國立臺灣大學工學院土木工程研究所

碩士論文

Department of Civil Engineering

College of Engineering

National Taiwan University

Master's Thesis

探索資料融合技術應用於多元天氣資料預測太陽能發電之潛力

Predicting Solar Power Potential Using Multiple Weather Data Sources: Exploring Data Merging Techniques

江照新

Chao-Hsin Chiang

指導教授: 汪立本 博士

共同指導教授: 謝依芸 博士

Advisor: Li-Pen Wang, Ph.D.

Co-advisor: I-Yun Lisa Hsieh, Ph.D.

中華民國 114 年 7 月

July, 2025

doi:10.6342/NTU202503010

國立臺灣大學碩士學位論文
口試委員會審定書

NATIONAL TAIWAN UNIVERSITY
MASTER'S THESIS ACCEPTANCE CERTIFICATE

探索資料融合技術應用於多元天氣資料預測太陽能發電之潛力

Predicting Solar Power Potential Using Multiple Weather Data Sources: Exploring Data Merging Techniques

本論文係 江照新 (R12521608) 在國立臺灣大學土木工程學系電腦輔助工程組完成之碩士學位論文，於民國114年7月25日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Computer-Aided Engineering Division of the Department of Civil Engineering on July 25, 2025 have examined a Master's Thesis entitled above presented by Chao-Hsin Chiang (R12521608) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

汪立本 汪立本
(指導教授 Advisor)

謝依芸 謝依芸
(共同指導 Co-Advisor)

張書瑋 張書瑋

系主管 Director： 葛宇甯 葛宇甯

Acknowledgements

第一個想到要感謝的，就是謝謝汪老師在我畢業時限逼近時，沒有質疑、責罵，反而是持續支持、給予我意見，並幫助我修改論文，真的很感謝老師的大心臟，在論文還只有個雛型時仍替我安排口試，讓我能保持良好的狀態在這最後一個月安心的衝刺。

這一路上有好多好多的協助，謝謝依芸老師，幫助我熟悉研究領域，並協助我聯絡意凡學長；謝謝意凡在研究初期慷慨分享資料與程式碼，還開直播為我講解研究；謝謝炳璋，陪我討論思考生活與研究、和我分享模型訓練的知識；謝謝麒凌，幫助我學會使用 Linux 系統，並帶我進入程式、系統與 Server 的世界；謝謝敬淳，在我剛進研究室時，幫助我更快融入環境；謝謝筠曼，熱心的幫助我準備口試流程，真的感受到了你的善良熱心；謝謝宏明，在地理空間座標方面提供協助，幫助我初步了解坐標系統並學會使用地理空間資料；也謝謝 WangUp 研究室的大家，在開會時提供各種想法，你們的建議很寶貴而且有建設性，還記得那時候做研究海報，開了一次剛好汪老師不在的會，結束就跟老師說暫時不用幫忙了，多了很多方向去修正提升自己的海報，相信老師也很開心有這樣的大家組成我們這個團隊，有你們我們才能一起 Up！

也謝謝我的寶貝，給予我好多好多的鼓勵與支持，在半夢半醒中陪我討論研究、還幫我找文獻 (是有用在最終引用的喔)，沒有妳我想我不知道什麼時候才

doi:10.6342/NTU202503010

會順利把整個研究完成；謝謝上帝，雖然我看不到你到底有沒有幫忙，但我相信我這一路走來，從低谷中爬起，離不開你的幫助，也謝謝聖靈陪著我禱告；謝謝自己，真的很開心你在最需要動力、衝勁的這段時間，成功讓過往的一切自我認識、心理覺察幻化出實際的成效，陪伴自己在最後這一個月高產能且穩定的一步一步把論文堆砌起來；謝謝姊姊陪我吃飯聊天，也謝謝爸爸媽媽聽我報告論文、給我建議並鼓勵我；謝謝物治江老師還有主任，讓我舒緩疼痛、練習聆聽與照顧自己的身體；有你們的陪伴，讓我不是獨自奮鬥與前進。

再來就是要感謝廣大網民還有 OpenAI 啦 (還是我該感謝全體人類？)，謝謝你們，謝謝 ChatGPT，幫助我知道什麼是更好、正式的學術寫作，幫助我更快速的推進論文，把更多的精力用在把論文修得更好、讓整體更完善上。

最後，也謝謝一路上幫助過我的每一位同學與朋友，這兩年發生的太多，沒辦法一一描述，若有遺漏，也一併表達感謝，謝謝每一個你，謝謝每一份善。

doi:10.6342/NTU202503010

摘要

在太陽能全球化發展的趨勢下，太陽能潛勢預測面臨日益差異化的地面天氣資料觀測條件。地面觀測資料為傳統上最準確的天氣資料來源，然而其空間分布不均，使作為預測主要輸入的天氣資料在可靠性上產生疑慮。本研究旨在探討氣象輸入資料的品質如何影響預測結果，進而提升太陽能潛勢預測的整體可靠性。

本研究採用兩種天氣資料來源：中央氣象署地面觀測資料 (點資料) 與 ERA5 重分析資料 (格點資料)，並設計八種天氣輸入組合，涵蓋單一來源、雙資料來源使用策略 (包含平行化輸入、Kriging with External Drift 與 Kriging with Radar-Based Error Correction 兩種資料融合技術) 及關鍵地面變數的納入與否。預測模型採用長短期記憶 (LSTM) 神經網路，並在相同模型架構與訓練流程下進行系統性比較分析。為模擬全球不同地區的觀測資源差異，進一步設計十三種地面觀測密度情境，並評估各輸入組合於不同情境下之預測表現。

研究結果指出，在使用單一資料來源的情況下，地面觀測資料於太陽能潛勢預測的優勢範圍約為氣象站距離小於 75 公里；而當觀測資源稀缺時，格點資料則展現出良好的穩定性與應用潛力。在 50–75 公里這一關鍵距離範圍內，雙資料來源策略的表現優於僅使用格點資料，其中以平行化輸入法效果最佳，其次為 KRE 融合方法。此外，研究亦發現三項地面觀測專屬變數：日照率、日照時數與最大紫外線指數，對於提升預測精度具有正面效益。

doi:10.6342/NTU202503010

關鍵字：太陽能發電、發電潛勢、預測模型、資料融合、地面氣象資料、格點天氣資料、天氣站距離

doi:10.6342/NTU202503010

Abstract

Under the global trend of solar energy development, solar power potential prediction faces increasingly diverse ground weather observation conditions. Although ground-based weather data are traditionally the most accurate source, their uneven spatial distribution raises concerns regarding the reliability of weather inputs for prediction. This study aims to examine how the quality of weather inputs affects prediction outcomes, in order to improve the overall reliability of solar power potential prediction.

