooftops effectively and

reasonably design building rooftop structures.

4.1 Study area and data

This section includes the description of the data sources, the data structure design adaptive to machine learning models, the study area, and data pre-processing.

4.1.1 Study area

Hong Kong, China, is located at 22°15′ N, 114°15′ E, with a typical subtropical climate. It has a total land area of about 1, 110 km^2 with a hilly and mountainous topography, and around 75% of the land in Hong Kong is a mountainous area. High population density and limited land resources have formed the high-density urban morphology in downtown areas of Hong Kong associated with densely packed high-rise buildings (Leng et al., 2020). The territory is divided into 18 districts, with around 323,886 buildings by 2019. Because of the high building density and limited land resources, the rooftop solar PV system can be suitable and feasible for Hong Kong's renewable energy development in the future. Figure 29 (b-g) shows the change in annual clear sky surface solar irradiation in six locations from 2012 to 2021 in Hong Kong, and the data is obtained from NASA Power (NASA Power, 2021). To calculate the annual clear sky irradiation, we employed the monthly clear sky irradiation from the NASA Power dataset. NASA Power adopted the data from the Baseline Surface Radiation Network site observations. The monthly clear sky irradiation demonstrated a Bias of 0.03% and an RMSE of 5.7%. The resolution of the data is 1 degree latitude by 1 degree longitude. It is clear that the annual clear sky surface solar irradiation in different locations in Hong Kong is highly consistent, so this study does not use clear sky irradiation as the input variable. Figure 29 (h-i) shows the amount of annual solar irradiation observed by the King's Park Station and Kau Sai Chau Station (Hong Kong Observatory, 2023) from 2012 to 2021, which suggests that Hong Kong has a great potential for developing solar energy.

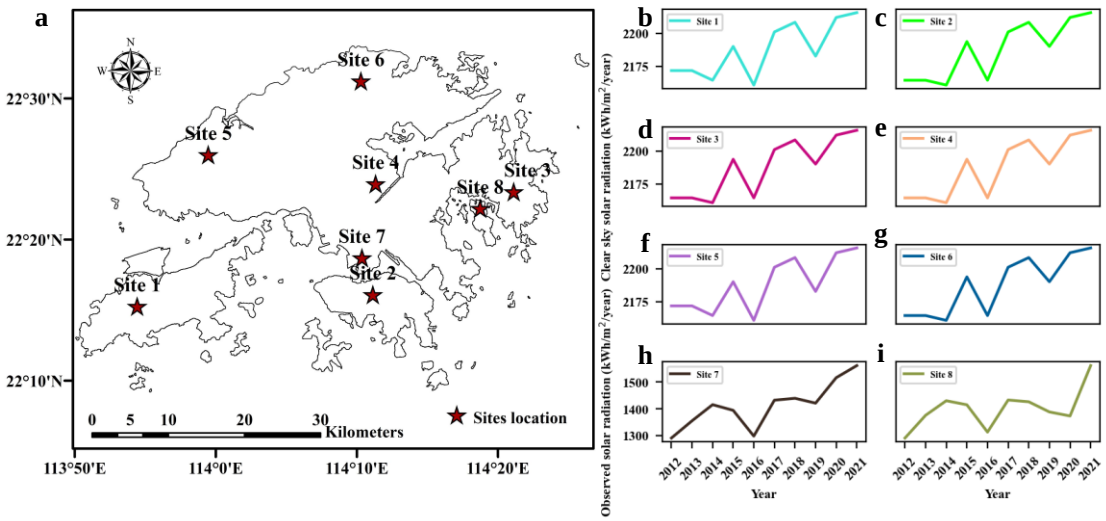


Figure 29 The change of annual solar irradiation from 2012 to 2021 in Hong Kong, China. (a)Locations of six sites. (b-g) Annual clear sky surface solar irradiation in six sites. (h-i) Annual solar irradiation from the King's Park Station and Kau Sai Chau Station

4.1.2 Dataset description

This study considers six influential factors that can affect spatio-temporal solar distribution on rooftops (Tong et al., 2005; Ko et al., 2015; Sarralde et al., 2015; Buffat et al., 2018; Mohajeri et al., 2018; Nelson et al., 2020; Walch et al., 2020), namely morphological data, DSM, building shadow, terrain shadow, tilted rooftop slope, and tilted rooftop aspect. Among them, the DSM at 1m resolution and building polygons enriched with the height attribute were obtained from the Civil Engineering and Development Department and the Lands Department of the Government of Hong Kong SAR in 2019. This study intends to investigate the specific impact of building shade and mountain shade on rooftop solar irradiation, respectively. Therefore, this study calculates the building shade using the building footprint with building height and uses DEM to calculate the mountain shade. The rooftop solar irradiation map with 1m resolution used for cross-validation is obtained from the project of Hong Kong Solar Irradiation Map for Building Rooftops which is conducted by the Electrical and Mechanical Services Department (Wong et al., 2016) and Remote Sensing Lab at Hong Kong Polytechnic University, and it is calculated by using Remote Sensing technologies and Geographic Information Systems (Wong et al., 2016).

4.1.2.1 Urban morphological data

Previous studies have proved that urban morphology can affect the building solar energy potential (Martins et al., 2014; Li et al., 2015; Sarralde et al., 2015; Zhu et al., 2019; Chatzipoulka et al., 2016; Zhu et al., 2022). Boccalatte et al. (Boccalatte et al., 2022) evaluated the impact of 40 urban morphological parameters on rooftop solar radiation. Additionally, many studies (Lopez et al., 2016; Chatzipoulka, et al., 2018; Tanu et al., 2021) suggest that the Sky View Factor (SVF) has a strong correlation with rooftop solar irradiation. Therefore, a total of 41 urban morphological parameters are calculated from building polygons in Hong Kong using a Python library named *Momepy* (Fleischmann et al., 2019). The *Momepy* library is based on several well-known Python packages for GIS-based data analysis, namely *GeoPandas* (Greenhall, 2019), *PySAL* (Rey, 2010), and *networkX* (Hagberg, 2008). These 41 morphological parameters can be divided into four categories, which can represent the building dimension, building shape, building intensity, and building spatial distribution. The list of these parameters, as well as the related equations and description, are displayed in Appendix 1.

4.1.2.2 Building shadow

Since skyscrapers in cities often cast shadows on each other (Ko et al., 2015), the mutual shadowing by buildings is considered in our study. A shadow polygon will be formed when the sunshine arrives at the rooftop. The direction of solar irradiation is described by the elevation and azimuth with a determined intensity at an instant of time.

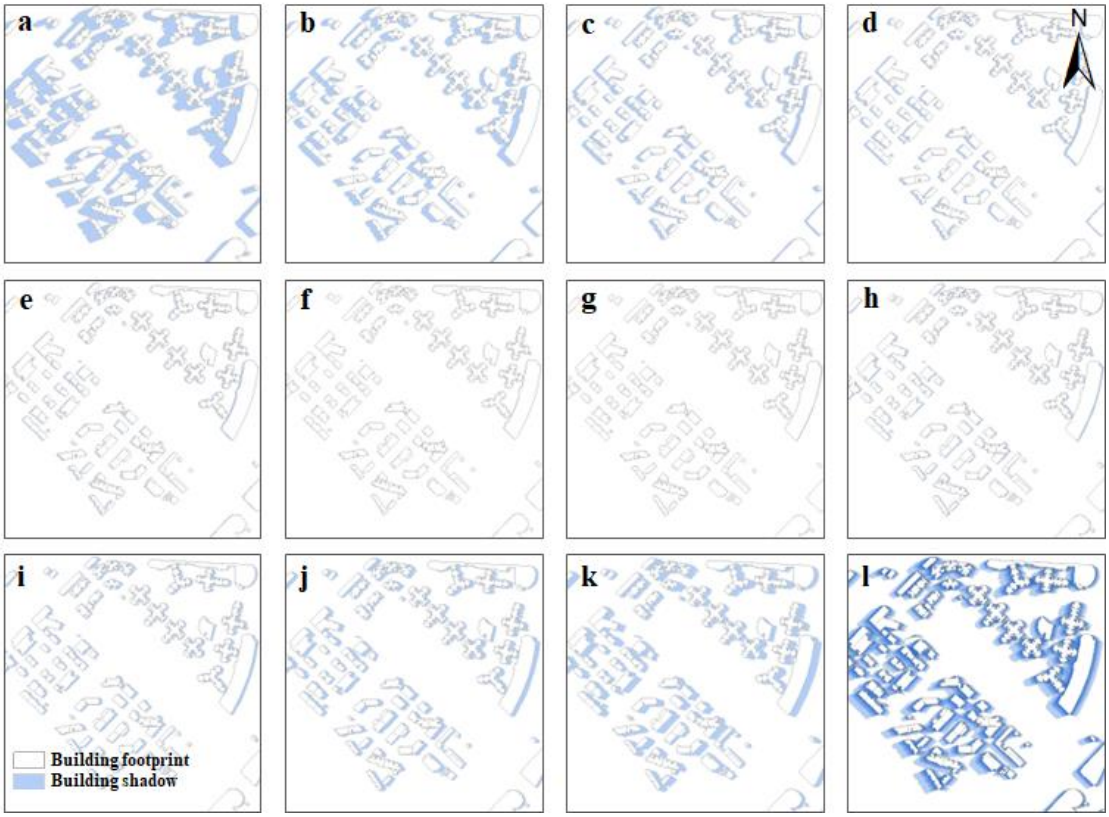


Figure 30 Hourly shadow distribution in an urban area of Hong Kong on 15th August 2019. (a-k) Hourly shadow distribution from 7 am – 5 pm. (l) Accumulated shadow distribution

4.1.2.3 Terrain shadow

Hong Kong is characterized by complex topography with high mountains and dense urban developments (Tong et al., 2005). Therefore, the effect of terrain variation on rooftop solar irradiation is considered in this study.

4.1.2.4 Rooftop slope and aspect

Previous studies have suggested that tilted rooftops with various orientations significantly affect the site selection of solar PV arrays (Buffat et al., 2018; Mohajeri et al., 2018; Nelson et al., 2020; Walch et al., 2020). Thus, the rooftop characteristics, i.e., slope and aspect, are considered the input variables in our dataset. The *Aspect* and *Slope* toolsets in ArcMap generate the rooftop aspect image and slope image at 1-m resolution based on the DSM data.

4.1.3 Dataset pre-processing and data construction

Urban morphological data and rooftop slope and aspect are static data, which are directly calculated by Python Library and ArcMap. While building shadow and terrain shadow are dynamic data, both data need to be performed in accumulation processing for transforming into annual data. The building footprints and the height information are used to generate hourly 2D building shadow polygons from 7 am to 5 pm on 15th August 2019. Generated 2D building shadow polygons are transformed into Raster images, and these shadow images are overlayed into one day 2D building shadow image. The overlapped shadow image is considered daily shadow distribution for calculating the annual total building shadow image with a 1-m resolution. Figures 30 (a) to (k) demonstrate hourly building shadow changes from 7 am to 5 pm, and Figure 30 (l) presents the accumulated shadow distribution on that day. Calculation of terrain shadow faces a challenge because its shape is irregular, so the calculated values from the *Hillshade* toolset in ArcMap are used as hourly terrain shadow from 7 am to 5 pm on 15th August 2019 using the DEM data. Next, hourly shadow distributions are accumulated on a daily basis, and it is considered the average annual terrain shadow intensity at the resolution of 1 m.

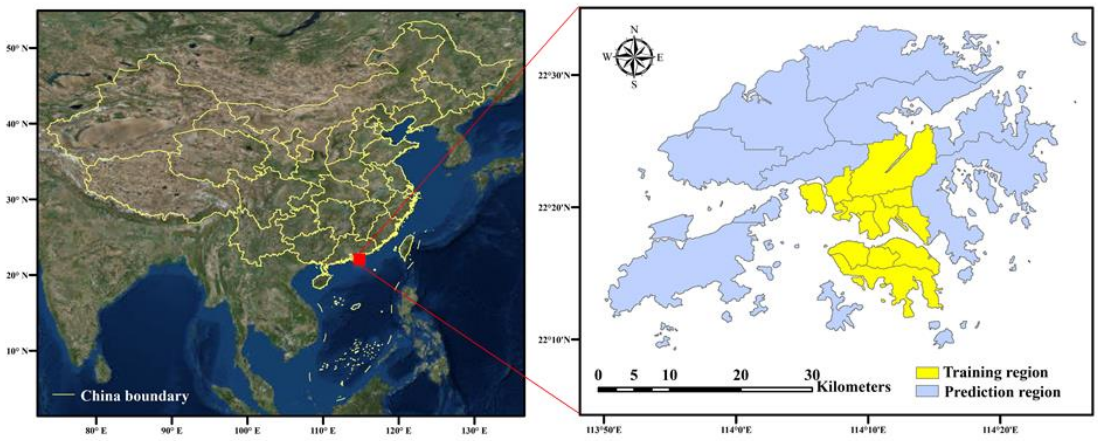


Figure 31 The specific distribution of the training and testing regions

All data are transformed into the raster files having the same resolution, orientation, and projection system (i.e., Hong Kong 1980 Grid coordinate system), which are deemed as multi-band images for training, validating, and testing. Our dataset contains 323,886 buildings that occupy 51-km^2 land surface in total. The whole data is organized according to the 18 districts. We divided these districts into the training and prediction regions (Figure 31), which respectively account for 45% and 55% of the total rooftop area. This is based on two

considerations. First, the training region covers high-density, middle-density, and low-density buildings, and the amount of building rooftops is well-sufficient for training and testing the models. Second, to evaluate the model performance, this study trains and validates the models and utilizes the resulting models to estimate the rooftop solar potential based on the testing dataset. The ratio of the training and testing datasets is 9 to 1. To improve the quality of the dataset, the outliers of all data values in datasets (i.e., the null value and infinite) which consumed 0.18% of the entire dataset are filtered out in the dataset.

4.2 Methodology

This study proposes a fast and accurate method based on the machine learning model for the estimation of annual rooftop solar irradiation over an urban area, with a flowchart presented in Figure 32. Firstly, the MT method (Fleischmann et al., 2020) is used to calculate morphological features. Secondly, as a preliminary analysis to investigate the relationship between solar irradiation and the 41 morphological features (appendix 1), Pearson correlation analysis has been performed to test the effectiveness of the proposed indices. To improve the training efficiency, Random Forest, a widely used machine learning model particularly useful for classification and prediction, is used to select suitable input variables. Furthermore, the estimation results are compared using different machine learning models to select the optimal model based on criteria of fast computation and the highest performance to estimate annual rooftop solar irradiation. Finally, this study analyzed the distribution of mean annual solar irradiation received by rooftops on different rooftop slopes and aspects for providing a reliable reference for the effective deployment of solar PV arrays.

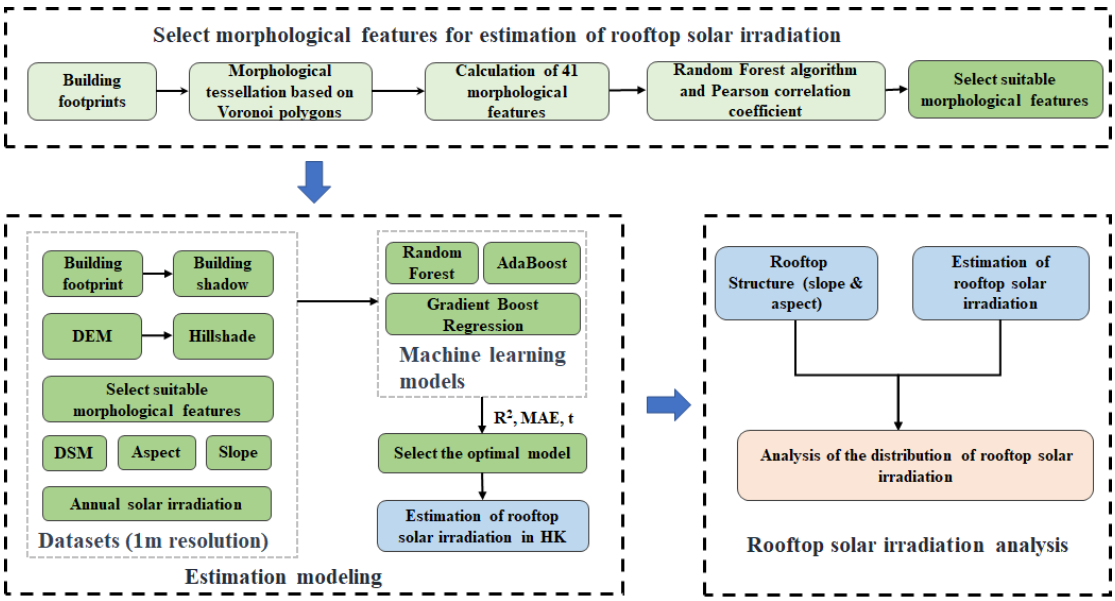


Figure 32 Flow chart for estimating rooftop solar irradiation

4.2.1 Calculation of morphological features

Forty-one morphological features used in this study are divided into four categories. The specific classification is shown in Appendix 1. The features related to the categories of building dimension and building shape are directly calculated based on the building footprint. Additionally, this study employs morphological tessellation cells (MTC) (Fleischmann et al., 2020) to define a reference boundary for calculating the features related to building intensity. The aim of spatial distribution analysis is to calculate the spatial relationship among buildings, so the spatial distance of 200m between a building and its adjacent buildings is employed based on the previous studies (Edussuriya et al., 2011; Ng et al., 2011) for calculating the features related to building spatial distribution.

4.2.1.1 Morphological tessellation

Using a boundary for calculating the building density information requires the selection of a specific spatial scale which is based on a grid or the administrative district for calculating building density (Leng et al., 2020). However, this selection of an appropriate spatial scale usually relies on empiricism (Wei et al., 2016; Yong et al., 2017; Javanroodi et al., 2018; Lima et al., 2018), and this leads to time-consuming for selecting the spatial scale and narrow scope of application. Compared to the above-mentioned traditional method, MTC is a geometric derivative of Voronoi polygons obtained from building footprints, and it represents the smallest

spatial unit that delineates the portion of land around each building (Boccalatte et al., 2022). This allows us to obtain the density information related to buildings for a better estimation of solar distribution. The process of generating morphological tessellation consists of five steps: (i) inward offset from building footprint; (ii) discretization of polygons' boundaries into points; (iii) generation of Voronoi cells; (iv) dissolution of Voronoi cells; (v) clip of preliminary tessellation. Figure 33 shows the morphological tessellation distribution in Hong Kong. Based on the MTC, it is feasible to capture the influence that each building exerts on the surrounding space as well as the building-related density information.

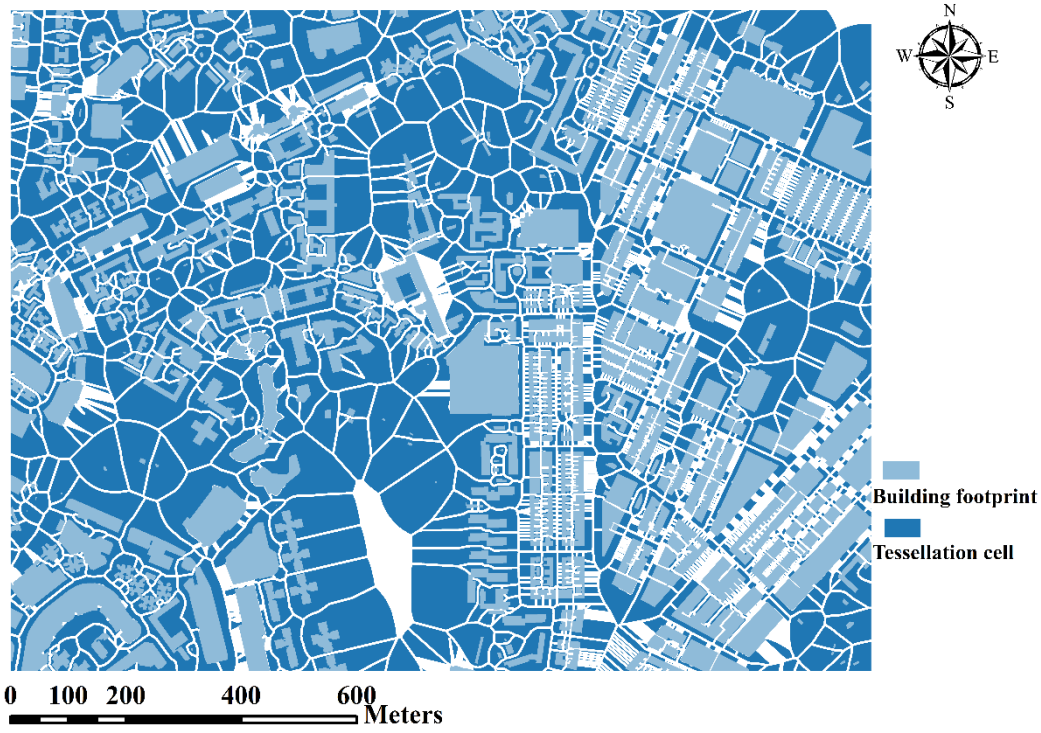


Figure 33 Building footprints and related tessellation cells of a specific area of Hong Kong

4.2.2 Machine learning models

4.2.2.1 Random Forest Regression

RF (Segal, 2004) is a combined regression model that is composed of a large amount of decision trees $h(X;\theta_k)$ $(k = 1,2,\dots, n)$, where θ_k is an identically distributed random vector, n is the number of decision trees, and X is the input variables, namely, morphological indices, DSM, building shadow, terrain shadow, rooftop slope, and rooftop aspect. We set Y as the output variable that denotes the solar irradiation values calculated by the physical model, then (X, Y)

composed of the original datasets. The RF model uses the Bootstrap method (Efron and Tibshirani, 1985) for sampling of input datasets, and then employs the complete splitting method to construct the decision trees.

4.2.2.2 Gradient Boost Regression Tree

GBRT (Friedman, 2001) is an ensemble model using forward addition based on the base function of the classification and regression tree (CARF). The process of constructing one CARF consists of two parts, generation of the decision tree and decision tree pruning. In the process of constructing the GBRF model, the squared error is used as the learning target of the model (Equation 5):

$$L(y_i, f(x)) = min \sum_{i=1}^{N} (y_i - f(x))^2 \quad (5)$$

Where *y* represents the solar irradiation values calculated by the physical model as ground truth, *x* denotes seven input variables in our dataset, *f(x)* represents the prediction value of the model, and *N* represents the size of the sample.

4.2.2.3 Adaboost Regression Tree

The Adaboost algorithm (Duraˇciov´a and Pruˇzinec, 2022, Schapire and Singer, 2000) is one of the best supervised learning methods with satisfactory prediction performance. This algorithm can inherit many weak regression models and finally form a strong model. The processing of the AdaBoost algorithm is that for the same sample points, their weight will be continually updated for training multiple weak regression models, then the weak regression models with different weights are composed of a final strong regression model.

4.2.3 Selection importance parameters

The importance of each variable in RF can be estimated by the random sampling method. The original sample size is set as *N*, and variables are $x_1, x_2, \dots, x_m$ (here *m*=46, because we have 46 input variables in our dataset). Each time a bootstrap method is used to randomly select a

sample from the total sample and randomly select *n* times with replacement. *n* bootstrap samples generate *n* regression trees. The samples that are not drawn each time are composed of *n* out-of-bag data as test samples, so that the importance order of each variable in the classification regression can be obtained.

4.2.4 Estimation model for annual rooftop solar irradiation

Physical modeling for solar irradiation estimation is usually performed based on physical principles and mathematical modeling, thus these methods are often referred to as model-driven methods (Kundur et al., 2004). Model-driven methods have the advantages of clear logistic and rigorous derivation. However, improving computing efficiency by simplifying the model would always lead to a decrease in the estimation accuracy (Baltas et al., 2018). Machine learning can extract knowledge from massive data with the advantages of high computing efficiency and high accuracy, thus these methods are known as data-driven methods. Nevertheless, these methods are dependent on a prior-knowledge and database. Since it is difficult to measure the ground truth of rooftop solar irradiation in large-scale regions, it faces challenges in obtaining ground truth for training the model. To rapidly and accurately estimate rooftop solar irradiation, a data and model dual-driven loosely coupled approach was proposed by the integrations of machine learning models and physical models. In this study, the estimation model consists of two modules, namely, the model-driven methods and the data-driven methods. The model-driven method uses the method proposed by Wong et al. (Wong et al., 2016), which is based on the hemispherical viewshed algorithm developed by Rich et al. (Rich et al., 1994), to calculate hourly rooftop solar irradiation. The specific hemispherical viewshed algorithm is introduced in Appendix 2. The hourly solar data is accumulated into annual rooftop solar irradiation, and the model-driven model passes the annual data onto the data-driven model as the ground truth. In the data-driven model, three machine learning methods are compared to select the optimal one for estimating all rooftop solar potential by using three evaluating indicators, including R^2, Mean Absolute Error (MAE), and computation time.

4.3 Results and Discussion

4.3.1 Correlation analysis between morphological features and rooftop solar irradiation

Pearson correlation analysis is performed to investigate the relationship between annual rooftop solar irradiation and 41 morphological features, and the Variance Inflation Factor (VIF) (Neter, et al., 1996) is used to diagnose multicollinearity for these morphological features. These features can be divided into four categories, namely, building dimension, building shape, building intensity, and building spatial distribution (Boccalatte et al., 2022). Table 4 shows the Pearson correlation coefficients and corresponding *p*-values between rooftop solar irradiation and each morphological feature, and Table 5 shows multicollinearity among these features. The most explainable parameters relevant to rooftop solar irradiation are related to the building shape (i.e., building shape index (Shp_{idx}), building Rectangularity (Rec), building equivalent rectangular index (ERI), building circular compactness (Com), building elongation (Elg), and building square compactness (Squ_{com})), with R>0.65. Furthermore, the parameters related to the building dimension (i.e., building fractal dimension (Fra), Tessellation longest axis length (LAL_{tess}), building volume to façade (VFR), and building longest axis length (LAL)) and ones related to the building spatial distribution, including, negative average neighborhood shading angle (Shd_{an}), building adjacency (Adj), sky view factor (SVF), and mean inter-building distance (IBD_{mean}). They show strong and positive correlations with R ranging between 0.65 and 0.7. On the contrary, the parameters with strong correlation coefficients do not consist of those related to the building intensity. For example, CAR has moderate correlations with rooftop solar irradiation, with R=0.61. The results indicate that the mentioned parameters related to the building shape, dimension, and spatial distribution can greatly affect the amount of the receiving rooftop solar irradiation. Only one feature, building shape index, has *p*-values between 0.01 and 0.005, which suggests that it is statistically significant at the 0.01 level. The *p*-values of four features (i.e., positive average neighborhood shading angle, average building height, rugosity, and alignment) are between 0.01 and 0.05, which suggests that they are statistically significant at 0.05 level. In addition, the *p*-values of the remaining 36 features are less than 0.005, which suggests that they are statistically significant at the 0.005 level. The

small values of VIF (VIF≤10) corresponding to 11 morphological features suggest there is no issue with multicollinearity. In addition, the values of VIF corresponding to 16 morphological features are between 10 and 100, which suggests that these features have moderate multicollinearity, while the remaining morphological features whose values of VIF are larger than 100 suggest that they have strong multicollinearity.

Table 4 Results of Pearson correlation analysis between rooftop solar irradiation and each morphological feature

Name	**R**	**Name**	**R**	**Name**	**R**
Fra	0.70(***)	CAR_{mean}	0.62(***)	A	0.47(***)
Shp_{idx}	0.70 (**)	P	0.62(***)	HD_p	0.47(***)
Rec	0.69(***)	CAR	0.61(***)	V_{mean}	0.47(***)
ERI	0.69(***)	Shd_{ap}	0.61(*)	TFA_{mean}	0.46(***)
Com	0.68(***)	H	0.58(***)	Squ	0.42(***)
Adj	0.68(***)	H_{mean}	0.57(*)	SWR	0.42(***)
Shd_{an}	0.68(***)	FAR_{mean}	0.54(***)	V	0.40(***)
SVF	0.67(***)	Ort	0.53(***)	Flr_{area}	0.39(***)
Elg	0.67(***)	A_{mean}	0.51(***)	HW	0.35(***)
IBD_{mean}	0.66(***)	Rug	0.50(*)	H_{abs}	0.11(***)
Squ_{com}	0.66(***)	HD_n	0.49(***)	N_{neigh}	0.019(***)
VFR	0.65(***)	FA	0.49(***)	HD	-0.28(***)
LAL_{tess}	0.65(***)	A_{tess}	0.47(***)	Shd_{mean}	-0.35(***)
LAL	0.65(***)	Ali	0.47(*)		

(*) p<=0.05, (**) p<=0.01, and (***) p<=0.005

Table 5 Results of multicollinearity analysis among morphological features

Group	Name	VIF	Name	VIF	Name	VIF	Name	VIF
VIF>100	V_{mean}	8110	Rug	4542	Com	1069	LAL	109
	TFA_{mean}	7712	FA	4334	Fra	763		
	Flr_{area}	5393	Shp_{idx}	3463	Squ_{com}	469		
	V	5388	ERI	1485	Rec	181		
10<VIF≤100	P	96	FAR_{mean}	50	CAR	23	A_{mean}	17
	H_{mean}	54	CAR_{mean}	38	LAL_{tess}	22	Adj	14
	Shd_{an}	52	VFR	35	Shd_{ap}	21	Nneigh	14
	Elg	51	H	26	SVF	17		
VIF≤10	IBD_{mean}	10	A_{tess}	6	Ort	3	HW	1
	HD	8	H_{abs}	6	A	3		
	HD_n	7	HD_p	5	Ali	2		
	SWR	7	Shd_{mean}	4	Squ	2		

4.3.2 Parameters selection and importance analysis

The RF model is often employed for calculating the parameter importance and selecting variables for training machine learning models in some studies (Gr¨omping et al., 2009; Bhanujyothi et al., 2014; Dewi et al., 2019). The results are sorted in descending order as shown in Table 6. It presents that DSM makes a significant contribution to our estimation model, with 0.55 importance, followed by the rooftop slope. However, the importance of other parameters is close to zero. To increase the efficiency of the computation, we conducted recursive parameter selection for selecting useful parameters. To improve the efficiency of the selection, the interval of eliminating parameters was flexibly adjusted based on three indicators, i.e., R^2, MAE, and computation time. The order for eliminating parameters is based on the values of importance from low values to high values. Table 7 shows R^2, MAE, and computation time for recursively eliminating parameters from the parameter set. The RF model with 46 parameters showed the highest estimation accuracy, with R^2=0.78. We further calculated the importance of these 46 parameters and arranged these parameters in descending order. The corresponding parameters are sequentially removed from the dataset parameter list, starting with the smallest value, based on their order of importance. Initially, the parameter removal interval was set at four. However, even with this interval, the R^2 value remained consistently at 0.78. As a result, the interval was adjusted to ten. However, this adjustment resulted in a slight increase in MAE. Consequently, the interval was further refined to eight. Considering the MAE and computation time, when the R^2 value decreased by 0.77, the interval was adjusted to one. Overall, as the number of parameters gradually decreases, corresponding R^2 and computation time also reduce, and MAE slightly increases. To balance the performance regarding the three indicators, models built by seven parameters are considered suitable for estimating rooftop solar irradiation, achieving high accuracy and fast computation. This is because the R^2 and MAE of the model with seven parameters are near that of the model with 14 parameters, and using seven parameters can save half the computation time than using 14 parameters. The final dataset consists of DSM, shadow from the surrounding buildings, shadow from natural terrain, rooftop aspect, rooftop slope, building volume to façade ratio, and SVF.

Table 6 The importance between rooftop solar irradiation and each parameter

Name	I	Name	I	Name	I	Name	I
DSM	0.55	H_{mean}	0.0049	HW	0.0034	SWR	0.0027
Slope	0.13	Elg	0.0047	V	0.0034	FA	0.0027
Shadow	0.055	IBD_{mean}	0.0045	Shd_{an}	0.0031	Rug	0.0025
Aspect	0.039	FAR_{mean}	0.0041	ERI	0.0031	TFA_{mean}	0.0024
SVF	0.026	LAL_{tess}	0.0040	HD_n	0.0030	Shp_{idx}	0.0024
Hillshade	0.021	Rec	0.0040	Flr_{area}	0.0030	V_{mean}	0.0023
VFR	0.016	CAR_{mean}	0.0039	HD	0.0030	Squ_{com}	0.0023
A_{tess}	0.0070	CAR	0.0038	Adj	0.0029	Com	0.0023
Ort	0.0064	HD_p	0.0038	Shd_{mean}	0.0028	H_{abs}	0.0023
Ali	0.0056	H	0.0038	Fra	0.0028	N_{neigh}	0.0021
Squ	0.0051	A_{mean}	0.0038	LAL	0.0027		
A	0.0051	Shd_{ap}	0.0037	P	0.0027		

Table 7 R^2, MAE, and time for recursively selecting parameters

No. of input parameters	R^2	MAE (kWh/m^2/year)	Time (h)
46	0.78	20.71	20.47
42	0.78	20.72	19.27
38	0.78	20.73	17.10
28	0.78	20.77	12.50
20	0.78	20.94	8.54
14	0.78	21.21	6.01
7	0.77	22.83	3.00
6	0.74	24.77	2.50
5	0.74	24.63	2.10
4	0.70	27.33	1.70
3	0.67	27.74	1.30

4.3.3 Estimation of annual rooftop solar irradiation using machine learning models

The experiments were performed on a desktop with Intel Core i7-9700K CPU and 32 GB memory. Five-fold cross-validation (Rodr´ıguez et al., 2009) was performed to train and test each model. Specifically, the original dataset was randomly divided into five equally-sized sub-datasets. Among these five sub-datasets, one was designated as the validation data for evaluating the performance of machine learning models, while the remaining four sub-datasets were utilized as the training data. The grid search method (Baltas et al., 2018) was used to optimize the hyper-parameters, and the optimization of hyper-parameters can be found in Table 8.