Two sources of weather data were used: point-based observations from the Central Weather Administration (CWA) and gridded ERA5-reanalysis data. Eight input combinations were designed to represent different strategies, including single-source inputs, dual-source approaches (parallel input and two data merging techniques: Kriging with External Drift [KED], and Kriging with Radar-Based Error Correction [KRE]), and inclusion of key ground-based variables. A Long Short-Term Memory (LSTM) neural network was employed as the prediction model, and all combinations were evaluated under the same

doi:10.6342/NTU202503010

model architecture and training procedure. To simulate varying observational resource conditions worldwide, 13 weather station density scenarios were further constructed to assess performance under different levels of data availability.

The results show that ground-based data outperform gridded data when the distance to the nearest station is less than 75 km. When observational resources are limited, ERA5 gridded data demonstrates better stability and applicability. Within the critical 50–75 km range, dual-source strategies showed better performance than gridded-only inputs, with the parallel input method yielding the best performance, followed by the KRE data merging method. In addition, three CWA-specific variables: sunshine duration, sunshine percentage, and maximum UV index were found to have positive effects on prediction accuracy.

Keywords: solar power, power potential, prediction model, data merging, ground weather data, gridded weather data, weather station distance

doi:10.6342/NTU202503010

Contents

doi:10.6342/NTU202503010

doi:10.6342/NTU202503010

doi:10.6342/NTU202503010

List of Figures

doi:10.6342/NTU202503010

doi:10.6342/NTU202503010

List of Tables

doi:10.6342/NTU202503010

Chapter 1 Introduction

1.1 Background and Motivation

In the context of the global transition to a sustainable society, solar energy has emerged as one of the fastest-growing forms of renewable energy. According to Chapter 3.0 "Electricity" in *Renewables 2024* by the International Energy Agency (IEA) [1], the global renewable energy capacity is expected to increase by more than 5,520 GW between 2024 and 2030, which is 2.6 times the increase observed between 2017 and 2023. Among this growth, solar energy alone is projected to account for over 4,200 GW, representing more than three-quarters of the total increase in renewable capacity.

Reliable solar power potential prediction plays a crucial role in the development of solar energy. Solar power potential prediction refers to the assessment of the future potential solar energy output at a target site, either at a pilot site prior to deploying a solar power plant or at an existing power plant during its operation. It is, respectively, used to evaluate the suitability of the pilot site and to forecast future power generation. Accurate prediction of potential solar power resources allows policy makers and energy planners to identify optimal sites for solar power plants, as well as to understand future energy supply trends, supporting decisions on supply adequacy and comprehensive renewable energy planning.

doi:10.6342/NTU202503010

Over the past few decades, a wide range of methods have been developed for solar power potential prediction. These can be broadly classified into four categories: statistical, physical, artificial intelligence (AI)-based, and hybrid methods. This classification is summarized in a recent review by Jannah et al. (2024) [2].

To facilitate discussion, this study simplifies solar power potential prediction into three key components:

(1) weather input, consisting of several weather variables such as temperature, solar radiation, and cloud cover; these variables describe the meteorological conditions at target site;

(2) potential prediction model, which, as defined in this context, corresponds to the previously introduced approaches (namely, statistical, physical, AI-based, and hybrid models); and

(3) predicted energy output of the target site.

A review of previous studies (e.g., [3], [4], [5], [6]) indicates that past studies have primarily focused on improving the accuracy of prediction models. However, this study posits that one critical aspect has received comparatively little attention: the reliability of weather input. Readers seeking a broader understanding of model-focused research in this domain may refer to Jannah et al. (2024) [2].

Concerns about the reliability of weather input are particularly relevant in regions where ground-based observations are limited or unavailable. Ground weather stations are traditionally considered the most accurate and stable source of weather data. However, due to high installation and maintenance costs, their number and spatial coverage are lim-

doi:10.6342/NTU202503010

ited. Moreover, as solar energy continues to expand globally, this limitation becomes increasingly significant. Earlier studies often focused on single regions, where limited ground observations had a relatively minor impact. In contrast, under the current global expansion, many regions have little to no ground weather station coverage, making this issue more critical, especially in remote or underdeveloped areas.

A common strategy to address this issue is to include gridded weather datasets as a complementary source, such as satellite images, radar data, and numerical weather prediction (NWP) outputs. They typically offer broad spatial coverage but lower accuracy compared to point-based ground observations. Combining these two complementary data sources can enhance both the completeness and applicability of the resulting weather input. Such complementarity has been discussed in studies such as [7]. Several other studies (e.g., [8]; [9]; [10]) have adopted both ground-based and gridded weather data to better support solar power potential prediction.

Although an increasing number of studies have incorporated the aforementioned complementary datasets in recent years, the author observes that research on solar power potential has rarely examined the reliability of weather input in depth. As solar energy expands globally, the diversity of weather data sources continues to increase, including various ground-based and gridded datasets. This also entails the use of various source-integration methods. Such growing complexity underscores the importance and necessity of systematically investigating how different sources and processing strategies of original weather data influence prediction outcomes.

doi:10.6342/NTU202503010

1.2 Research Aim and Objectives

This study aims to improve the reliability of solar power potential prediction by examining how weather input influences prediction outcome, with an initial focus on exploring data merging techniques for combining ground-based and gridded weather data. For clarity, weather input in this study refers to the combination of several estimated weather variables, such as temperature and solar radiation, at the target sites. The target sites in this study correspond to the 40 existing solar power plants (detailed in Section 2.3). These estimated variables represent the local weather conditions used as model input for solar potential prediction. The estimation methods are detailed in Sections 3.2 and 3.2.1.

For consistency and comparability, the same study area and overall data structure are adopted as in the research framework proposed by Feng (2023) [11]. In that study, two distinct solar power potential models were developed, each trained separately using ground-based observations and gridded reanalysis data as model input.

The scope of analysis is further narrowed by focusing on two key aspects of weather input design in this study:

1. The spatial availability of nearby ground weather stations; and
2. The estimation methods used to transform original weather data into location-specific input.

This study employs two types of meteorological datasets as input sources for solar power potential prediction: (1) point-based ground observations from Taiwan's Central Weather Administration (CWA), and (2) gridded ERA5-reanalysis datasets, which serve

doi:10.6342/NTU202503010

as complementary data to ground-based observations.

Building on these data sources, eight weather variables input combinations are constructed based on three experimental dimensions (the variables used in each combination and their estimation methods are detailed in Section 3.2.2):

1. Use of a single data source (either ground or gridded).
2. Integration of two data sources, including two data merging techniques: Kriging with External Drift (KED) and Kriging with Radar-based Error correction (KRE).
3. Inclusion or exclusion of three key ground-only variables.

An LSTM (Long Short-Term Memory) neural network is adopted as the prediction model, as it was identified as the most effective approach in the framework proposed by Feng (2023) [11]. Each input combination is trained and tested using the same model architecture, hyperparameter settings, and training procedure, ensuring a consistent basis for comparative analysis. Detailed model configurations and implementation steps are provided in Chapter 3.