Table 8 The hyper-parameters of the different machine learning models

Model	The used hyper-parameters
RF_{46}	$n\_estimators$ ： 200, $min\_samples\_leaf$ ： 1, $min\_samples\_split$ ： 2
RF_7	$n\_estimators$ ： 100, $min\_samples\_leaf$ ： 1, $min\_samples\_split$ ： 2
GBM_7	$n\_estimators$ ： 100, $learning\_rate$ ： 0.1, max_depth ： 3, $subsample$ ： 0.8
$AdBoost_7$	$n\_estimators$ ： 50, $learning\_rate$ ： 1.0, $base\_estimator$ ： $deprecated$, $loss$ ： $linear$

After selecting the final dataset, this study compared the estimation performance of three machine learning models in the Kowloon district (Table 9). The RF model using 46 parameters (RF_{46}) obtains the highest estimation accuracy with R^2=0.79 and MAE=20.71 kWh/m^2/year. However, it costs 21.16 hours to train a robust model. The value of R^2 of the RF model using seven parameters (RF_7) is close to that of the RF_{46} model, while the computation time of the RF_{46} model is around seven times longer. This suggests that the RF_{46} model is a complex estimation model with redundant parameters, which can lead to low computational efficiency.

Although the AdaBoost model utilizing seven parameters ($AdBoost_7$) spends the least time for training, R^2 is only 0.58. This means that this model has a low capability to accurately estimate rooftop solar irradiation in the study area. The performance of RF_7 and GBM utilizing seven parameters (GBM_7) is followed by that of RF_{46}. The R^2 and MAE of the RF_7 datasets are higher than those of GBM_7, while the RF_7 model takes twice as long as the GBM_7 model.

To investigate the estimation accuracy of the models, this study also calculated the absolute errors. Table 10 shows the absolute error distribution in different ranges using four models in the Kowloon district. The percentages of the absolute errors within 20 kWh/m^2/year are 94.15% for RF_{46}, 93.79% for RF_7, 92.67% for GBM_7, and 92.53% for $AdBoost_7$, respectively. For all models, the absolute errors over 500 kWh/m^2/year account for less than 6%. Compared with the four models, $AdBoost_7$ shows a slightly worse estimation performance, with around 8% of the absolute errors over 20 kWh/m^2/year. Overall, these four machine learning models show satisfactory estimation performance. This means that the estimation accuracies of all models are high and these models can provide reliable estimation results.

Combined with the results of Table 9 and Table 10, the performance of estimation accuracy (i.e., R^2, MAE, and absolute error) of two RF models is better than the GBM model and AdaBoost model. To greatly investigate the computation efficiency of two RF models, this study compared calculation time for the calculation of the dataset, training model, and prediction using two RF models in Hong Kong. Table 11 is the result. Overall, the calculation time for each part using the RF_7 model is obviously less than that using the RF_{46} model. Especially, the training time using RF_{46} is more than 26 times longer than that of using RF_7. Thus, considering the estimation accuracy and computation time, we selected the RF model with seven parameters dataset for estimating the rooftop solar irradiation in the whole area of Hong Kong to obtain a balance between time cost and estimation accuracy.

The annual rooftop solar irradiation map was created by using the RF_7 model. Figure 34 (a) displays the annual rooftop solar map in Hong Kong, and Figures 34 (b) to (e) show annual rooftop solar maps in Hong Kong Island, Central and West, Yuen Long, and Kowloon, respectively. The high-density area shows smaller rooftop solar potential, while the low-density

area shows larger solar potential. This is because buildings in dense areas are greatly affected by the shadow from surrounding buildings. Therefore, the shadow effect is a significant factor in estimating rooftop solar irradiation in dense cities

Table 9 R^2, MAE and time of different models in Kowloon

Model	R^2	Mean absolute error (MAE) (kWh/m^2/year)	Time of training models (h)
RF_{46}	0.79	20.71	21.16
RF_7	0.77	22.83	3.00
GBM_7	0.71	28.72	1.47
$AdaBoost_7$	0.58	42.25	0.87

Table 10 Absolute error distribution in different models in Kowloon

Model	Range of the absolute error (kWh/m^2/year)			
	0-20	20-500	500-1000	>1000
$AdBoost_7$	92.53%	1.62%	5.86%	0.00%
GBM_7	92.67%	5.14%	1.87%	0.31%
RF_7	93.79%	4.39%	1.58%	0.24%
RF_{46}	94.15%	4.22%	1.36%	0.27%

Table 11 The comparison of calculation time for calculation of the dataset, training model, and prediction using two RF models in Hong Kong

Model	Time for calculation of the dataset (h)	Time for training model (h)	Time for prediction (h)
RF_{46}	32.79	319.79	2.41
RF_7	20.84	12.13	0.85

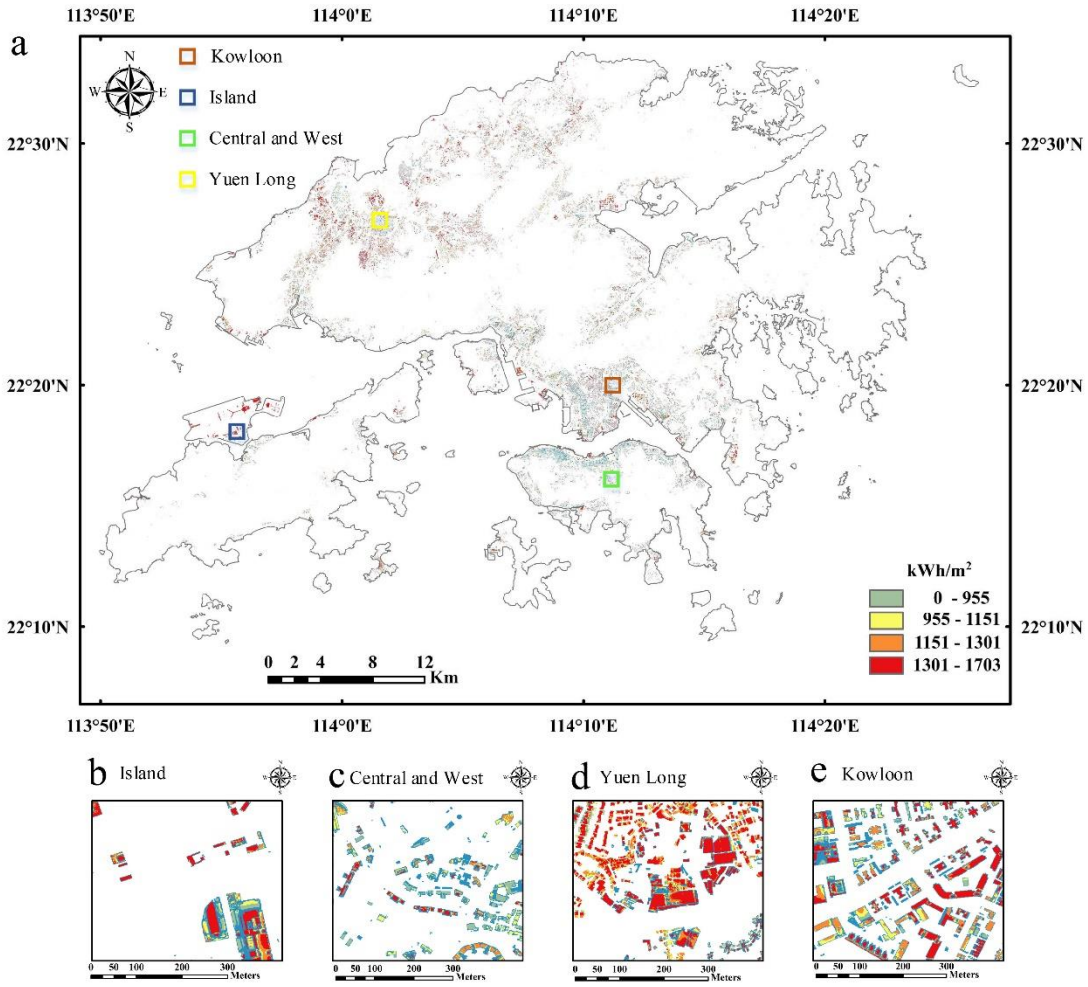


Figure 34 Annual rooftop solar irradiation map in Hong Kong. (a) The whole region. (b) Hong Kong Island. (c) Central and West. (d) Yuen Long. (e) Kowloon.

To evaluate the usability and generalization ability of our model, this study compared Mean Relative Error (MRE) between the training and prediction regions (Table 12 and Table 13). For training the model, the MRE of all the training regions is within 7%, and the time for training the model is approximately 12 hours. For estimating the whole of Hong Kong, the MRE varies from about 9% to 5%, and the computation time is about 0.85 hours. Although the MRE of the prediction regions is slightly higher than that in the training regions, this is a high estimation accuracy for the prediction regions which are not trained. The results indicate that our model has good generalization capability.

Table 12 The prediction accuracy in training regions

Region	MRE	Region	MRE	Region	MRE	Region	MRE
Central and Western	6.20%	Kwai Tsing	3.88%	Sham Shui Po	4.19%	Wong Tai Sin	4.26%

East	4.57%	Kwun Tong	3.93%	Southern	4.50%	Yau Tsim Mong	4.71%
Kowloon	4.58%	Sha Tin	4.73%	Wan Chai	5.69%		

Table 13 The prediction accuracy in prediction regions

Region	MRE	Region	MRE	Region	MRE	Region	MRE
Hong Kong Island	9.11%	Sai Kung	7.17%	Tai Po	8.37%	Tsuen Wan	9.27%
Tuen Mun	9.30%	Yuen Long	9.59%	North	5.16%		

4.3.4 Accuracy assessment of physical model

The estimation of rooftop solar irradiation from the physical model was employed as the ground truth to cross-validate the machine learning models. To assess the accuracy of the physical model, field verification was conducted at five different sites, including a single-house rooftop in Kam Tin, a 20th-floor rooftop of private housing in Sha Tin, a sky garden at the Hong Kong Polytechnic University (PolyU), the lawn at the HKO King's Park Station, and a secondary school rooftop in Tseung Kwan O. Measurements were taken using MS-802, CM21, and CMP11 pyranometers. Table 14 shows the details of field verification, including data collection periods, site names and locations, and the equipment used. Additionally, Table 15 presents the comparison between validation field data and the estimated global horizontal solar irradiation using the physical model. Overall, the model achieves a high accuracy of 95.99% with an MRE of 4.01%. These results affirm the highly accurate performance of the physical model, validating the reliability of estimation values derived from it as ground truth.

Table 14 Details of field verification

Site	Period	Location Name	Coordinates	Equipment Used
1	22 Feb 2020−25 Feb 2020	Kam Tin	(22.24, 114.07)	MS-802, CM21
2	25 Feb 2020−28 Feb 2020	Sha Tin	(22.38, 114.20)	MS-802, CM21
3	29 Apr 2020−6 May 2020	PolyU	(22.31, 114.18)	MS-802, CM21
4	27 Aug 2020−7 Sep 2020	King's Park Station	(22.31, 114.17)	MS-802, CM21, CMP11

5	30 Dec 2020-6 Jan 2021	Tseung Kwan O	(22.32, 114.26)	CMP11

Table 15 Comparison between validation field data with the estimated result at the five validation sites

Site	Estimated result (Wh/m2)	Measurement (Wh/m2) (MRE)		
		MS-802	**CM21**	**CMP11**
1	18,979	19,092 (−0.59%)	19,711 (−3.71%)	N/A
2	12,771	11,331 (12.71%)	11,408 (11.95%)	N/A
3	35,914	35,964 (−0.14%)	37,508 (−4.25%)	N/A
4	42,631	43,038 (−0.94%)	44,412 (−4.01%)	44,264 (−3.69%)
5	24,357	24,357	24,357	24,232 (0.51%)

4.3.5 Comparison between physical model and machine learning model

The physical model demonstrated highly accurate results, with hourly estimations exhibiting 4.01% MRE for the entire year. Although the accuracy of the RF_7 model at 7.72% MRE is slightly lower than that of the physical model, both models can provide high accuracy and reliable estimation results. For comparison, the physical model and RF_7 model were used to estimate rooftop solar irradiation on 5,334 buildings. These buildings cover 423,876 square meters which are selected randomly from the 18 districts in Hong Kong. The physical model spent around 927 seconds, whereas the RF_7 model cost 6 seconds. From Table 8, it is clear that all calculation time required for estimating the whole rooftop solar potential using the RF_7 model is approximately 33.82 h. In contrast, the physical model needs to spend nearly a year to complete the same estimation. This demonstrates that our model can greatly reduce computation time and thus overcome the low-efficiency problem when using the physical model. For the input parameters, the physical model utilized DSM, locational information (latitude and longitude), slope, aspect, and direct and diffuse solar radiation to calculate the rooftop solar irradiation, obtaining high spatial-temporal resolution of solar radiation. Compared with the physical model, the RF_7 model only uses seven parameters (i.e., DSM,

shadow from the surrounding buildings, shadow from natural terrain, rooftop aspect, rooftop slope, and SVF) to achieve highly accurate and efficient estimation, and these parameters are relatively easy to obtain a reliable generalization.

4.3.6 Analysis of rooftop solar irradiation distribution

After the estimation of annual mean solar irradiance by applying our trained model, we conducted an analysis to explore the relationship between annual mean solar irradiance and slopes as well as aspects. This analysis involved associating the estimated solar irradiance values with their corresponding slopes and aspects, which were determined based on the geographical locations of the rooftops. Figure 35 visualizes the average annual solar irradiation received by the rooftop as a function of roof slope and aspect. Overall, the annual mean solar irradiation received by rooftops is high, from 1,120 kWh/m^2 to 1, 280 kWh/m^2. This suggests that solar potential on rooftops in Hong Kong could generate a considerable amount of electricity efficiently. The distribution of rooftop solar irradiation is east-west symmetry, and the values of solar irradiation gradually decrease from north to south. Furthermore, the largest irradiation is found for south-facing rooftops with a slope between 30 and 40 degrees, while north-facing rooftops have the lowest solar irradiation, smaller than 1, 200 kWh/m^2. The results are in line with the order of nature that the sun shines mainly from the south to the north in the northern hemisphere.

The team also investigated annual mean solar irradiation on rooftops surfacing different aspects and slopes, respectively. Results in Figure 36 (a) show that the rooftops facing south receive the greatest solar irradiation, followed by rooftops facing west and east; in particular, the north receives the least irradiation, as expected. Figure 36 (b) shows that the flat to gently sloping rooftops, with the slope ranging from 0-40 degrees, receive the greatest solar irradiation. The results illustrate the steeper the slope, the smaller the received solar irradiation.