Based on the above setup, each input combination is independently used to train a unified LSTM architecture with fixed settings, enabling this study to assess how different weather inputs affect solar potential prediction by comparing their prediction performance.

To further simulate the global diversity of ground weather station availability, in Section 3.3, this study defines 13 distance thresholds for each target solar power plant. For each threshold, all ground stations locate within the specific distance are excluded from the input construction process. Seven of the eight weather input combinations are generated based on these filtered ground observations. In line with this framework, comparing

doi:10.6342/NTU202503010

models trained using the same input combination but under different filtering scenarios enables a systematic assessment of how the availability of nearby stations affects prediction performance.

doi:10.6342/NTU202503010

Chapter 2 Study Area and Datasets

2.1 Study Area

Taiwan is selected as the study area for this research, covering a geographic range from 20.8°N to 26.5°N latitude and 118°E to 123°E longitude, including both the main island and surrounding islets. Taiwan lies between the subtropical and tropical zones and receives abundant solar energy. In addition, the high density of ground weather stations enables the effective simulation of various ground observation scenarios worldwide, ranging from dense to sparse distributions.

2.2 Weather Data Sets

2.2.1 Ground Weather Data

The ground weather dataset used in this study is provided by the Central Weather Administration of Taiwan [12, 13]. This dataset includes both ground weather observations and the coordinates of each weather station. The data spans the period from 2017 to 2021 and has a daily temporal resolution. It includes the following eight weather variables: temperature (°C), relative humidity (%), total precipitation (mm), global solar radiation

doi:10.6342/NTU202503010

(MJ/m²), total cloud cover (0 – 10), sunshine duration (hour), sunshine percentage (%), and daily maximum UV index (0–15). Figure 2.1 shows the spatial distribution of ground weather stations that recorded global solar radiation data on January 1, 2017.

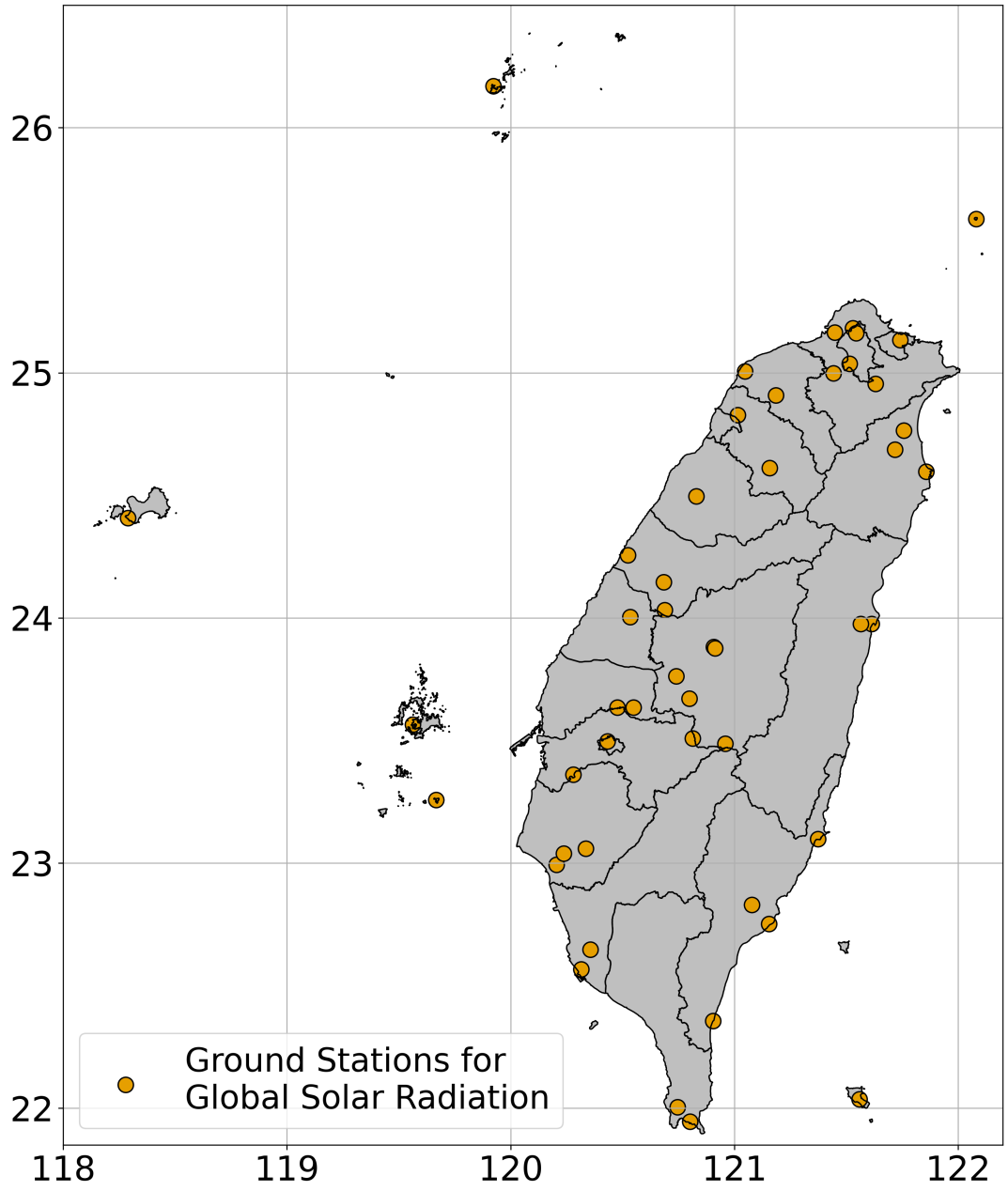


Figure 2.1: Geographic distribution of ground weather stations in Taiwan (with available global solar radiation data on January 1, 2017).

doi:10.6342/NTU202503010

2.2.2 Gridded Weather Data

The gridded weather dataset consists of three ERA5 reanalysis products: ERA5-Land [14], ERA5 on Single Level [15], and ERA5 on Pressure Level [16]. All datasets are obtained from the Climate Data Store and cover the period from 2017 to 2021. Each dataset is described in detail below.

ERA5-Land: Covers only the land areas of Taiwan's main island and provides variables including 2-meter temperature (K), total precipitation (m), surface solar radiation downwards (J/m²), surface latent heat flux (J/m²), surface sensible heat flux (J/m²), surface net thermal radiation (J/m²), evaporation (m), potential evaporation (m). The temporal resolution is hourly, with a spatial resolution of 0.1° × 0.1°.

ERA5 on Single Level: Covers the entire study area. Provides 2-meter temperature (K), total precipitation (m), and surface solar radiation downwards (J/m²) data for coastal and islets areas. Additionally, total cloud cover (0–1) is extracted across the entire study area to enrich variable diversity. This dataset has an hourly temporal resolution and a spatial resolution of 0.25° × 0.25°.