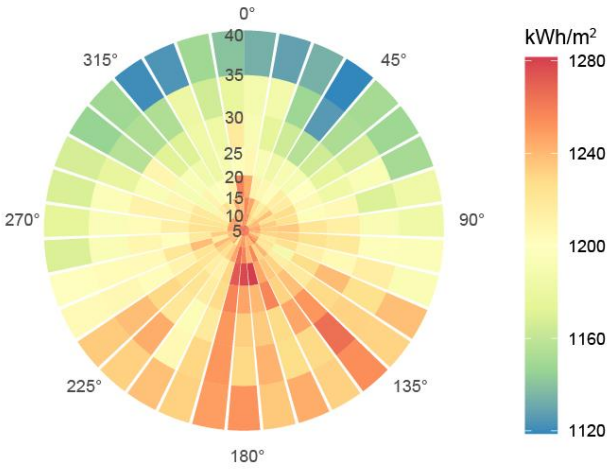
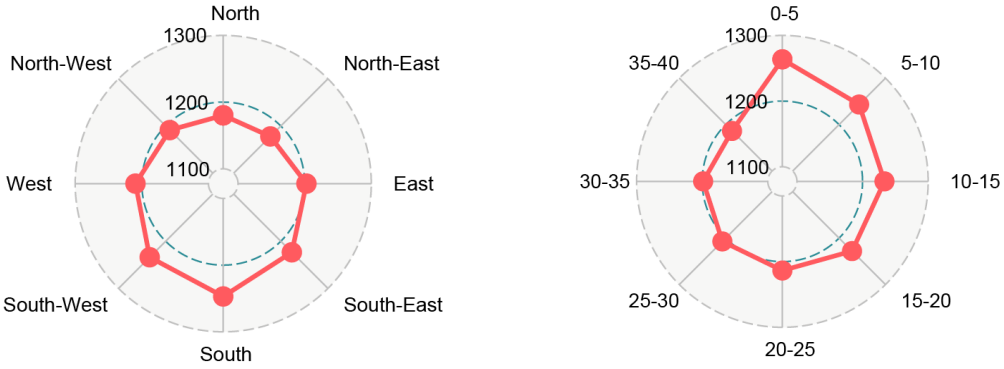


Figure 35 Annual mean solar radiation (kWh/m^2) as a function of slope and aspect of roof surfaces for buildings in Hong Kong, China



(a) (b)

Figure 36 Annual mean solar irradiation (kWh/m^2) of roof surfaces for different ranges of (a) aspect and (b) slope

4.4 Conclusion

This study proposes a fast and accurate method to estimate annual rooftop solar irradiation at a spatial resolution of 1 m in Hong Kong, China. This study parameterizes the influential factors (i.e., morphological features, building rooftop structures, DSM, the shadow from buildings, and the shadow from terrain) and quantifies the importance of these features on rooftop solar irradiation. Compared between RF, GBM, and AdaBoost, the RF model is used for the estimation of annual rooftop solar irradiation for individual buildings since its estimation accuracy is high (R^2=0.77), and the computation speed is fast. Compared with the physical model, the machine learning models developed in this study can greatly reduce the computation time for rooftop solar estimation at fine spatio-temporal scales. These results suggest that our

method can estimate rooftop solar irradiation on individual buildings, which is useful for solar related applications, such as planning rooftop PV arrays. As our developed models are well-trained and validated with satisfactory scalability in various spatial and temporal resolutions, it is possible to apply these models to other regions having a similar built environment, and the proposed method is also deliverable for entirely different areas.

The traditional methods of calculating building density usually require the definition of a reference boundary, which is generated by a grid or administrative limits of a district. However, these methods just calculate the average value in a certain portion and fail to capture site-specific and density information related to buildings. The morphological tessellation method used in our study can overcome this limitation, which makes it possible to capture the specific impact of surrounding space on each building.

This study approximated the estimated rooftop solar irradiation from the physical model as the ground truth for cross-validation. This is because getting field measurements by installing high-density of solar sensors on all the rooftops in Hong Kong is almost impossible. This is one of the feasible solutions as previous studies also utilized a similar method for validation (Gastli et al., 2010; Duraˇciov´a et al., 2022).

Compared with the conventional physical models, such as the upward-looking hemispherical viewshed algorithm, our approach is 5,592 times faster in computing annual solar potential on all rooftops in Hong Kong. Therefore, we conduct recursive parameter selection to filter out redundant parameters based on the balance of estimation accuracy and computation time. Results of the model with seven parameters show high accuracy with fast computation, and this indicates that this model satisfies the requirements for estimating rooftop solar irradiation in terms of accuracy and computation speed.

However, the developed method outperformed when compared with others, it has some limitations, i.e. we calculated building shadow and terrain shadow on one specific day to represent annual shadow distribution, which would affect the estimation accuracy to some

extent. This is because the calculation of hourly building shadow and terrain shadow for one day with high spatio-temporal resolution requires a computation time of around 24 hours. In this regard, this study uses this estimation method for shadow data. From the final results of the estimation of rooftop solar irradiation, the method proposed in this study can provide high estimation accuracy. Therefore, using generalized shadow data can decrease the computation time and confirm estimation accuracy at the same time.

In conclusion, the author proposes a fast and accurate parametric method for estimating rooftop solar irradiation based on the machine learning method using seven parameters (DSM, SVF, shadow from buildings, shadow from mountains, VRF, slope, and aspect). The results demonstrate that the proposed method can provide a reliable, fast, highly accurate reference for potential applications, including solar PV installation planning, financial analysis and investment decision-making, and urban planning. Specifically, our results can help relevant parties identify suitable locations for solar PV installations and make decisions on the feasibility and optimal placement of solar panels on rooftops. The method also can help to assess the potential solar energy generation and associated cost savings, enabling governments and companies to evaluate the viability and profitability of rooftop solar PV installations. Additionally, the highly accurate estimation results can provide reliable references to optimize the orientation and layout of future constructions and maximize solar energy utilization which can help to effectively reduce the Urban Heat Island.

Chapter 5 A data-model dual-driven loosely coupled approach for fast and accurate estimation of hourly rooftop solar irradiation at the building scale

The annual rooftop solar irradiation can be estimated using the method proposed in Chapter 4, which is crucial to evaluate the solar development potential for the city. However, dynamic solar radiation data can help power companies monitor changes in PV output in real-time and predict short-term generation trends. This enables them to take proactive measures to address potential supply fluctuations, thereby enhancing grid stability. Therefore, this Chapter proposes a combined approach using a physical model and machine learning decoupling for the hourly refined estimation of rooftop solar irradiation. Due to the lack of real-time observational data for urban rooftop solar radiation, a partial calculation of rooftop solar radiation values is initially performed using a physical model, with these values serving as ground truth for deep learning. Rooftop solar radiation is influenced by various factors, including shading from buildings, meteorological conditions, and urban morphology. This chapter employs the quantification parameter approach from the previous chapter to parameterize these data and investigates the relationship between static and dynamic parameters and the hourly dynamic changes in rooftop solar energy. Finally, this paper employs the DGTFT deep learning method to estimate urban rooftop solar radiation on an hourly basis. The application of this method contributes to a more comprehensive understanding of urban rooftop solar potential and provides robust support for urban photovoltaic development planning.

5.1 Study area and data

5.1.1 Study area

We also choose Hong Kong as the study area in this study. Since the meteorological data is considered as the inputs for estimating hourly rooftop solar irradiation, we selected the five sites which buildings are the nearest meteorological stations. The distribution of five sites is shown in Figure 37. There are two reasons why this study selects these five sites to generate the dataset for modeling. First, the meteorological data shows little difference in the small

region, so the distance among these sites is quite far. Second, the locations of selected sites are representative. We selected two sites in high-density buildings in the urban region and three sites in low-density buildings in rural regions.

Table 16 The details of the five sites

Near station name	Site ID	Location	Near station name	Site ID	Location
Hong Kong International Airport	HKA	(22º18'34",113º55'19")	Hong Kong Observatory	HKO	(22º18'07",114º10'27")
King's Park	KP	(22º18'43",114º10'22")	Lau Fau Shan	LFS	(22º28'08",113º59'01")
Ta Kwu Ling	TKL	(22º31'43",114º09'24")			

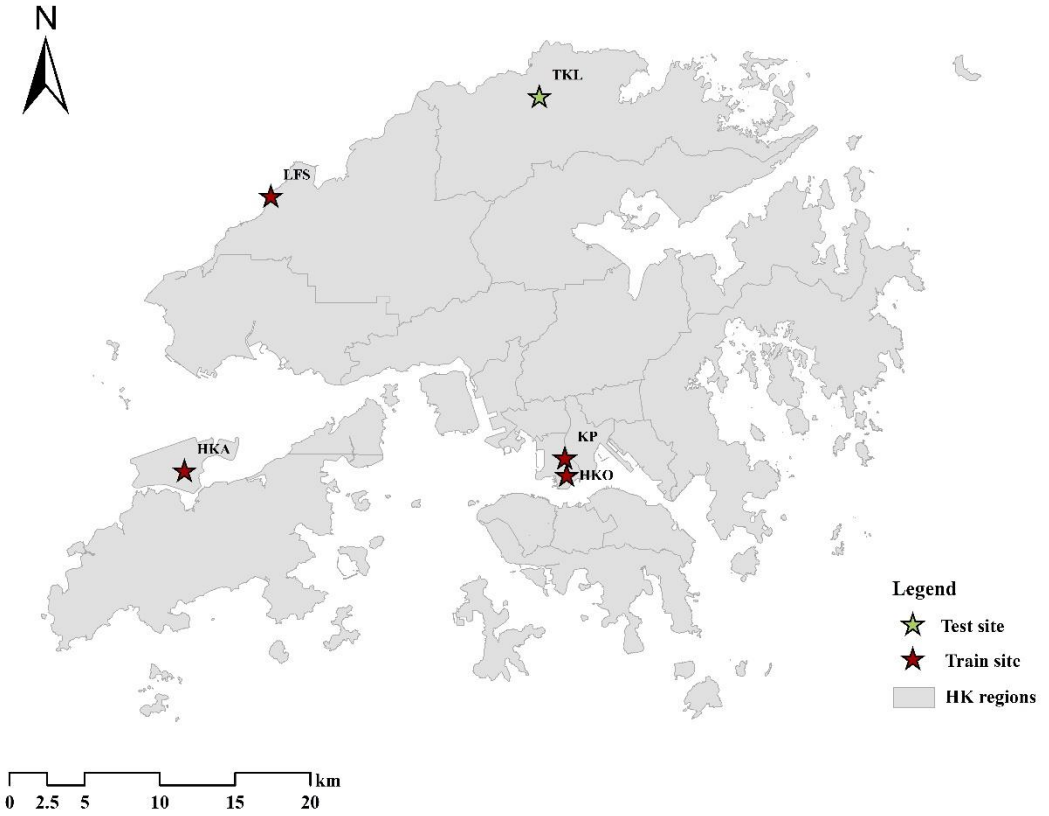


Figure 37 The distribution of five sites in Hong Kong

5.1.2 Data

The amount of the receiving solar irradiation on the building rooftop can be affected by many factors. Specifically, weather factors play a significant role in estimating solar irradiation. For example, Liao et al. (Liao et al., 2023) suggest that meteorological conditions can affect land surface solar irradiation to some extent. Furthermore, many studies (Schunder et al., 2020; Boccalatte et al., 2022; Huang et al., 2022; Liao et al., 2023) suggest that complex shadow conditions from surrounding buildings and terrain and urban morphology are the main limiting factors to decrease rooftop solar irradiation. Additionally, DSM is an important input for calculating solar radiation using the physical model mentioned in Appendix 2. Therefore, this study used DSM, shadow from buildings and terrain, morphological data, and mereological data (i.e., hourly prevailing direction in degree, mean sea level pressure, hourly mean wind speed, relative humidity, and temperature) to generate the dataset for training the TFT model. DSM and the morphological data used in this Chapter are the same as the Chapter 4. Other data are detailed as follows. The methods for calculating the shadow from the buildings and terrain are the same as that in Chapter 4. However, Since the shadow shows the differences in different seasons, we calculated the monthly shadow from the buildings and terrain. Specifically, we calculated the shadow from the buildings on the 15th of each month from 7 am to 5 pm and the calculation values of the middle day of the month are assigned as the shadow values for other days of that month. The method of assigning values for shadows from the terrain is consistent with the method for shadows from the buildings. Furthermore, meteorological data were obtained from the Hong Kong Observatory, including hourly prevailing direction in degree, mean sea level pressure, hourly mean wind speed, relative humidity, and temperature. The shadows from buildings and terrain and the meteorological data are from January 1st, 2019 to December 31st, 2020 at the interval of one hour from 7 am to 5 pm.

5.2 Methodology

In Chapter 4, the author proposed a data-model dual-driven method to estimate the annual rooftop solar irradiation, which only investigates the impact of the static parameters on the annual rooftop solar irradiation. In this chapter, we also employ the data-model dual-driven

framework to estimate the hourly rooftop solar irradiation at the building scale. This framework contains a model-driven method and a data-driven model. The model-driven method employs the upward-looking hemispherical viewshed algorithm proposed by Rich et al. (Rich et al., 1994), and the mechanism of this algorithm is detailed in Appendix 2. The aim of the model-driven method is to calculate the hourly rooftop solar irradiance by the function of Solar Radiation Tool in ArcGIS, and the calculation values are regarded as the ground truth for modeling using the data-driven model. The data-driven model uses the DGTFT method to estimate the hourly rooftop solar irradiation at 1m resolution, and it uses the solar irradiation values calculated by the model-driven method to train the DGTFT model. The specific methodology is detailed as follows.

5.2.1 The physical model for estimation of rooftop solar irradiation

The hemispherical viewshed algorithm explained in Appendix 2 was used to calculate the solar radiation over a geographic area or for specified point locations (longitude, latitude). To obtain the precise rooftop solar irradiation at 1m resolution, we convert the building polygons into the building points, and we perform the hemispherical viewshed algorithm to calculate the solar irradiation of these points.

This algorithm takes location, elevation, slope, aspect, and atmospheric transmission as the most relevant inputs. Since this study just investigates the rooftop solar potential on flat surfaces, the constant values of zero are used for slope and aspect. Furthermore, there are two crucial parameters, i.e. diffuse proportion (D) and transmittivity (T), which denote the proportion of global normal radiation flux that is diffuse and the fraction of radiation that passes through the atmosphere (averaged over all wavelengths), respectively. The author proposed improving the accuracy of the modeling results by fine-tuning the input parameters of the model to co-match the direct and diffuse solar irradiations obtained from the Hong Kong Observatory (HKO). Therefore, the hourly global, direct, and diffuse irradiation data measured at the King's Park and Kau Sai Chau stations were obtained for parameter optimization. The total amount of radiation calculated for a given location is given as global radiation in the (energy) units of

Wh/m^2. Since the research objective in this study focuses on the building rooftop, we use the DSM in Hong Kong as the input raster data for calculating rooftop solar irradiation, which contains the elevation information. Other parameters used for calculating the rooftop solar irradiation by the physical model are shown in Table 17.

Table 17 The parameters used for calculating the rooftop solar irradiation by the physical model

Parameter name	value
Skysize	400
DayInterval	1
HourInterval	1
ZFactor	1
CalcDirections	32
ZenithDivisions	16
AzimuthDivisions	16

5.2.2 Dual-gate Temporal Fusion Transformer for estimating solar irradiation

The aim of this study is to investigate the non-linear relationship between the shadow from the buildings and terrain, the morphological parameters, the meteorological parameters, and the rooftop solar irradiation. These parameters can be classified into static parameters and time-varying parameters. However, it is difficult for traditional machine learning methods to extract the time series characters from these data. Therefore, this study employs the temporal fusion transformer method to integrate the static features from the morphological data with the dynamic features from the shadow of buildings and terrain and the meteorological data for estimating hourly rooftop solar irradiation at the building scale.

In this study, Let I represents unique entities in rooftop solar irradiation. Each entity i consists of static metadata s_i, time series inputs $X_{i,t}$, and solar targets $y_{i,t}$ at time step t, $t \in [0,\ T_i]$. In the case of this study, s_i contains the value of DSM and the morphological data, $X_{i,t}$

contains the shadow from the buildings and terrain and the meteorological data, and the $y_{i,t}$ is the hourly rooftop solar irradiation. As the mentioned in Chapter 3, $X_{i,t}$ can be classified into two categories, $X_{i,t} = [z_{i,t}^T,\ x_{i,t}^T]^T$. Past inputs $z_{i,t}$ consists of relative time index, hill shadow, building shadow, hourly prevailing direction in degree, mean sea level pressure, hourly mean wind speed, relative humidity, temperature, and rooftop solar irradiation, and know future inputs $x_{i,t}$ contains relative time index, hill shadow, building shadow, hourly prevailing direction in degree, mean sea level pressure, hourly mean wind speed, relative humidity, and temperature. The prediction function is defined as follows:

$$\hat{y}_i(t,\tau) = f(\tau, y_{i,t-k:t}, z_{i,t-k:t,} x_{i,t-k:t+\tau}, s_i) \tag{5.2.2}$$

Where $y_{i,t-k:t} = \{y_{i,t-k}, \dots, y_{i,t}\}$ denotes targets of hourly rooftop solar irradiation until the time t, and τ represents the prediction time point.

5.2.3 The data-model dual-driven mechanism

To obtain the accurate and fast estimation results of rooftop solar irradiation at a fine resolution, we employ the data-model dual-driven mechanism. Specifically, the data-driven model and the model-driven method are loose coupling. The whole processing of this mechanism is detailed as follows.

i) Selection sites for generating datasets. We selected five representative weather stations and searched for the nearest building polygon. Then, we converted these building polygons into the five points. We performed the estimation of solar irradiation on these five sites.

ii) Calculation of rooftop solar irradiation by the model-driven method. We use the hemispherical viewshed algorithm to calculate the solar irradiation on five sites from January 1st,2019 to December 31st, 2020 at the interval of one hour. Since only solar radiation in the daytime has the research values, we just calculate the

values from 7 am to 5 pm within a day. The validation of the model-driven method has been demonstrated in Chapter 4, and the results suggest the accuracy of the results of the model-driven method enables the calculation values to as the ground truth for modeling the data-driven model.

iii) Generation of the datasets for the data-driven model. The formats of the shadow from the buildings and terrain, parameterized urban morphology, and DSM are Tiff images. So, we extracted the above parameter values from the corresponding sites. The solar irradiation calculated from the model-driven method was used as the ground truth for training the DGTFT model. All data was organized in chronological order. These data were labeled as the static data, time-varying data, and the target data.

iv) Division of the datasets. We divided the entire time-series dataset into a series of samples, and each sample is a subsequence of a full-time series. The subsequence consists of encoder and decoder/prediction time points for a given time series. In the case of our study, the length of the encoder is 66 hours (there are 11 hours time points within a day, here we use the data of six days), and the length of the decoder is 11 hours. We use the dataset from four sites as the training and validation datasets and the dataset from the TKL site as the test dataset.

v) TFT modeling. After generating the datasets, we train the DGTFT model using the train and validation datasets and use the test dataset to evaluate the accuracy. And then, we employ the trained model to estimate the rooftop solar irradiation on individual buildings.