ERA5 on Pressure Level: Covers the entire study area and provides relative humidity (%) at the 1000 hPa pressure level. The temporal resolution is hourly, and the spatial resolution is 0.25° × 0.25°.

All ERA5 variables are processed to ensure a consistent daily temporal resolution and standardized units. Among these, 2-meter temperature (K), relative humidity (%), and total cloud cover (0–1) are instantaneous variables, and their daily values are obtained by calculating the 24-hour mean. The remaining variables are accumulated variables,

doi:10.6342/NTU202503010

and their daily values are derived by summing hourly values over a full day. However, for ERA5-Land, accumulated variables follow a different convention: the value at 00:00 UTC on day D + 1 corresponds to the accumulation from 00:00 to 24:00 UTC on day D (ECMWF, 2023) [17].

To align ERA5 variables with ground-based ones, in this study, surface solar radiation downwards (J/m²) from ERA5 is used as the counterpart to the ground variable global solar radiation (MJ/m²), and 2-meter temperature (K) from ERA5 is used as the counterpart to ground variable temperature (℃). In addition, relative humidity (%) at the 1000 hPa pressure level from ERA5 is used to represent near-surface humidity, corresponding to the ground-based relative humidity (%) records.

The final set of variables used in this study, along with their abbreviations, sources, and unified units, is summarized in Table 2.1.

Table 2.1: Summary of Weather Variables Used in This Study

Abbreviation	Corresponding Variable	Ground	ERA5	Unified Unit
T	Temperature	v	v	℃
RH	Relative Humidity	v	v	%
P	Total Precipitation	v	v	mm
GR	Global Solar Radiation	v	v	MJ/m²
TC	Total Cloud Cover	v	v	Index (0–10)
SD	Sunshine Duration	v	-	hour
SP	Sunshine Percentage	v	-	%
UVmax	Daily Maximum UV Index	v	-	Index (0–15)
LF	Surface Latent Heat Flux	-	v	MJ/m²
SF	Surface Sensible Heat Flux	-	v	MJ/m²
NR	Surface Net Thermal Radiation	-	v	MJ/m²
E	Evaporation	-	v	mm
PE	Potential Evaporation	-	v	mm

doi:10.6342/NTU202503010

2.3 Solar Power Generation Data

Solar power generation data are obtained from the open datasets of Taiwan Power Company. These include daily electricity generation [18], installed capacity [19], and location information [20] for 40 solar power plants between 2017 and 2021. The period of valid data varies across the power plants. The exact coordinates of each site are manually identified using Google Maps [21]. In this study, the 40 solar power plants are used for model training and testing. Their locations are shown in Figure 2.2.

To eliminate the influence of different installed capacities on comparison results, the daily generation values are converted to capacity factor using the following formula:

$$\text{Capacity Factor (0-1)} = \frac{\text{Daily Generation (kWh)}}{\text{24 (hour)} \times \text{Installed Capacity (kW)}} \tag{2.1}$$

doi:10.6342/NTU202503010

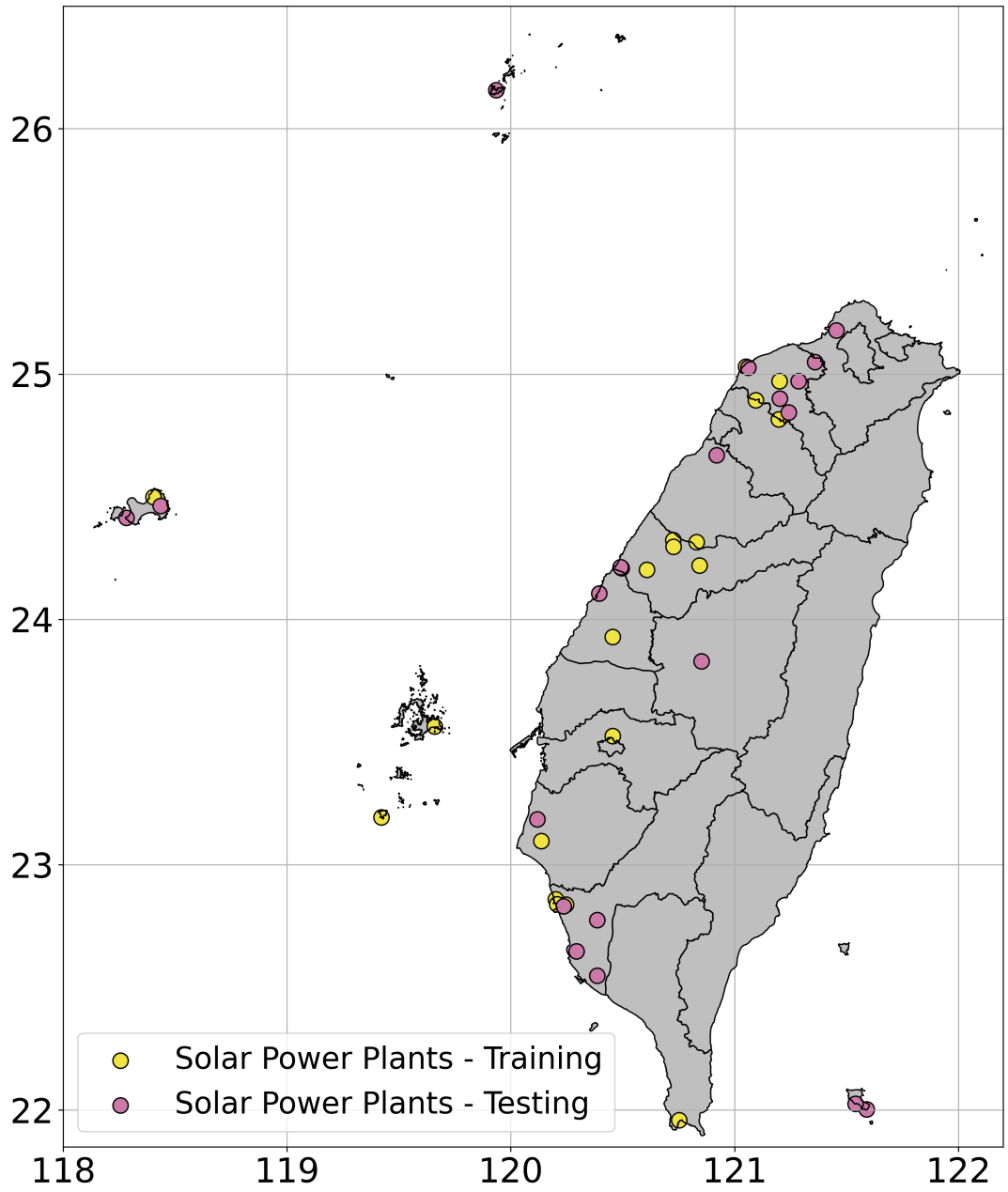


Figure 2.2: Geographic distribution of solar power plants in Taiwan. Plants used for model training and testing are marked separately.

doi:10.6342/NTU202503010

Chapter 3 Methodology

3.1 Overview

This study aims to examine how weather input influences the prediction outcome. A detailed description of the weather input is provided in Sections 1.1 and 1.2. As mentioned in Section 1.2, weather input refers to a combination of several estimated weather variables. To better reflect this concept, an expanded term *weather variables input combination* is occasionally used in this and the following chapters.

In practical applications, because ground weather stations and target sites are typically located at different places, the variables that constitute a weather input are often not directly available from original ground observations at the target prediction location. Instead, these variables are typically estimated using spatial interpolation and other processing methods based on nearby observations or gridded datasets. In response to this issue, consistent with the problem formulation described in Section 1.2, this study focuses on two key aspects of weather input design: (1) the spatial availability of nearby ground weather stations, and (2) the estimation methods used to transform original weather data into location-specific input.

These two critical aspects form the foundation of the experimental design in this

doi:10.6342/NTU202503010

study, and the overall workflow is illustrated in Figure 3.1.

First, estimation strategies are addressed in Section 3.2. Eight distinct weather input combinations are developed based on the two data sources: CWA (ground observation) and ERA5 (gridded data). These combinations represent various data usage approaches, including single-source input (e.g., CWA or ERA5), dual-source parallel input, and two data merging methods.

Second, ground station availability is addressed in Section 3.3. Thirteen distance-based filtering scenarios are designed to reflect different spatial distributions and data availability conditions. Among the eight input combinations, seven rely on ground-based data and are therefore affected by the removal of nearby stations. One input combination, using only ERA5 reanalysis, is independent of ground-based data and is thus not subject to the impacts of the station filtering.

In total, 92 synthetic datasets are generated (7 combinations × 13 filtering scenarios + 1 ERA5-only combination). All datasets are trained and evaluated using the same Long Short-Term Memory (LSTM) model architecture and training pipeline. This design ensures comparability across experiments.

doi:10.6342/NTU202503010

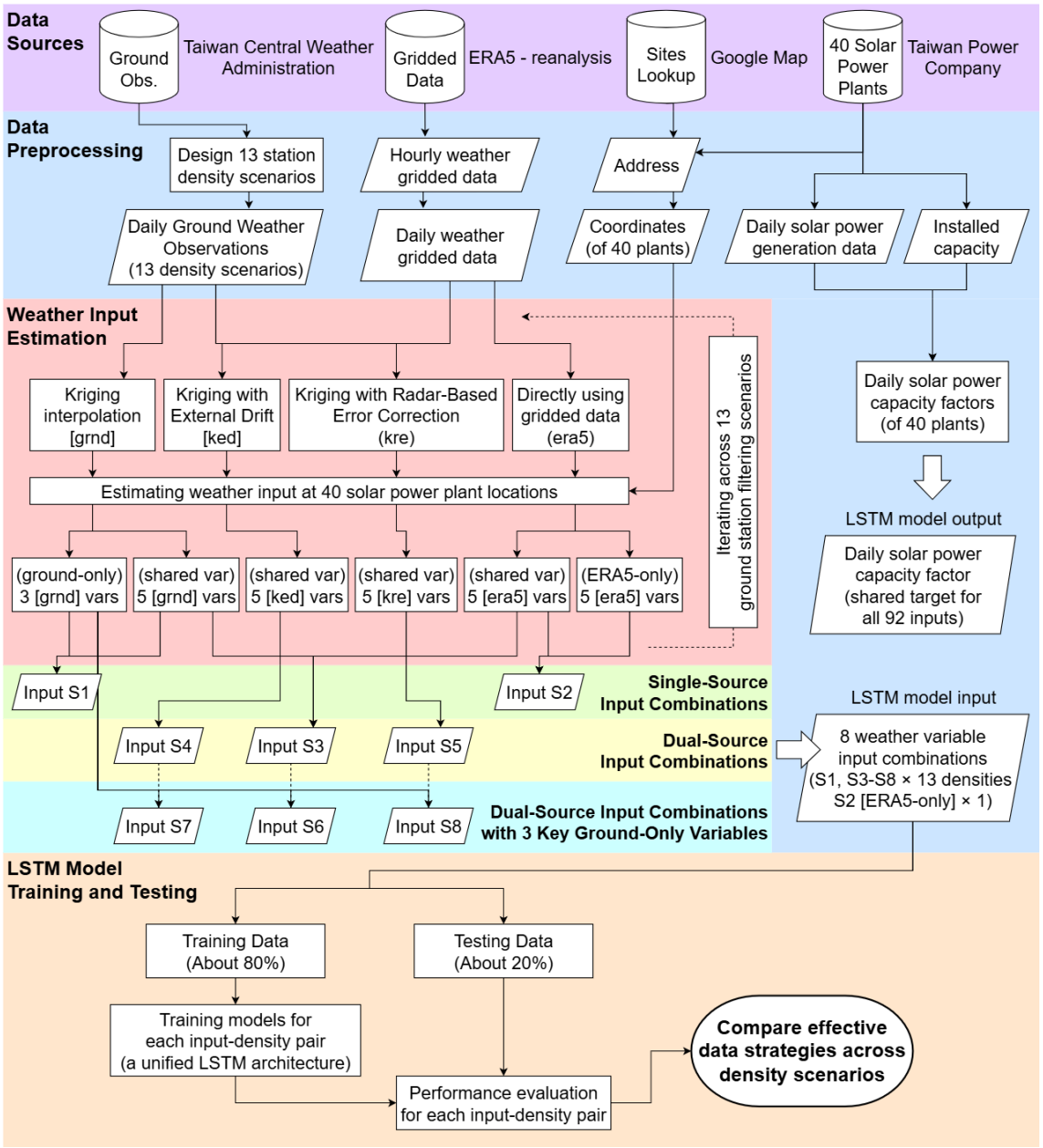


Figure 3.1: Overall workflow of the proposed methodological framework

doi:10.6342/NTU202503010

3.2 Weather Variables Input Combinations

In this study, the locations of 40 solar power plants in the dataset are used as the target sites for weather variables estimation. At each site, eight weather variables input combinations are constructed.

3.2.1 Weather Variable Estimation Methods

Across all weather variables input combinations, four methods are used to estimate weather variables. When a variable is derived solely from ground-based data, Kriging Interpolation (KG), a best linear unbiased estimator, is applied for spatial interpolation. When a variable is derived solely from ERA5-reanalysis data, the grid value corresponding to the target site is directly extracted. When a variable is estimated using both data sources, two data merging techniques extended from the KG method are employed: Kriging with External Drift (KED) and Kriging with Radar-Based Error Correction (KRE). These four estimation methods are described in detail in the following sections.

3.2.1.1 Kriging Interpolation

Kriging interpolation is a linear and unbiased estimation method. It was originally introduced by Matheron in 1963 [22] and has been widely used in geostatistics.