5.2.4 Evaluation metrics

In this Chapter, four indicators were used to evaluate the estimation performance, including R^2,

MAE, RMSE, and Mean Absolute Percentage Deviation (MAPE). The formulas for R^2, MAE, and RMSE are listed in Chapter 3.1.6, and the formula of MAPE is as follows:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{|y_i - \hat{y}_i|}{|y_i|} \quad (5.2.4)$$

5.3 Results

5.3.1 Data cleaning

To confirm the quality and dependability of the data, data cleaning is performed on all the data to filter the outliers, including missing values and non-numeric values. We counted the number of outliers as a percentage of the total number of data in five datasets, and the statistical results are shown in Table 18. From this table, we found that the ratios of the outliers in all datasets are low. However, solar irradiation is the classic time-series data, and the integrity of the time series of data can help the DGTFT model greatly capture the time series features from the dataset. Therefore, we use the zero values to replace the values of these outliers to ensure the integrity of the time series of data.

Table 18 The ratio of the outliers in five datasets

Dataset name	Outliers ratio	Dataset name	Outliers ratio
HKA	0%	HKO	0.012%
KP	0.90%	LFS	0.11%
TKL	0.79%		

5.3.2 The result from the physical model

In this study, the calculation results of the physical model are the ground truth for generating the datasets for training and testing the TFT deep learning model. Our research period covers from January 1st, 2019 to December 31st, 2020 at the interval of one hour from 7 am to 5 pm. So the total time point of each dataset is 8041. We used the hemispherical viewshed algorithm to calculate the rooftop solar irradiation on the selected five sites and record the computation time. The experiment was performed on a desktop with Intel (R) Core (TM) i7-9700K CPU and 32 GB memory. The results of the computation time of each site point using the physical

model are shown in Table 19. The computation time for one site point is about seven hours, which suggests that the computation speed is low and this leads to huge time costs when the physical model is applied for calculating the large amount of rooftop solar irradiation.

Table 19 The computation time of each site point using the physical model

Site name	Time (h)	Site name	Time (h)
HKA	7.16	HKO	6.62
KP	7.26	LFS	6.14
TKL	7.29		

5.3.3 DGTFT model results

We employ the datasets generated from HKA, HKO, KP, and LFS sites as the training datasets, and the dataset from the TKL site is the test dataset. The four indicators are used to evaluate the accuracy, including R^2, MAE, RMSE, and MAPE. The results are shown in Table 20. These results indicate that the DGTFT model displays high-accuracy estimation performance, with R^2=0.90, MAE=26.90 (MJ/m^2), RMSE=32.39 (MJ/m^2), and MAPE=18%. Additionally, we also calculate the computation time of training the model and testing the model. Training the DGTFT model costs about 0.20 hours and using the trained model to estimate the test dataset just costs 0.13 hours. Compared to the results of the physical model, the time cost of the DGTFT model is far less than that of the physical model. Specifically, the physical model requires 7.29 hours to calculate the hourly rooftop solar irradiation covering two years using TKL dataset, while the DGTFT model just needs 0.13 hours to complete the estimation task under the equivalent computation. These results demonstrate that the DGTFT model can provide a fast and accurate estimation result for hourly rooftop solar irradiation.

Table 20 The test accuracy results

Model	R^2	MAE (MJ/m^2)	RMSE (MJ/m^2)	MAPE
DGTFT	0.90	26.90	32.39	18%

One of the innovations of the DGTFT model is to calculate the importance of the static variables and time-varying variables. From Figure 38, it is noticed that the hill shadow accounts for the

most importance in the Decoder, at about 80 %, following by the mean wind speed, at about 8%. Other variables account for a very small percentage. These results suggest that hill shadow plays a significant role in helping the Decoder layer to predict rooftop solar irradiation. Furthermore, we also found that the relative time index accounts for the most importance at around 35%, and the hill shadow and building shadow account for about 17% and 16% in the Encoder layer from Figure 39. The importance of the rooftop solar irradiation, prevailing direction in degree, mean sea level pressure, mean wind speed, relative humidity, and temperature decrease in percentage from about 9% to 3%. These results suggest that the relative time index plays an important role in estimating the time-series solar data, and the shadow from the buildings and terrain are two important factors to affect the amount of rooftop solar irradiation. Furthermore, the DGTFT model calculates the importance of the static variables, as shown in Table 21. It is noted that the most important variable is solar radiation$_{center}$, at 26.77% of importance, while the other remaining variables account for far less importance. This may suggest that the morphological variables are not important factors in estimating hourly rooftop solar irradiation, because the importance of all morphological variables is far lower than that of the solar radiation$_{center}$.

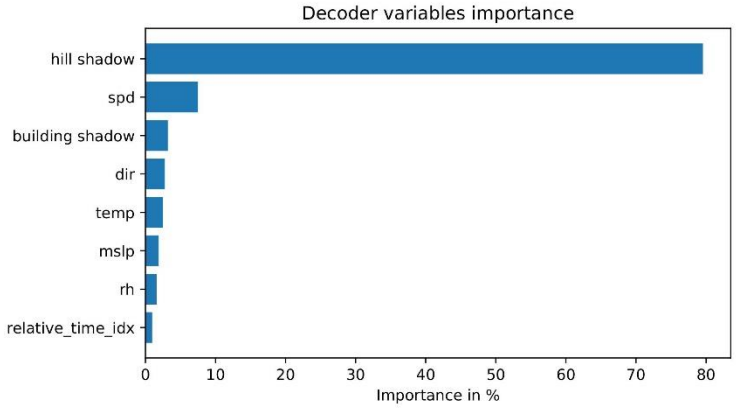


Figure 38 The importance of the variables in Decoder

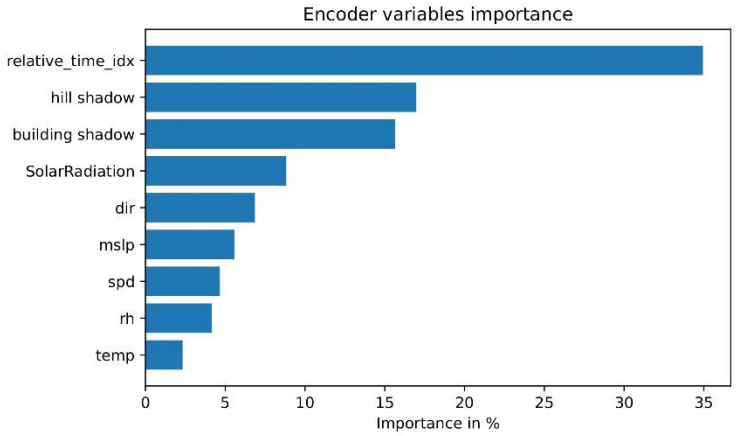


Figure 39 The importance of the variables in Encoder

Table 21 Importance of the static variables in the DGTFT model

Parameter	Importance (%)	Parameter	Importance (%)	Parameter	Importance (%)	Parameter	Importance (%)
SolarRadiation_scale	26.77	t_lal	1.74	mean_heigh	1.23	squareness	0.66
CAR	8.56	hd_n	1.68	squ_comp	1.21	hd	0.62
mean_ta	6.62	orientatio	1.66	mean_FAR	1.21	Rug	0.6
FAR	5.03	mean_volum	1.64	perimeter	1.14	rect	0.48
neighbours	4.84	mean_inter	1.56	elongation	1.01	floor_area	0.46
adjacency	3.82	station_x	1.55	encoder_length	0.95	vfr	0.44
SolarRadiation_center	3.31	alignment	1.48	Dsm	0.92	mean_CAR	0.32
eri	3.13	comp	1.36	fractal	0.89	roof_heigh	0.23
hd_p	2.98	volume	1.34	height	0.85	lal	0.1
hw	1.92	swr	1.3	mean_fa	0.66	squareness	0.66
svf	1.8	shape_inde	1.27	mean_area	0.66	hd	0.62

5.3.4 Estimation hourly map

The author utilized the trained model to generate hourly rooftop solar irradiation maps at a 1-m spatial resolution for the building in Hong Kong on October 1st, 2020, spanning from 7 am to 6 pm. The solar maps generated by the trained model were compared with the results calculated by the physical model. The comparison result is presented in Figure 40.

In general, solar irradiation progressively increases from 7 am to 12 pm, reaching its peak at noon. Subsequently, it gradually decreases, and by 6 pm, solar irradiation is at its lowest. Additionally, from Figure 40, the maps generated by the DGTFT model are diametrically similar to those generated by the physical model. To greatly evaluate the estimation performance of the DGTFT model, we calculated the mean absolute hourly error of the results between the proposed method and the physical model from 6 am to 5 pm. The result is shown in Figure 41. It is noted that the mean absolute hourly error is relatively small, with a maximum value of 59.1 MJ/m². This suggests that the DGTFT model provides highly accurate estimations for rooftop solar irradiation. These hourly results offer a precise and reliable reference for optimizing the dispatch and management of solar power systems.

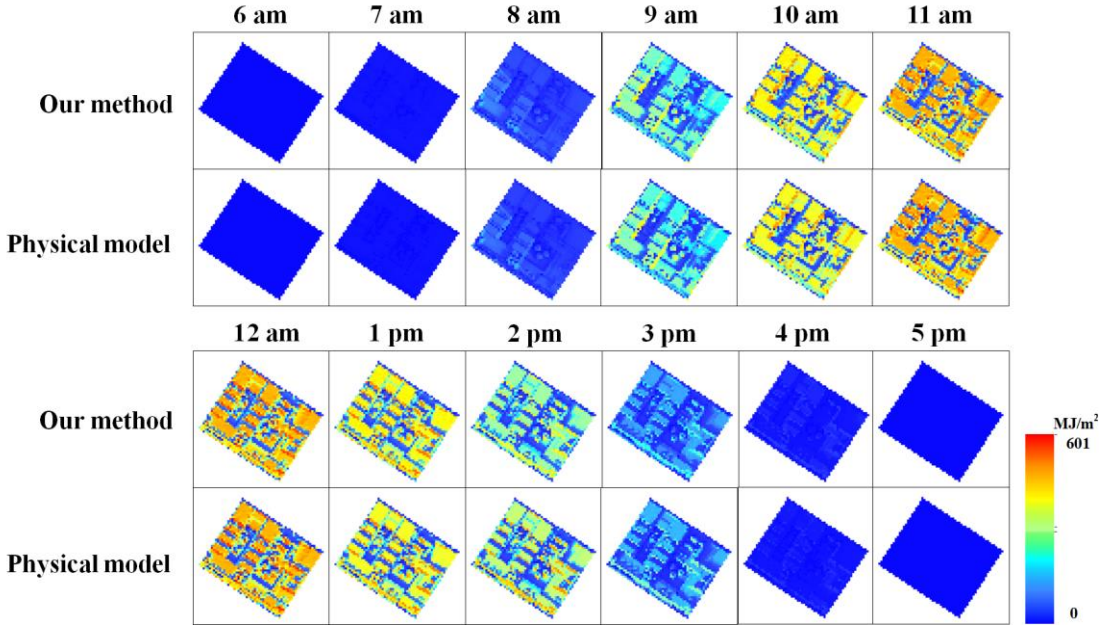


Figure 40 The comparison of hourly estimated rooftop solar irradiation from 6 am to 5 pm between the proposed method and the physical model.

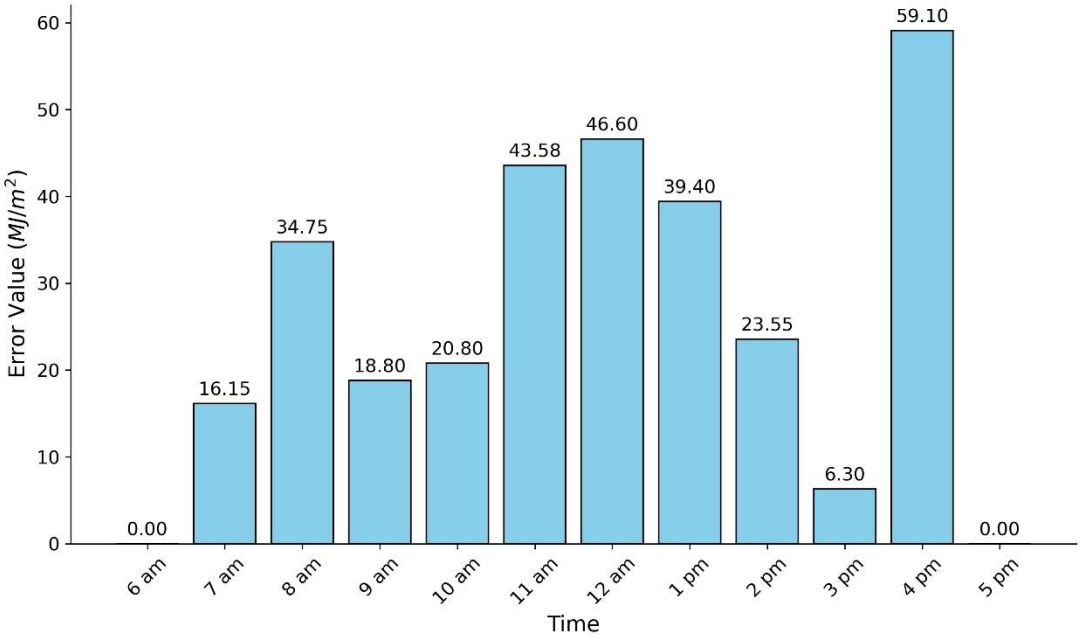


Figure 41 The mean absolute hourly error of the results between the proposed method and the physical model from 6 am to 5 pm.

5.4 Discussion and Conclusion

This chapter proposed a data-model dual-driven loosely coupled approach for fast and accurate prediction of hourly rooftop solar irradiation at the building scale. Firstly, the model-driven method (physical model) was used for calculating the hourly rooftop solar irradiation from January 1st, 2019 to December 31st, 2020 at the interval of one hour from 7 am to 5 pm. Since the estimation accuracy of the physical model was evaluated and demonstrated in Chapter 4, this study also uses the calculation values of rooftop solar irradiation as the ground truth for training the data-driven model. After that, this study parameterized the relative parameters, including hourly hill shadow and building shadow, morphological data, and integrated DSM, meteorological data, and rooftop solar irradiation calculated by the model-driven method with these parameterized data to generate the datasets. The author selected the four sites for training the data-driven model (DGTFT model), and one site for testing. The results show the data-model dual-driven approach can provide a fast and accurate estimation result of hourly rooftop solar irradiation, with R^2=0.90, MAE=26.90 (MJ/m^2), RMSE=32.39 (MJ/m^2), and MAPE=0.18. And the model-driven method takes 56 times longer than the data-driven model. The proposed dual-driven approach integrates the merits of highly accurate estimation using the model-driven method with the merits of fast computation capability using the DGTFT data-

driven model. Furthermore, the DGTFT model calculated the importance of static variables and time-varying variables in the static layer, Encoder layer, and Decoder layer. The results suggest that shadow from buildings and terrain can greatly affect rooftop solar irradiation, and the relative time index is an important factor for modeling time series data using DGTFT. Additionally, the most significant static variable is solar radiation$_{center}$, while morphological variables and DSM seems to make fewer contributions to training the DGTFT model. It is possible that this study just employs four sites for training the model and the DGTFT model is difficult to extract the static features from small static data samples.

In conclusion, this study provides a fast and high-accuracy estimation method for hourly rooftop solar irradiation based on the data-model dual-driven approach. The hourly result estimated by this method can provide a reliable reference for the government to make decisions for solar PV installation on rooftops and give an accurate suggestion for how to realize 2050 carbon neutrality.

Chapter 6 Conclusions and Recommendations

In this Chapter, I discuss the results of the land surface solar potential and the rooftop solar potential. I also conclude the machine learning-based methods used for the estimation of solar potential from the perspective of the macro-geographical level to the spectator-building level. Finally, I present several recommendations for the current limitations and how to extend the existing context into broad practice.