It estimates the value at an unknown location based on the observed values at known locations. In this study, Kriging is applied to interpolate weather conditions at solar power plant locations, which serve as both training and testing data points.

This method assumes that the spatial variability of a variable depends only on the dis-

doi:10.6342/NTU202503010

tance between locations. To model this spatial dependence, Kriging requires a variogram, which quantifies the degree of spatial correlation among observed data points.

To construct the variogram, the empirical variogram is first calculated from the ground observation data using the following formula:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \tag{3.1}$$

where $\gamma(h)$ is the semivariance at distance h, $z(x_i)$ is the observed value at location x_i, and $N(h)$ is the number of pairs separated by distance h.

In this study, the empirical variogram is computed using the `gstools` package, and an exponential model is then fitted to capture the underlying spatial pattern. Figure 3.2 shows an example of the empirical and fitted variogram.

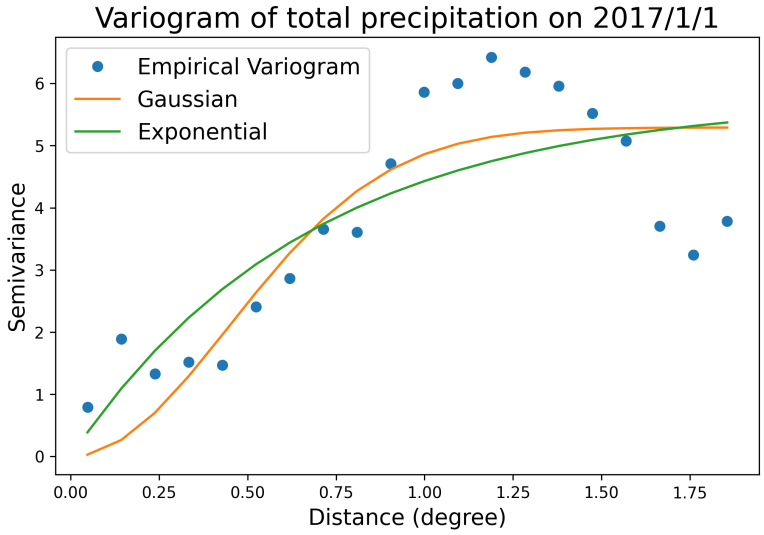


Figure 3.2: Example of empirical variogram and fitted models for precipitation on January 1, 2017.

The fitted variogram is then applied to the Kriging equations to estimate spatial weights. Kriging solves a system of equations to determine the weights λ_i assigned to

doi:10.6342/NTU202503010

each known observation. The Kriging matrix system is:

$$\mathbf{M}\boldsymbol{\lambda} = \boldsymbol{\gamma} \tag{3.2}$$

where:

- $\mathbf{M}$ is the Kriging matrix constructed using the fitted variogram model,
- $\boldsymbol{\lambda}$ is the vector of Kriging weights,
- $\boldsymbol{\gamma}$ is the vector of semivariances between each known point and the unknown point.

Their elements are shown below:

$$\underbrace{\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}}_{\mathbf{M}} \cdot \underbrace{\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix}}_{\boldsymbol{\lambda}} = \underbrace{\begin{bmatrix} \gamma_{1o} \\ \gamma_{2o} \\ \vdots \\ \gamma_{no} \\ 1 \end{bmatrix}}_{\boldsymbol{\gamma}} \tag{3.3}$$

Here, γ_{ij} is the semivariance between the i-th and j-th known locations, calculated by applying their distance to the fitted variogram model. γ_{io} in the right-hand side vector represents the semivariance between the i-th known location and the unknown location x_o.

The system is solved to obtain the weights as follows:

doi:10.6342/NTU202503010

$$\boldsymbol{\lambda} = \mathbf{M}^{-1}\boldsymbol{\gamma} \tag{3.4}$$

This gives the weight vector $\boldsymbol{\lambda}$, which represents the spatial influence of each observed point on the unknown point.

Once the weights are obtained, the value at the unknown location x_o is estimated by the weighted sum:

$$z(x_o) = \sum_{i=1}^{n} \lambda_i \cdot z(x_i) \tag{3.5}$$

Note that the μ in weight vector $\boldsymbol{\lambda}$ does not participate in the weighted sum.

where:

- $z(x_o)$ is the interpolated value at unknown location x_o,
- λ_i is the Kriging weight for the i-th observed location,
- $z(x_i)$ is the observed value at location x_i.

The value of $z(x_o)$ thus represents the estimated value of a weather variable at the target site. Considering that the ground variables used in this study have non-negative nature, a correction method proposed by Deutsch (1996) [23] is applied to ensure that the Kriging estimates for each ground variable remain non-negative.

Variables estimated using this method are denoted with the subscript `grnd` (indicating ground-based estimates).

doi:10.6342/NTU202503010

3.2.1.2 Kriging with External Drift

Kriging with External Drift (KED) is the first data merging method used in this study. It is an extension of the standard Kriging interpolation technique that incorporates external drift variables (in this study, the ERA5 gridded data) into the Kriging system. Other applications and detailed discussions of KED can be found in Hudson and Wackernagel (1994) [24].

The KED system modifies the Kriging equations (Equation 3.2) to account for the external drift, and in this study, its matrix system is expressed as:

$$\underbrace{\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 & f(x_1) \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 & f(x_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 & f(x_n) \\ 1 & 1 & \cdots & 1 & 0 & 0 \\ f(x_1) & f(x_2) & \cdots & f(x_n) & 0 & 0 \end{bmatrix}}_{\mathbf{M}_{\text{KED}}} \cdot \underbrace{\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \\ \nu \end{bmatrix}}_{\boldsymbol{\lambda}_{\text{KED}}} = \underbrace{\begin{bmatrix} \gamma_{1o} \\ \gamma_{2o} \\ \vdots \\ \gamma_{no} \\ 1 \\ f(x_o) \end{bmatrix}}_{\boldsymbol{\gamma}_{\text{KED}}} \tag{3.6}$$

where:

- $f(x_i)$ represents the external drift information at the i-th known location; in this study, it corresponds to the ERA5 gridded value at the location of the i-th observed point,
- $f(x_o)$ is the external drift value at the unknown location x_o,
- μ, ν are auxiliary multipliers. These terms do not participate in the final weighted

doi:10.6342/NTU202503010

sum for estimating the value at the unknown location,

- The notations γ_{ij}, γ_{io}, and λ_i are consistent with those defined in the Kriging interpolation section (Section 3.2.1.1).

As with Kriging interpolation, the system is solved to obtain the weights as follows:

$$\boldsymbol{\lambda}_{\text{KED}} = \mathbf{M}_{\text{KED}}^{-1}\boldsymbol{\gamma}_{\text{KED}} \tag{3.7}$$

This gives the weight vector $\boldsymbol{\lambda}_{\text{KED}}$, which represents the influence of each observed point on the unknown point, while also incorporating external drift information.

Once the weights are obtained, the value at the unknown location x_o is estimated by the same weighted sum as in Kriging interpolation:

$$z(x_o) = \sum_{i=1}^{n} \lambda_{i,\text{KED}} \cdot z(x_i) \tag{3.8}$$

where the notations are consistent with those in Equation 3.5, except that the weights $\lambda_{i,\text{KED}}$ are derived from the KED system.

As with the Kriging method, a correction method proposed by Deutsch (1996) [23] is applied to ensure that the interpolated values of non-negative variables remain valid.

Variables estimated using this method are denoted with the subscript `ked`.

doi:10.6342/NTU202503010

3.2.1.3 Kriging with Radar-Based Error Correction (KRE)

Kriging with Radar-Based Error Correction (KRE) is the second data merging method employed in this study (Sinclair and Pegram, 2005 [25]). This approach extends the Kriging method by incorporating correction based on the ERA5 gridded data (in this study, the radar-based error correction is replaced by ERA5 gridded data error correction).

First, ERA5-reanalysis gridded values at ground weather station locations are interpolated using Kriging. The difference between the original gridded value and the Kriging interpolated gridded value is then calculated. A larger difference indicates poorer Kriging quality at that location. Finally, the Kriging interpolated ground observation value at the target site is adjusted by adding this difference, resulting in the KRE estimate. The procedure for KRE interpolation involves the following steps:

1. **Gridded Error Estimation**

   The difference between the original gridded value (ERA_{grid}) and the Kriging interpolated gridded value (ERA_{Kriging}) is computed at the unknown point (the target site). This step captures the gridded error correction term:

$$\Delta ERA = ERA_{\text{grid}} - ERA_{\text{Kriging}}$$

2. **Error-Corrected Observation Interpolation**

   The gridded error correction term (ΔERA) is added to the Kriging interpolated ground observation value (G_{Kriging}), which is computed using the method described in Section 3.2.1.1, resulting in the final corrected value (KRE) at the target location:

$$G_{\text{KRE}} = G_{\text{Kriging}} + \Delta ERA$$

doi:10.6342/NTU202503010

Variables estimated using this method are denoted with the subscript `kre`.

3.2.1.4 Directly Using Gridded Data

The gridded data used in this study are obtained from three ERA5 reanalysis datasets. The extraction method is relatively straightforward: for each grid cell, the value at its central coordinate is assigned as the representative value. The grid cell containing the location of each solar power plant is identified, and the corresponding grid value is used as the weather variable estimate for that location. If no ERA5 grid value is available at the location, the nearest grid point is used instead.

Variables estimated using this method are denoted with the subscript `era5` (for ERA5-reanalysis data).

3.2.2 Variables Using and Strategy for Each Combinations

The design strategies for the eight weather input combinations can be categorized into three types. The specific weather variables used and their corresponding estimation methods are summarized in Tables 5–7. Abbreviations of the weather variables are defined in Table 2.1, and the estimation method for each variable is indicated by the subscript attached to the variable abbreviation.

Single-source input combinations: Combinations S1 and S2 utilize weather data solely from one source. S1 is from the CWA ground-based observation dataset, and S2 is from the ERA5-reanalysis dataset. All available variables from the respective datasets are included. The detailed weather variables and their estimation methods are listed in Table 3.1.

doi:10.6342/NTU202503010

Table 3.1: Input combinations of single weather data source

No.	Weather Variables and Estimation Methods
S1	T_{grnd}, RH_{grnd}, P_{grnd}, GR_{grnd}, TC_{grnd}, SD_{grnd}, SP_{grnd}, $UVmax_{grnd}$
S2	T_{era5}, RH_{era5}, P_{era5}, GR_{era5}, TC_{era5}, LF_{era5}, SF_{era5}, NR_{era5}, E_{era5}, E_{era5}

Dual-source input combinations: Combinations S3, S4, and S5 are designed to investigate how to simultaneously utilize ground and gridded weather data. These combinations include the five weather variables that are shared between the CWA ground dataset and the ERA5-reanalysis dataset.

The initial focus of this study is to explore the potential of data merging in solar power potential prediction. This study first examines how dual weather data sources have traditionally been used in the field. Previous studies, such as [8] and [9], typically treat ground-based and gridded weather data as separate input variables for the models. This study refers to this approach as *parallel input*. To evaluate whether merging these two data sources can lead to different outcomes, three dual-source input combinations (S3, S4, and S5) are designed, representing pre- and post-merging weather input combinations.

S3 represents parallel input, a strategy commonly used in previous studies, where the five variables derived solely from CWA ground observations and the five from ERA5-reanalysis are combined to form a ten-variable input combination.

S4 and S5 represent data merging methods using KED (Kriging with External Drift) and KRE (Kriging with Radar-Based Error Correction), respectively, to merge the five matched variable pairs from S3. The details of these combinations are presented in Table 3.2.

Dual-source input + 3 key ground-only variables: This strategy extends the dual-source design by including three key weather variables that are exclusively available from

doi:10.6342/NTU202503010

Table 3.2: Dual-source input combinations using shared variables from both datasets

No.	Weather Variables and Estimation Methods
S3	T_{grnd}, RH_{grnd}, P_{grnd}, GR_{grnd}, TC_{grnd}, T_{era5}, RH_{era5}, P_{era5}, GR_{era5}, TC_{era5}
S4	T_{ked}, RH_{ked}, P_{ked}, GR_{ked}, TC_{ked}
S5	T_{kre}, RH_{kre}, P_{kre}, GR_{kre}, TC_{kre}

the CWA dataset: sunshine duration, sunshine rate, and maximum UV index. These variables were identified in Feng's study (2023) [11] as highly correlated with solar power generation. Since these variables are unavailable in the ERA5 dataset, they are excluded from the previous dual-source combinations (S3–S5). To assess whether omitting these variables would significantly affect prediction performance, combinations S6, S7, and S8 are developed by adding the three ground-only variables to S3, S4, and S5 using a parallel input strategy. The detailed weather variables and estimation methods are provided in Table 3.3.

Table 3.3: Dual-source input combinations with additional ground-only variables

No.	Weather Variables and Estimation Methods
S6	T_{grnd}, RH_{grnd}, P_{grnd}, GR_{grnd}, TC_{grnd}, T_{era5}, RH_{era5}, P_{era5}, GR_{era5}, TC_{era5}, SD_{grnd}, SP_{grnd}, $UVmax_{grnd}$
S7	T_{ked}, RH_{ked}, P_{ked}, GR_{ked}, TC_{ked}, SD_{grnd}, SP_{grnd}, $UVmax_{grnd}$
S8	T_{kre}, RH_{kre}, P_{kre}, GR_{kre}, TC_{kre}, SD_{grnd}, SP_{grnd}, $UVmax_{grnd}$

3.3 Ground Weather Station Filtering Scenarios

Benefiting from the high density of ground weather stations in Taiwan, this study proposes a customized method to construct 13 ground weather station filtering scenarios. The goal is to directly assess the impact of the distance between ground stations and the target estimation site. To simulate varying levels of ground observation availability, we progressively remove nearby ground stations in a stepwise manner. This procedure is performed individually for each solar power plant and iteratively applied to the original

doi:10.6342/NTU202503010

ground observation dataset.