6.1 Discussion

To realize the fast and highly accurate estimation of solar potential at a high spatio-temporal resolution, several Geo-AI methods were used for estimating hourly/daily land surface solar irradiation in Australia, China, and Japan, and a data and model dual-driven method was employed for estimating annual and hourly rooftop solar irradiation in Hong Kong. The key findings are as follows:

1) Among the traditional machine learning methods, the GBM model is the optimal model to estimate land surface solar irradiation over large-scale regions. However, these traditional methods cannot consider the impact of the geographic variation on solar irradiation. therefore, an interpretable DGTFT deep learning method was proposed and applied for solving this issue, the results suggest this method has good estimation performance and transferability.

2) A data and model dual-driven method can greatly integrate the merits of fast computation of the data-driven method and the high-accuracy results of the model-driven method, and the results suggest that this method can offer a fast and accurate estimation result for rooftop solar potential.

3) The DGTFT method is the optimal method to estimate hourly rooftop solar irradiation because it not only can consider the impact of static geographic variation and dynamic

factors on rooftop solar irradiation but also overcomes the limitation of the "black-box" nature to give the interpretability of all inputs. It also has the capability of fast computation and highly accurate results.

6.2 Limitations

There are some limitations in this thesis. Firstly, to test the transferability of the methods used for estimating land surface solar irradiation, three countries, including Australia, China, and Japan, were selected as the research regions. However, since some inputs for estimating rooftop solar irradiation, such as building polygon and DSM, are not easy to obtain, the research region just focuses on Hong Kong. This increases the difficulty in testing the transferability of this method in other cities. Secondly, some data used for estimating rooftop solar irradiation, such as building polygon and DSM, are historical data, and this would lead to the estimated bias because urban buildings change slowly over time. Finally, this thesis just estimates rooftop solar irradiation on the whole rooftop. However, not all areas of the rooftops are suitable for installing PV panels. Therefore, accurately estimating the availability of rooftop areas for solar irradiation is further work.

6.3 Conclusions

Accuracy evaluation of solar potential is greatly significant in providing a reliable and reasonable reference for urban designers and the government to effectively generate renewable energy and mitigate energy-related emissions. The hierarchy for rooftop photovoltaic energy comprises three levels: (i) the physical potential, which encompasses the total amount of energy received from the Sun in the area of study; (ii) the geographic potential, which restricts the locations where this energy can be captured; and (iii) the technical potential, which further takes into account the technical characteristics (including performance) of the equipment used for transforming the resource into electrical energy (Izquierdo wt al., 2008). This thesis focuses on the physical potential and the geographic potential. In this thesis, I propose a framework to estimate the land surface solar potential and rooftop solar potential. Specifically, firstly, I propose a simple and effective method for the estimation of land surface solar irradiation based

on machine learning models using meteorological data, Himawari-8 satellite cloud and aerosol products, and solar observation data in Australia and China. The estimation of solar irradiation based on four machine learning models, i.e., RF, SVR, MLP, and GBM, is effective and reliable, and GBM has achieved the best performance in terms of accuracy and computational efficiency. The estimation of seasonal and annual solar irradiation at nationwide levels is useful for planning solar-related applications. Furthermore, to solve the "black box" problem of the traditional machine learning methods and enhance the capabilities of extracting information from spatio-temporal sequence data, an interpretable DGTFT deep learning method was designed to improve the estimation performance for land surface solar irradiation. Thirdly, from the perspective of geographic potential, the author investigated the impact of multi-source data on the amount of annual solar irradiation on building rooftops and constructed a machine learning model to fast and accurately estimate rooftop annual solar irradiation with a 1m resolution. In addition, the author also analyzed the mean annual solar irradiation received by the rooftops as a function of rooftop slope and aspect. The results are particularly useful for designers, investors, owners, and stockholders in providing quantitative information on the effects of roof slope and aspect on the PV solar potential at the design stage. In the final part, I used a data-model dual-driven framework to estimate hourly solar irradiation on rooftops with a 1m resolution. Its results with high spatial and temporal resolution can be used for precisely calculating the specific electrical energy generated by the PV panels on rooftops.

This thesis provides the following insights on the estimation of solar potential:

1) Among the traditional machine learning methods, GBM achieved the highest accuracy with R^2 at all stations, followed by RF, SVR, and MLP. It suggests that the proposed method can provide an accurate and reliable estimation of land surface solar irradiation, compared with the theoretical solar irradiation without the obstacle of the atmosphere. While the RF model outperformed GBRT and AdaBoost for estimating annual rooftop solar irradiation, with R^2=0.77 and MAE=22.83kWh/m^2/year. Additionally, the time for training and prediction of rooftop solar irradiation is within 13 hours, achieving a 99.32% reduction in time compared to the physical-based hemispherical viewshed algorithm. These results suggest that the proposed method can provide an accurate and

fast estimation of annual rooftop solar irradiation for large datasets.

2) Meteorological parameters play a significant role in estimating spatio-temporal solar irradiation, including hourly/daily land surface solar irradiation and hourly rooftop solar irradiation. These parameters can greatly help increase the estimation accuracy.

3) The proposed DGTFT deep learning method is effective for large geographical regions and can be used worldwide when similar datasets are obtained. Also, the DGTFT model provides a fast and accurate estimation result of hourly rooftop solar irradiation, with R^2=0.90, MAE=26.90 (MJ/m^2), RMSE=32.39 (MJ/m^2), and MAPE=0.18. The physical-based hemispherical viewshed algorithm takes 56 times longer than the DGTFT model. Hourly rooftop solar irradiation at a fine resolution can aid in the accurate assessment of solar energy resources at specific locations, which is crucial for the planning and design of solar energy projects to maximize renewable energy utilization.

4) Compared to the solar resources in China and Japan, Australia has the most abundant physical potential, suggesting this country is the optimal region to develop and promote the solar industry. In addition, Hong Kong has abundant rooftop solar potential, with annual power generation of 3.3 TWh.

6.4 Recommendations for the Future Work

Although this thesis has proposed a framework to estimate the land surface solar irradiation and rooftop solar irradiation at a fine spatio-temporal resolution, this study also has some limitations:

5) This paper only considers the maximum potential of solar photovoltaic electricity generation, and in practical applications, there are still numerous factors to be taken into account. These include the economic feasibility of installing solar panels, the angle of installation on building rooftops, vegetation cover, building materials, local

atmospheric conditions, and other factors. Therefore, it is recommended that future research further expands consideration of the factors related to the application of urban rooftop photovoltaics to achieve energy savings and emissions reduction in urban areas.

6) The estimation of solar geographical potential includes assessing the potential of building rooftops and facades. This study specifically focuses on rooftop potential but can be extended in the future to encompass a comprehensive estimation of the entire rooftops and facades of buildings. Additionally, not all rooftop areas are suitable for solar panel installation. Therefore, estimating the area of rooftops suitable for solar panel installation is crucial for a detailed evaluation of rooftop solar potential. This estimation can provide valuable insights for optimizing recommendations for solar panel installation. In future research, remote sensing technology and deep learning methods can be employed to more accurately estimate the available area on rooftops for solar utilization.

Appendix 1

Name	Description	Symbol	Category	Equation
Building height	Building height	H	D	-
Building area	Building footprint area	A	D	-
Building volume	Building volume	V	D	-
Building perimeter	Sum of lengths of the building exterior walls	P	D	-
Building longest axis length	Diameter of the minimal circumscribed circle around the building footprint	LAL	D	-
Building volume to façade ratio	Ratio between building volume and the total area of façades	VFR	D	$vfr = \frac{volume}{perimeter \cdot height}$
Building fractal dimension	Statistical index of the complexity of a geometry	Fra	D	$fractal = \frac{2\log(perimeter/2)}{\log(area)}$
Building circular compactness	Index of the similarity of a shape with a circle. It is based on the area of the minimal enclosing circle (Ac)	Com	S	$comp = \frac{area}{Ac}$
Building square compactness	Measure of the compactness of the building footprint	Squ_{com}	S	$squ_{comp} = (\frac{4\sqrt{area}}{perimeter})^2$
Building squareness	Mean deviation μ of each i corner of the building from 90°. Ncor is the number of corners	Squ	S	$squareness = \frac{\sum_{i=1}^{Ncor} \mu_i}{Ncor}$
Building Rectangularity	Index of the similarity of a	Rec	S	$rect = \frac{area}{AMBR}$

	shape with a rectangle. It is based on the area of the minimal rotated bounding rectangle of the building (AMBR)			
Building shape index	Shape index of the building footprint	Shp_{idx}	S	$shape_{index} = \frac{\sqrt{\frac{area}{\pi}}}{0.5 \cdot lal}$
Building equivalent rectangular index	Measure of shape complexity based on the area of the minimal rotated bounding rectangle of a building (AMBR) and its perimeter (PMBR)	ERI	S	$eri = \sqrt{\frac{area}{AMBR}} \cdot \frac{PMBR}{perimeter}$
Building elongation	Measure of the deviation of the building shape from a square based on the length of the minimal rotated bounding rectangle of a building (LMBR) and its width (lMBR)	Elg	S	$elongation = \frac{LMBR}{lMBR}$
Floor area ratio	Ratio between the building total floor area and the area of the related tessellation cell	Flr_{area}	I	$floor_{area} = \frac{area}{t\_area}$
Shared walls ratio of adjacent buildings	Ratio between the length of the perimeter shared with adjacent buildings (Pshared) and the building perimeter	SWR	SD	$swr = \frac{Pshared}{perimeter}$

Building orientation	Building orientation	Ort	D	-
Alignment	Mean deviation of solar orientation (devsol) of neighboring buildings	Ali	SD	$alignment = \frac{\sum_{j \in neigh} devsol(j)}{Nneigh}$
Building adjacency	Ratio between the number of joined adjacent structures (Nneigh,join) and the number of neighboring buildings (Nneigh)	Adj	SD	$adjacency = \frac{Nneigh\_join}{Nneigh}$
Mean inter-building distance	Mean distance between the building and the adjacent buildings	IBD_{mean}	SD	$mean_{in} = \frac{1}{Nneigh} \sum_{j \in neigh} d(j)$
Average building area	Mean footprint area of building neighboring constructions	A_{mean}	SD	$mean_{area} = \frac{1}{Nneigh} \sum_{j \in neigh} area(j)$
Average building height	Mean height of building neighboring constructions	H_{mean}	SD	$mean_{height} = \frac{1}{Nneigh} \sum_{j \in neigh} height(j)$
Average building volume	Mean volume of building neighboring constructions	V_{mean}	SD	$mean_{volume} = \frac{1}{Nneigh} \sum_{j \in neigh} volume(j)$
Average building total floor area	Mean total floor area of building neighboring constructions	TFA_{mean}	SD	$mean_{fa} = \frac{1}{Nneigh} \sum_{j \in neigh} floor\_area(j)$
Average Height to Width ratio	Mean ratio between building height and width of building neighboring constructions	HW	SD	$hw = \frac{1}{Nneigh} \sum_{j \in neigh} \frac{H}{d(j)}$

Distance–weighted average height difference	Mean height difference with distance weighted between the reference building and its neighboring buildings	HD	SD	$hd = \frac{\sum_{j \in neigh}(H(j) - H) \cdot w(j)}{\sum_{j \in neigh} w(j)}$
Average neighborhood shading angle	Mean shading angle between the reference building and its neighboring buildings	Shd_{mean}	SD	$shade_a = \arctan\left(\frac{1}{Nneigh}\sum_{j \in neigh}\frac{H(j) - H}{d(j)}\right)$
Positive distance–weighted average height difference	Mean height difference with distance weighted between the reference building and its neighboring buildings (H(j)>H)	HD_p	SD	$hd\_p = \frac{\sum_{j \in neigh}(H(j) - H) \cdot w(j)}{\sum_{j \in neigh} w(j)}$
Negative distance–weighted average height difference	Mean height difference with distance weighted between the reference building and its neighboring buildings (H(j)<H)	HD_n	SD	$hd\_n = \frac{\sum_{j \in neigh}(H(j) - H) \cdot w(j)}{\sum_{j \in neigh} w(j)}$
Positive average neighborhood shading angle	Mean shading angle between the reference building and its neighboring buildings (H(j)>H)	Shd_{ap}	SD	$shade_{ap} = \arctan\left(\frac{1}{Nneigh}\sum_{j \in neigh}\frac{H(j) - H}{d(j)}\right)$
Negative average neighborhood shading angle	Mean shading angle between the reference building and its neighboring	Shd_{an}	SD	$shade_{an} = \arctan\left(\frac{1}{Nneigh}\sum_{j \in neigh}\frac{H(j) - H}{d(j)}\right)$

	buildings (H(j)<H)			
Rugosity	Ratio between the building volume and the area of the related tessellation cell	Rug	I	$Rug = \frac{volume}{area\_t}$
Floor area	Floor area of each object based on height and area	FA	S	$FA = \frac{height \cdot area}{3}$
Coverage area ratio	Ratio between the building footprint area and the area of the related tessellation cell	CAR	I	$CAR = \frac{area}{area\_t}$
Mean coverage area ratio	Mean coverage area ratio of the neighboring tessellation cells	CAR_{mean}	SD	$mean\_CAR = \frac{1}{Nneigh}\sum_{j \in neigh} CAR(j)$
Mean floor area ratio	Mean floor area ratio of the neighboring tessellation cells	FAR_{mean}	SD	$mean\_FAR = \frac{1}{Nneigh}\sum_{j \in neigh} FAR(j)$
Sky view factor	Sky view factor	SVF	SD	
Number of neighbors	Number of neighbors	N_{neigh}	SD	-
Tessellation longest axis length	Diameter of the minimal circumscribed circle around the tessellation cell	LAL_{tess}	D	-
Average tessellation area	Mean tessellation area of building neighboring tessellation cells	A_{tess}	SD	$mean\_ta = \frac{1}{Nneigh}\sum_{j \in neigh} area\_t(j)$

D=dimension, S=shape, I=intensity, SD=spatial distribution

Appendix 2

1 Upward-looking hemispherical viewshed algorithm

The estimates of rooftop solar irradiation calculated by the *Solar Analyst Tool* in *ArcMap* are as the ground truth in our machine learning model training. The *Solar Analyst Tool* is based on methods from the hemispherical viewshed algorithm developed by Fu and Rich (Fu and Rich, 2022). The specific process of the upward-looking hemispherical viewshed algorithm is as follows.

1.1 Viewshed calculation

A viewshed is the angular distribution of sky visibility versus obstruction and represents the proportion of the obstructed sky in a specific location on a DEM, which is similar to the view from upward-looking hemispherical photographs. Viewsheds are calculated by searching a specified set of directions around an interesting location on DEM in each direction and determining the maximum angle of sky obstruction. For the unsearched directions, the interpolation method is used to calculate the horizon angles. The horizon angles are projected into a two-dimensional (2D) grid using an equiangular hemispherical projection. A value with visible versus obstructed sky directions is assigned to each corresponding grid unit. The grid cell location (i.e., row and column) represents a zenith angle θ and an azimuth angle α on the hemisphere of directions.

1.2 Sunmap calculation

After generating a viewshed for a specific location on a DEM, a sunmap is created to represent the amount of direct solar radiation from each sky direction in the same 2D grid system. The sunmap consists of specified suntracks, and it represents the apparent position of the sun as it varies through time. Zenith and azimuth angles are used to represent the position of the sun, and they are calculated based on latitude, day of year, and time of day using standard astronomical formulae (Rich et al., 1994). Zenith and azimuth angles are projected to 2D grids with the same resolution as the viewsheds. Two sunmaps are created, namely, a sunmap for

winter solstice to summer solstice, and sunmap for summer solstice to the winter solstice. For each sky sector of the sunmap, the associated time duration and the position of the sun are calculated, and each sector is assigned a unique identification number.

1.3 Skymap calculation

To achieve the skymap calculation, the whole sky is divided into a series of sky sectors defined by zenith and azimuth divisions. The skymap is used in the final solar radiation calculation to estimate diffuse solar radiation. The sky sectors in the skymap are required to be small enough that the centroid zenith and azimuth angles can reasonably represent the direction of the sky sector in subsequent calculations. The skymap is also projected into the 2D grid for the final solar radiation calculation.

1.4 Overlay of viewsheds with sunmaps and skymaps

After creating sunmap and skymap, two maps are overlayed to enable calculation of the direct and diffuse solar radiation received from each sky direction. For gap fraction in the skymap or sunmap sector, it is calculated by dividing the number of unobstructed units by the total of units in that sector.

1.5 Global solar radiation calculation

Since reflection radiation accounts of the small proportion of the global solar radiation, global radiation G_R is calculated as the sum of direct and diffuse radiation of all sectors in the sunmap and skymap. The formula is as follows:

$$G_R = D_R + F_R \tag{1}$$

where D_R denotes the total direct solar radiation for all sunmap sectors, F_R represents the total diffuse solar radiation for all skymap sectors. The formula is as follows:

$$D_R = \sum D_{\theta,\alpha} \tag{2}$$

where $D_{\theta,\alpha}$ denotes the direct insolation from the sunmap sector with a centroid at zenith angle θ and azimuth angle α. The formula for calculation of $D_{\theta,\alpha}$ is as follows:

$$D_{\theta,\alpha} = S_{const} \times \tau^{m(\theta)} \times SunDur_{\theta,\alpha} \times SunGap_{\theta,\alpha} \times \cos\left(AngIn1_{\theta,\alpha}\right) \quad (3)$$

where S_{const} denotes a solar constant and its range is from 1338 to 1368 WM^{-2}, τ is transmittivity of the atmosphere for the shortest path, m (θ) is the relative optical path length, $SunDur_{\theta,\alpha}$the time duration represented by the sky sector, $SunGap_{\theta,\alpha}$ is the gap fraction for the sunmap sector, and $AngIn1_{\theta,\alpha}$ is the angle of incidence between the centroid of the sky sector and the axis normal to the surface. The formula for calculation of F_R is as follows:

$$F_R = R_{glb} \times P_{dif} \times Dur \times SkyGap_{\theta,\alpha} \times Weight_{\theta,\alpha} \times \cos\left(AngIn2_{\theta,\alpha}\right) \quad (4)$$

where R_{glb} is the global normal radiation, P_{dif} is the proportion of global normal radiation flux that is diffused, Dur is the time interval for analysis, $SkyGap_{\theta,\alpha}$ is the gap fraction (proportion of visible sky) for the sky sector, $Weight_{\theta,\alpha}$ is proportion of diffuse radiation originating in a given sky sector relative to all sectors. $AngIn2_{\theta,\alpha}$ is the angle of incidence between the centroid of the sky sector and the intercepting surface.

Reference

Ad Ministration, U. (2020). International Energy Outlook 2020. National Engineer.

Alhamwi A, Medjroubi W, Vogt T, Agert C. (2019). Development of a GIS-based platform for the allocation and optimisation of distributed storage in urban energy systems. Applied Energy; 251:113360.

Allen, R. G. (1997). Self-calibrating method for estimating solar radiation from air temperature. Journal of Hydrologic engineering, 2(2), 56-67.

Amrouche B, L. P. (2014). Artificial neural network based daily local forecasting for global solar radiation. Applied energy, 130, 333-341.

Amrouche, B., & Le Pivert, X. (2014). Artificial neural network based daily local forecasting for global solar radiation. Applied energy, 130, 333-341.

Angstrom, A. (1924). Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. Quarterly Journal of the Royal Meteorological Society, 50, 121-126.

Arriagada, P., Karelovic, B., & Link, O. (2021). Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. Journal of Hydrology, 598, 126454.

Assouline D, Mohajeri N, Scartezzini JL. (2018). Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests. Applied energy, 217, 189-211.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Babbar SM, Lau CY, Thang KF.(2021). Long Term Solar Power Generation Prediction using Adaboost as a Hybrid of Linear and Non-linear Machine Learning Model. International Journal of Advanced Computer Science Applications, 1, 12-11

Bailek, N., Bouchouicha, K., Al-Mostafa, Z., El-Shimy, M., Aoun, N., Slimani, A., & Al-Shehri, S. (2018). A new empirical model for forecasting the diffuse solar radiation over Sahara in the Algerian Big South. Renewable Energy, 117, 530-537.

Baltas NG, Mazidi P, Ma J, de Asis Fernandez F, Rodriguez P. (2018). A comparative analysis of decision trees, support vector machines and artificial neural networks for on-line

transient stability assessment. In 2018 International Conference on Smart Energy Systems and echnologies (SEST), 1, 1-6.

Behrang, M. A., Assareh, E., Ghanbarzadeh, A., & Noghrehabadi, A. R. (2010). The potential of different artificial neural network (ANN) techniques in daily global solar radiation modeling based on meteorological data. Solar Energy, 84(8), 1468-1480.

Benatiallah, D., Bouchouicha, K., Benatiallah, A., Harrouz, A., & Nasri, B. (2019). Forecasting of solar radiation using an empirical model. Algerian Journal of Renewable Energy Sustainable Development, 1(2), 212-219.

Besharat F, D., AA, Faghih AR (2013). Empirical models for estimating global solar radiation: A review and case study. Renewable Sustainable Energy Reviews, 21, 798-821.

Besharat, F., Dehghan, A. A., & Faghih, A. R. (2013). Empirical models for estimating global solar radiation: A review and case study. Renewable and sustainable energy reviews, 21, 798-821.

Bhanujyothi K, Himabindu K, Suryanarayana D. (2014). A Comparative Study of Random Forest & K – Nearest Neighbors on HAR dataset Using Caret. International Journal of Innovative Research Technology, 3, 6-9.

Biazar, S. M., Rahmani, V., Isazadeh, M., Kisi, O., & Dinpashoh, Y. (2020). New input selection procedure for machine learning methods in estimating daily global solar radiation. Arabian Journal of Geosciences, 13(12), 1-17.

Bird, R. E. (1984). A simple, solar spectral model for direct-normal and diffuse horizontal irradiance. Solar energy, 32(4), 461-471.

Bishop, J. K., Rossow, W. B., & Dutton, E. G. (1997). Surface solar irradiance from the international satellite cloud climatology project 1983–1991. Journal of Geophysical Research: Atmospheres, 102(D6), 6883-6910.

Black, J. N. (1956). The distribution of solar radiation over the earth's surface. Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B, 7, 165-189.

Boccalatte A, Thebault M, M´en´ezo C, Ramousse J, Fossa M. (2022). Evaluating the impact of urban morphology on rooftop solar radiation: A new city-scale approach based on Geneva GIS data. Energy Buildings, 260, 111919.

Bristow, K. L., & Campbell, G. S. (1984). On the relationship between incoming solar radiation

and daily maximum and minimum temperature. Agricultural and forest meteorology, 31(2), 159-166.

Buffat R, Grassi S, Raubal M. (2018). A scalable method for estimating rooftop solar irradiation potential over large regions. Applied energy, 216, 389-401.

Ceballos, J. C., Bottino, M. J., & De Souza, J. M. (2004). A simplified physical model for assessing solar radiation over Brazil using GOES 8 visible imagery. Journal of Geophysical Research: Atmospheres, 109(D2).

Chatzipoulka C, Compagnon R, Kaempf J, Nikolopoulou M. (2018). Sky view factor as predictor of solar availability on building fa¸cades. Solar Energy, 170, 1026-1038.

Chen JL, X. B., Chen CD, Wen ZF, Jiang Y, Lv MQ, Wu SJ, Li GS. (2014). Estimation of monthly-mean global solar radiation using MODIS atmospheric product over China. ournal of Atmospheric Solar-Terrestrial Physics, 110, 63-80.

Chen S, Wong NH, Zhang W, Ignatius M. (2023). The impact of urban morphology on the spatiotemporal dimension of estate-level air temperature: A case study in the tropics. Building and Environment, 228, 109843

Chen, C. F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF international conference on computer vision, 357-366.

Chen, J. L., Xiao, B. B., Chen, C. D., Wen, Z. F., Jiang, Y., Lv, M. Q., ... & Li, G. S. (2014). Estimation of monthly-mean global solar radiation using MODIS atmospheric product over China. Journal of Atmospheric and Solar-Terrestrial Physics, 110, 63-80.

Cheng V, Steemers K, Montavon M, Compagnon R. (2006). Urban Form, Density and Solar Potential. The 23rd Conference on Passive and Low Energy Architecture, 2006, 1-6.

China Meteorological News Press, (2021). China Climate Bulletin 2020.

Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Cogliani, E., Ricchiazzi, P., & Maccari, A. (2007). Physical model SOLARMET for determinating total and direct solar radiation by meteosat satellite images. Solar Energy, 81(6), 791-798.

Cortes C, V. V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Dahmani, K., Notton, G., Voyant, C., Dizene, R., Nivet, M. L., Paoli, C., & Tamas, W. (2016). Multilayer Perceptron approach for estimating 5-min and hourly horizontal global irradiation from exogenous meteorological data in locations without solar measurements. Renewable Energy, 90, 267-282.

De S JL, L., GB, Dos S CM, Junior RAF, Tiba C, Lyra GB, Lemes MAM. (2016). Empirical models of daily and monthly global solar irradiation using sunshine duration for Alagoas State, Northeastern Brazil. Sustainable Energy Technologies Assessments, 14, 35-45.

De Souza, J. L., Lyra, G. B., Dos Santos, C. M., Junior, R. A. F., Tiba, C., Lyra, G. B., & Lemes, M. A. M. (2016). Empirical models of daily and monthly global solar irradiation using sunshine duration for Alagoas State, Northeastern Brazil. Sustainable Energy Technologies and Assessments, 14, 35-45.

Deo RC, Ş. M., Adamowski JF, Mi JC. (2019). Universally deployable extreme learning machines integrated with remotely sensed MODIS satellite predictors over Australia to forecast global solar radiation: A new approach. Renewable Sustainable Energy Reviews, 104, 235-261.

Deo, R. C., Wen, X., & Qi, F. (2016). A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Applied Energy, 168, 568-593.

Dewi C, Chen RC. (2019). Random forest and support vector machine on features selection for regression analysis. . International Journal of Innovative Computing Information Control, 15, 2027-2037.

Drucker H, B. C., Kaufman L, Smola A, Vapnik V. (1997). Support vector regression machines. Advances in neural information processing systems, 9, 155-161.

Duraˇciov´a R, Pruˇzinec F. (2022). Effects of Terrain Parameters and Spatial Resolution of a Digital Elevation Model on the Calculation of Potential Solar Radiation in the Mountain Environment: A Case Study of the Tatra Mountains. ISPRS International Journal of Geo-Information, 11, 389

Edussuriya A, Chan B, Ye C. (2011). Urban morphology and air quality in dense residential environments in Hong Kong. PartI: District-level analysis, Atmos.Environ. 45, 4789–

4803

Efron B, Tibshirani R. (1985). The bootstrap method for assessing statistical accuracy. Behaviormetrika; 12:1-35

EL, M. (1998). METSTAT—The solar radiation model used in the production of the National Solar Radiation Data Base (NSRDB). Solar energy, 62(4), 263-279.

Engström, G., Gars, J., Krishnamurthy, C., Spiro, D., Calel, R., Lindahl, T., & Narayanan, B. J. N. c. (2020). Carbon pricing and planetary boundaries. 11(1), 1-11.

Ertekin, C., & Yaldız, O. (1999). Estimation of monthly average daily global radiation on horizontal surface for Antalya (Turkey). Renewable energy, 17(1), 95-102.

Fathizad H, Mobin MH, Gholamnia A, Sodaiezadeh H. (2017). Modeling and mapping of solar radiation using geostatistical analysis methods in Iran. Arabian Journal of Geosciences, 10, 1-13.

Feng F, W. K. (2021). Merging ground-based sunshine duration observations with satellite cloud and aerosol retrievals to produce high-resolution long-term surface solar radiation over China. Earth System Science Data, 13(3), 907-922.

Fleischmann M, Feliciotti A, Romice O, Porta S. (2020). Morphological tessellation as a way of partitioning space: Improving consistency in urban morphology at the plot scale. Computers, Environment Urban Systems, 80, 101441.

Freund, Y., & Schapire, R. E. (1996,). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

G, L. (1983). Estimates of irradiance over Zimbabwe. Sol. Energy, 31(6).

Gao L, C. L., Li CC, Li J, Che HZ, Zhang YP. (2021). Evaluation and possible uncertainty source analysis of JAXA Himawari-8 aerosol optical depth product over China. Atmospheric Research, 248, 105248.

Garg HP, G. S. (1982). Prediction of global solar radiation from bright sunshine hours and other meteorological parameters. Paper presented at the Solar-India, Proceedings of the National Solar Energy Convention.

Garg, H. P., & Garg, S. N. (1983). Prediction of global solar radiation from bright sunshine

hours and other meteorological data. Energy Conversion and Management, 23(2), 113-118.

Gassar AAA, Cha, SH. (2010). Review of geographic information systems-based rooftop solar photovoltaic potential estimation approaches at urban scales. Applied Energy, 291, 116817.

Gastli A, Charabi Y. Solar electricity prospects in Oman using GIS-based solar radiation maps. Renewable Sustainable Energy Reviews, 14, 790-797.

Ghimire S, D. R., Raj N, Mi JC. (2019). Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Applied energy, 253, 113541.

Ghimire, S., Deo, R. C., Raj, N., & Mi, J. (2019). Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Applied Energy, 253, 113541.

Gilbert, M. M. (2004). Renewable and efficient electric power systems: John Wiley & Sons.

Gr¨omping U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 63, 308-319.

Greenhall, geopandas/geopandas: v0.6.1; 2019. https://doi.org/10.5281/ZENODO. 3483425 [Accessed 8 February 2023].

Guermoui M, G. K., Rabehi A, Djafer D, Benkaciali S. (2018). Estimation of the daily global solar radiation based on the Gaussian process regression methodology in the Saharan climate. The European Physical Journal Plus 133(6), 1-17.

Gürel, A. E., Ağbulut, Ü., Bakır, H., Ergün, A., & Yıldız, G. (2023). A state of art review on estimation of solar radiation with various models. Heliyon, 9(2).

Hagberg A, Swart PS, Chult D. Exploring network structure, dynamics, and function using NetworkX. Springer 2008, 8, 5495.

Hargreaves, G. H., & Samani, Z. A. (1982). Estimating potential evapotranspiration. Journal of the irrigation and Drainage Division, 108(3), 225-230.

Holden, P. B., Edwards, N. R., Ridgwell, A., Wilkinson, R. D., Fraedrich, K., Lunkeit, F., ... & Viñuales, J. E. (2018). Climate–carbon cycle uncertainties and the Paris Agreement. Nature Climate Change, 8(7), 609-613.

Hong Kong Building Rooftop Solar Map; 2022. https://solarmap.emsd.gov.hk/map [Accessed 8 February 2023].

Hong Kong Energy End-use Data, 2023. https://data.gov.hk/en-data/dataset/hk-emsd-emsd1-energy-end-use-data-2023

Hong Kong Observatory; 2023. https://www.hko.gov.hk/en/cis/climat.htm [Accessed 8 February 2023].

Hong T, Lee M, Koo C, Jeong K, Kim J. Development of a method for estimating the rooftop solar photovoltaic (PV) potential by analyzing the available rooftop area using Hillshade analysis. Applied Energy 2017; 194:320-332.

Huang Y, S. S., Manton M, Protat A, Majewski L, Nguyen H. (2019). Evaluating Himawari-8 cloud products using shipborne and CALIPSO observations: Cloud-top height and cloud-top temperature. Journal of Atmospheric Oceanic Technology, 36(12), 2327-2347.

International Energy Agency; 2021. https://www.iea.org/reports/renewables-2021/executive-summary [Accessed 8 February 2023].

Ismail, M. T., & Karim, S. A. A. (2020). Time Series Models of High Frequency Solar Radiation Data. Practical Examples of Energy Optimization Models, 79-89.

Izquierdo S, Monta˜n´es C, Dopazo C, Fueyo N. (2011). Roof-top solar energy potential unde performance-based building energy codes: The case of Spain. Solar Energy; 85:208-213.

Izquierdo S., Rodrigues M. and 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.

Jalil-Vega F, Kerdan I, Hawkes A. (2020). Spatially-resolved urban energy systems model to study decarbonisation pathways for energy services in cities. Applied Energy, 262, 114445.

Javanroodi K, Mahdavinejad M, Nik VM. (2018). Impacts of urban morphology on reducing cooling load and increasing ventilation potential in hot-arid climate.. Applied energy; 231:714-746.

JH, F. (2001). Greedy function approximation: a gradient boosting machine. Annals of

statistics, 1189-1232.

Ji, W., & Chee, K. C. (2011). Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Solar energy, 85(5), 808-817.

JK, P. (1997). Proposed quality control procedures for the meteorological office data tapes relating to global solar radiation, diffuse solar radiation, sunshine and cloud in the UK. Report FCIBSE.

Kammen DM, Sunter DA. (2016). City-integrated renewable energy for urban sustainability. Science, 352, 922-928.

Kannan, N., & Vakeesan, D. (2016). Solar energy for future world:-A review. Renewable and sustainable energy reviews, 62, 1092-1105.

Khosravi, A., Koury, R. N. N., Machado, L., & Pabon, J. J. G. (2018). Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms. Journal of Cleaner Production, 176, 63-75.

Ko L, Wang JC, Chen CY, Tsai HY. (2015). Evaluation of the development potential of rooftop solar photovoltaic in Taiwan. Renewable Energy; 76:582-595.

Kong, X., Du, X., Xu, Z., & Xue, G. (2023). Predicting solar radiation for space heating with thermal storage system based on temporal convolutional network-attention model. Applied Thermal Engineering, 219, 119574.

Kumari P, T. D. (2021). Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. Journal of Cleaner Production, 279, 123285.

Kumari, P., & Toshniwal, D. (2021). Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. Journal of Cleaner Production, 279, 123285.

Kundur P, Paserba J, Ajjarapu V, Andersson G, Bose A, Canizares C, Hatziargyriou N, Hill D, Stankovic A, Taylor C. (2004). Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions. IEEE transactions on Power Systems, 19, 1387-1401.

L, B. (2001). Random forests. Machine learning, 45(1), 5-32.

L´opez C, Sala M, Tagliabue L, Frontini F, Bouziri S. (2016). Solar Radiation and Daylighting

Assessment Using the Sky-view Factor (SVF) Analysis as Method to Evaluate Urban Planning Densification Policies Impacts. Energy Procedia, 91, 989-996.

Le TB, K. D., Xie HJ, Dong B, Vega RE. (2016). LiDAR-based solar mapping for distributed solar plant design and grid integration in San Antonio, Texas. Remote Sensing, 8(3), 247.

Lee J, W. W., Harrou F, Sun Y. (2020). Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. Energy Conversion Management, 208, 112582.

Leng H, Chen X, Ma Y, Wong N, Ming T. (2020). Urban morphology and building heating energy consumption: Evidence from Harbin, a severe cold region city. Energy Buildings, 224, 110143.

Lewis, G. (1983). Estimates of irradiance over Zimbabwe. Sol. Energy;(United Kingdom), 31(6).

Li D, Liu G, Liao S. (2015). Solar potential in urban residential buildings. Solar Energy, 111, 225-235.

Li, R., Ma, T., Xu, Q., & Song, X. (2018). Using MAIAC AOD to verify the PM2.5 spatial patterns of a land use regression model. Environmental Pollution, 243, 501-509.

Liao X, Zhu R, Wong, MS. (2022). Simplified estimation modeling of land surface solar irradiation: A comparative study in Australia and China.. Sustainable Energy Technologie and Assessments, 52, 102323.

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764.

Lima I, Scalco V, Lamberts R. (2018). Estimating the impact of urban densification on high-rise office building cooling loads in a hot and humid climate. Energy Buildings; 182:30-44.

Madlener, R., & Sunak, Y. (2011). Impacts of urbanization on urban structures and energy demand: What can we learn for urban energy planning and urbanization management? Sustainable Cities Society, 1(1), 45-53.

Makade, R. G., Chakrabarti, S., & Jamil, B. (2019). Prediction of global solar radiation using a

single empirical model for diversified locations across India. Urban Climate, 29, 100492.

Martins T, Adolphe L, Bastos L. (2014). From solar constraints to urban design opportunities: Optimization of built form typologies in a Brazilian tropical city. Energy Buildings, 76, 43–56.

Masters, G. M. (2013). Renewable and efficient electric power systems: John Wiley & Sons.

Maxwell, E. L. (1998). METSTAT—The solar radiation model used in the production of the National Solar Radiation Data Base (NSRDB). Solar Energy, 62(4), 263-279.

Mishra T, Rabha A, Kumar U, Arunachalam K, Sridhar V. (2020). Assessment of solar power potential in a hill state of India using remote sensing and Geographic Information System-ScienceDirect. Remote Sensing Applications: Society Environment, 19, 100370

Mohajeri N, Assouline D, Guiboud B, Bill A, Gudmundsson A, Scartezzini, JL. (2018). A city scale roof shape classification using machine learning for solar energy applications. Renewable Energy, 121, 81-93.

Mohajeri N, Upadhyay G, Gudmundsson A, Assouline D, K¨ampf J, Scartezzini J. (2016). Effects of urban compactness on solar energy potential. Renewable Energy, 93, 469-482.

Mohammadi, K., Shamshirband, S., Anisi, M. H., Alam, K. A., & Petković, D. (2015). Support vector regression based prediction of global solar radiation on a horizontal surface. Energy Conversion and Management, 91, 433-441.

Morganti M, Salvati A, Coch H, Cecere C. (2017). Urban morphology indicators for solar energy analysis. Energy Procedia, 134, 807-814.

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. Neurocomputing, 2(5-6), 183-197.

NASA Power; 2021. https://power.larc.nasa.gov/data-access-viewer/ [Accessed 8 February 2023].

Nelson JR, Grubesic TH. (2020). The use of LiDAR versus unmanned aerial systems (UAS) to assess rooftop solar energy potential. Sustainable Cities Society; 61:102353.

Neter J, Wasserman W, Kutner MH. (1996). Applied Linear Statistical Models. Technometrics,

39, 880-880.

Ng E, Chao Y, Liang C, Chao R, Fung J. (2011). Improving the wind environment in high-density cities by understanding urban morphology and surface roughness: A study in Hong Kong. Landscape Urban Planning, 101, 59-74.

Nikitidou E, Z. A., Salamalikis V, Kazantzidis A. (2019). Short-term cloudiness forecasting for solar energy purposes in Greece, based on satellite-derived information. Meteorology Atmospheric Physics, 131(2), 175-182.

Nikitidou, E., Zagouras, A., Salamalikis, V., & Kazantzidis, A. (2019). Short-term cloudiness forecasting for solar energy purposes in Greece, based on satellite-derived information. Meteorology and Atmospheric Physics, 131, 175-182.

Outlook, Energy. (2010) International Energy Outlook. Outlook.

Ozgoren, M., Bilgili, M., & Sahin, B. (2012). Estimation of global solar radiation using ANN over Turkey. Expert systems with applications, 39(5), 5043-5051.

Page, J. K. (1997). Proposed quality control procedures for the meteorological office data tapes relating to global solar radiation, diffuse solar radiation, sunshine and cloud in the UK. Report FCIBSE.

Paltridge, G. W., & Proctor, D. (1976). Monthly mean solar radiation statistics for Australia. Solar Energy, 18(3), 235-243.

Park J, M. J., Jung S, Hwang E. (2020). Multistep-ahead solar radiation forecasting scheme based on the light gradient boosting machine: A case study of Jeju Island. Remote Sensing, 12(14), 2271.

Paulescu M, T. P. E., Stefu N. (2011). A temperature-based model for global solar irradiance and its application to estimate daily irradiation values. International Journal of Energy Research, 35(6), 520-529.

Poon KH, K¨ampf JH, Tay SER, Wong NH, Reindl TG. (2020). Parametric study of URBAN morphology on building solar energy potential in Singapore context. Urban Climate, 33, 100624.

Prada, J., & Dorronsoro, J. R. (2018). General noise support vector regression with non-constant uncertainty intervals for solar radiation prediction. Journal of Modern Power Systems and Clean Energy, 6(2), 268-280.

Premalatha, M., & Naveen, C. (2018). Analysis of different combinations of meteorological parameters in predicting the horizontal global solar radiation with ANN approach: A case study. Renewable Sustainable Energy Reviews, 91, 248-258.

Rabehi, A., Guermoui, M., & Lalmi, D. (2020a). Hybrid models for global solar radiation prediction: a case study. International Journal of Ambient Energy, 41(1), 31-40.

Ramedani Z, O. M., Keyhani A, Khoshnevisan B, Saboohi H. (2014). A comparative study between fuzzy linear regression and support vector regression for global solar radiation prediction in Iran. Solar Energy, 109, 135-143.

RE, B. (1984). A simple, solar spectral model for direct-normal and diffuse horizontal irradiance. Solar energy, 32(4), 461-471.

Reda, I., & Andreas, A. (2004). Solar position algorithm for solar radiation applications. Solar energy, 76(5), 577-589.

Rey SJ, Anselin L. (2010). PySAL: A Python library of spatial analytical methods. In Handbook of applied spatial analysis. Springer, 1, 175-193.

Rich P, Dubayah R, Hetrick W, Saving S. (1994). Using viewshed models to calculate intercepted solar radiation: applications in ecology. American Society for Photogrammetry and Remote Sensing Technical Papers.. In American Society of Photogrammetry and Remote Sensing, 1, 524-529.

Robinson D. (2006). Urban morphology and indicators of radiation availability. Solar Energy, 80, 1643-1648

Rodriguez JD, P. A., Lozano JA . (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis machine intelligence, 32(3), 569-575.

Rohani, A., Taki, M., & Abdollahpour, M. (2018). A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I). Renewable Energy, 115, 411-422.

Saadaoui H, Ghennioui A, Ikken B, Rhinane H, Maanan M. (2019). Using GIS and photogrammetry for assessing solar photovoltaic potential on Flat Roofs in urban area case of the city of Ben Guerir/Morocco. International Archives of the Photogrammetry, Remote Sensing Spatial Information Sciences; 42:155-166

Sarralde J, Quinn D, Wiesmann D, Steemers K. (2015). Solar energy and urban morphology: Scenarios for increasing the renewable energy potential of neighbourhoods in London. Renewable energy, 73, 10-17.

Schapire RE, Singer Y. (2000). BoosTexter: A Boosting-based System for Text Categorization. Machine Learning; 39:p.135-168.

Segal MR. (2004). Machine Learning Benchmarks and Random Forest Regression. Center for Bioinformatics Molecular Biostatistics, 1, 1-14.

Shadab, A., Ahmad, S., & Said, S. (2020). Spatial forecasting of solar radiation using ARIMA model. Remote Sensing Applications: Society and Environment, 20, 100427.

Sharafati, A., Khosravi, K., Khosravinia, P., Ahmed, K., Salman, S. A., Yaseen, Z. M., & Shahid, S. (2019). The potential of novel data mining models for global solar radiation prediction. International Journal of Environmental Science and Technology, 16, 7147-7164.

Silva, V. L. G. D., Oliveira Filho, D., Carlo, J. C., & Vaz, P. N. (2022). An approach to solar radiation prediction using ARX and ARMAX models. Frontiers in Energy Research, 10, 822555.

Srivastava, R., Tiwari, A., & Giri, V. (2019). Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. Heliyon, 5(10), e02692.

Statistical Review of World Energy 2023, 2023. https://www.energyinst.org/statistical-review

Sun, S., Wang, S., Zhang, G., & Zheng, J. (2018). A decomposition-clustering-ensemble learning approach for solar radiation forecasting. Solar Energy, 163, 189-199.

Swartman RK, O. O. (1967). Solar radiation estimates from common parameters. Solar energy, 11(3-4), 170-172.

Swartman, R. K., & Ogunlade, O. (1967). Solar radiation estimates from common parameters. Solar energy, 11(3-4), 170-172.

Tabik S, Villegas A, Zapata EL, Romero LF. (2012). A Fast GIS-tool to Compute the Maximum Solar Energy on Very Large Terrains. Procedia Computer Science, 9, 364-372.

Tanu M, Amponsah W, Yahaya B, Bessah E, Ansah SO, Wemegah CS, Agyare WA. (2021). Evaluation of global solar radiation, cloudiness index and sky view factor as potential

indicators of Ghana's solar energy resource. . Scientific African, 14, e01061.

Tarasova, T. A., & Fomin, B. A. (2000). Solar radiation absorption due to water vapor: Advanced broadband parameterizations. Journal of Applied Meteorology, 39(11), 1947-1951.

Tariq, G. H., Ashraf, M., & Hasnain, U. S. (2021). Solar technology in agriculture. Technology in Agriculture, 387.

The Government of the Hong Kong (SAR) Press Releases; 2022. https://www.info.gov.hk/gia/general/202204/26/P2022042600448.htm [Accessed 8 February 2023].

Tong H, Walton A, Sang J, Chan JC. (2005). Numerical simulation of the urban boundary layer over the complex terrain of Hong Kong. Atmospheric environment, 39, 3549-3563.

Urraca, R., Antoñanzas, J., Antoñanzas-Torres, F., & Martinez-de-Pison, F. J. (2016). Estimation of daily global horizontal irradiation using extreme gradient boosting machines. Paper presented at the International Joint Conference SOCO'16-CISIS'16-ICEUTE'16.

Vaka, M., Walvekar, R., Rasheed, A. K., & Khalid, M. (2020). A review on Malaysia's solar energy pathway towards carbon-neutral Malaysia beyond Covid'19 pandemic. Journal of cleaner production, 273, 122834.

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.

Walch A, Castello R, Mohajeri N, Scartezzini JL. (2020). A fast machine learning model for large-scale estimation of annual solar irradiation on rooftops. ISES Solar World Congress, 45, 1-10.

Wang J, Li P, Ran R, Che Y, Zhou Y. (2018). A short-term photovoltaic power prediction model based on the gradient boost decision tree. Applied Sciences, 8, 689.

Wang MZ, C. H., Merrick JR, Amati M. (2016). Assessment of solar radiation reduction from urban forests on buildings along highway corridors in Sydney. Urban forestry urban greening, 15, 225-235.

Wang, J., Jiang, H., Wu, Y., & Dong, Y. (2015). Forecasting solar radiation using an optimized

hybrid model by Cuckoo Search algorithm. Energy Conversion Management, 81, 627-644.

Wei R, Song D, Wong NH, Martin M. (2016). Impact of Urban Morphology Parameters on Microclimate. Procedia Engineering, 169, 142-149.

Wong MS, Z. R., Liu ZZ, Lu L, Peng JQ, Tang ZQ, Lo CH, Chan WK. (2016). Estimation of Hong Kong's solar energy potential using GIS and remote sensing technologies. Renewable Energy, 99, 325-335.

Wu, L., Huang, G., Fan, J., Zhang, F., Wang, X., & Zeng, W. (2019). Potential of kernel-based nonlinear extension of Arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions. Energy Conversion and Management, 183, 280-295.

Y, F. (1995). Boosting a weak learning algorithm by majority. Information computation, 121(2), 256-285.

Yagli, G. M., Yang, D., & Srinivasan, D. (2019). Automatic hourly solar forecasting using machine learning models. Renewable and Sustainable Energy Reviews, 105, 487-498.

Yang J, Yang Y, Sun D, Jin C, Xiao X. (2021). Influence of urban morphological characteristics on thermal environment. Sustainable Cities Society, 72, 103045.

Yang K, H. G., Tamai N. (2001). A hybrid model for estimating global solar radiation. Solar energy, 70(1), 13-22.

Yeom, J. M., Seo, Y. K., Kim, D. S., & Han, K. S. (2016). Solar radiation received by slopes using COMS imagery, a physically based radiation model, and GLOBE. Journal of Sensors, 2016.

Yildirim, A., Bilgili, M., & Ozbek, A. (2023). One-hour-ahead solar radiation forecasting by MLP, LSTM, and ANFIS approaches. Meteorology and Atmospheric Physics, 135(1), 10.

Yong X, Chao R, Ma P, Ho J, Ng E. (2017). Urban morphology detection and computation for urban climate research. Landscape Urban Planning, 167, 212-224.

Yu, B., Liu, H., Wu, J., & Lin, W. (2009). Investigating impacts of urban morphology on spatio-temporal variations of solar radiation with airborne LIDAR data and a solar flux model: a case study of downtown Houston. International Journal of Remote Sensing, 30(17),

4359-4385.

Zang, H., Cheng, L., Ding, T., Cheung, K. W., Wang, M., Wei, Z., & Sun, G. (2020). Application of functional deep belief network for estimating daily global solar radiation: A case study in China. Energy Conversion Management, 191, 116502.

Zhang YL, L. X., Bai YL. (2015). An integrated approach to estimate shortwave solar radiation on clear-sky days in rugged terrain using MODIS atmospheric products. Solar energy, 113, 347-357.

Zhang, J., Zhao, L., Deng, S., Xu, W., & Zhang, Y. (2017). A critical review of the models used to estimate solar radiation. Renewable and Sustainable Energy Reviews, 70, 314-329.

Zhang, Y., Li, X., & Bai, Y. (2015). An integrated approach to estimate shortwave solar radiation on clear-sky days in rugged terrain using MODIS atmospheric products. Solar Energy, 113, 347-357.

Zhou Y, L. Y., Wang D, Liu X, Wang Y. (2021). A review on global solar radiation prediction with machine learning models in a comprehensive perspective. Energy Conversion Management, 235, 113960.

Zhu R, Cheng C, Santi P, Chen M, Zhang X, Mazzarello M, Wong M, Ratti C. (2022). Optimization of photovoltaic provision in a three-dimensional city using real-time electricity demand. Applied Energy, 316, 119042.

Zhu R, Man SW, You L, Santi, Ratti C. (2020). The effect of urban morphology on the solar capacity of three-dimensional cities. Renewable Energy, 153, 1111-1126.

Zhu R, You L, Santi P, Man S, Ratti C. (2019). Solar accessibility in developing cities: A case study in Kowloon East, Hong Kong. Sustainable Cities Society, 51, 101738