Global solar radiation is selected as the reference variable for defining the distance thresholds, as it is a key variable for solar power generation and is available in both ground-based and gridded datasets. For each solar power plant, 13 filtering scenarios are created. The 0th scenario includes all available ground observations without any removal. For the 1st to 12th scenarios, the distance thresholds are defined based on the distance to the 1st, 2nd, ..., 12th nearest ground radiation stations, respectively. It is important to note that the actual distance threshold for the same scenario level may vary between power plants.

Using the corresponding threshold for each scenario, all observation data from the ground station within that distance are removed from the dataset. The original data and the filtered datasets for each threshold thus constitute the 0th to 12th ground weather station filtering scenario datasets for each power plant.

3.4 Unified Potential Prediction Model Training Process

To ensure fairness and comparability across eight input combinations, a unified prediction model and training configuration is adopted. The model used is a Long Short-Term Memory (LSTM) neural network, with its architecture and hyper-parameters based on the optimized design proposed by Feng (2023) [11], as illustrated in Figure 3.3.

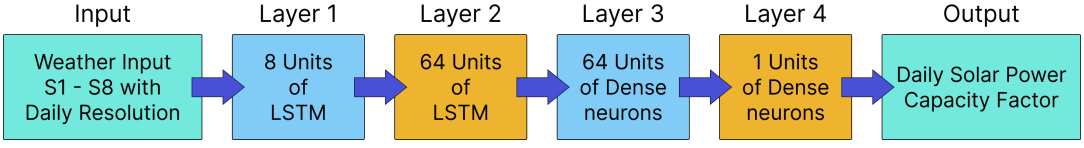


Figure 3.3: Architecture of the LSTM model used in this study.

Each input combination is split into a training set and a testing set in a fixed ratio (approximately 4:1), with the training set further divided into training and validation subsets.

doi:10.6342/NTU202503010

To ensure consistent data ordering across all combinations, a fixed random seed is applied during all shuffling steps before training. Additionally, to ensure consistent input length, days with zero generation, missing values (NaN), or anomalous weather variable values are removed. These steps are employed to help reduce the influence of non-weather factors. The actual solar power capacity factor is used as the shared prediction target for all input combinations.

Except for input combination S2, which relies only on ERA5 gridded data, all other seven input combinations depend on ground-based observations for variables estimation. Therefore, each of these combinations is iteratively trained and tested under 13 different ground station filtering scenarios. This design enables a detailed comparison of how input combinations perform under varying levels of ground observation availability.

Following the training process, model performance is evaluated during the testing phase to assess the impact of different weather input combinations.

3.5 Evaluation of Prediction Performance under Different Weather Input Combination

To evaluate the performance of each weather input combination, this study directly compares the model predictions with the actual capacity factors recorded at solar power plants. The goal is to quantify the prediction accuracy of each input combination and identify the relative strengths and weaknesses among the eight input combinations.

This evaluation is performed across all 13 ground weather station filtering scenarios as well, ensuring that the influence of observation availability is considered. The following

doi:10.6342/NTU202503010

subsection introduces the error metrics used in this comparison.

3.5.1 Evaluation Metric

A commonly used statistical metric is employed in this study to assess the deviation between model predictions and the true values of the solar power capacity factor. This metric is Mean Squared Error (MSE), defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.9}$$

where $\hat{y}_i$ denotes the predicted capacity factor at time step i for a specific solar power plant, and y_i denotes the corresponding true value. The total number of time steps is denoted by n.

These metrics are calculated individually for each of the 20 testing plants by comparing the predicted and true values over the entire testing time period. Each metric value represents the plant-level prediction error across all time steps.

doi:10.6342/NTU202503010

Chapter 4 Results and Discussion

4.1 Overview

This study evaluate solar power potential prediction at the individual solar power plant level. Specifically, the error metrics are calculated separately for each of the 20 testing plants by comparing the predicted and true historical values (of solar power capacity factor) over the entire testing time period.

To investigate the influence of ground station availability on the prediction performance, at chapter 3, 13 ground weather station filtering scenarios are designed. These scenarios are defined based on the distance to the k-th nearest ground station that provides global solar radiation data. For each solar power plant, the distance thresholds in each scenario vary depending on its spatial relationship to the surrounding ground stations.

In each scenario, model errors (e.g., MSE) are computed individually for all 20 testing power plants. The corresponding distance to the nearest available ground station is also recorded. These paired values (distance, error) from all 13 ground weather station filtering scenarios are combined and grouped according to distance intervals.

Therefore, for each input combination except S2 (which includes only gridded data), a total of 260 paired samples (20 plants × 13 filtering scenarios) of capacity factor error

doi:10.6342/NTU202503010

and corresponding nearest ground station distance are obtained. These paired data are aggregated into distance bins of 25 km and visualized using box plots to display the error distributions.

The results are presented following this structure: (1) Comparison between single-source weather inputs (S1 and S2), (2) Comparison of inputs using different combinations of ground observations and ERA5 reanalysis data (S3–S5), (3) Investigation of the effects of including or excluding three key ground variables: sunshine rate, sunshine duration, and maximum UV index, by comparing pairs such as S3 and S6, S4 and S7, and S5 and S8.

Note that the purely gridded data case, S2, is unaffected by the availability of ground observations. As a result, the prediction errors from the S2 case are identical across all observational filtering scenarios. Therefore, the first quartile (Q1), median, and third quartile (Q3) of the 20 testing plants' errors are calculated and presented as three constant reference lines in each comparison.

4.2 Comparison of Single-Source Weather Inputs: S1 vs. S2

This study first compares the prediction results of two single-source input combinations, S1 and S2, to examine the basic performance of ground-based and gridded weather data in solar power potential prediction. The results of S1 (ground-based observations) and S2 (ERA5 reanalysis gridded data) serve as baseline cases and provide a reference for evaluating dual-source input strategies discussed in subsequent sections.

doi:10.6342/NTU202503010

The comparison is presented in Figure 4.1. The results show that S1 outperforms S2 under short distance scenarios (0–75 km). However, as the distance to the nearest ground station exceeds approximately 50–75 km, the performance of the two begins to reverse. At even lower observation densities, the error of S1 increases at a slower rate and eventually stabilizes.

These findings confirm that ground-based weather data offer higher accuracy in dense observational settings, aligning with the conventional view that such data are generally more reliable. However, the performance of ground data drops significantly long distance scenarios, presumably due to the limited spatial coverage of ground observations. In such cases, ERA5 can serve as a useful alternative to mitigate this limitation. This trend align with the expected characteristics of ground-based versus gridded weather datasets.

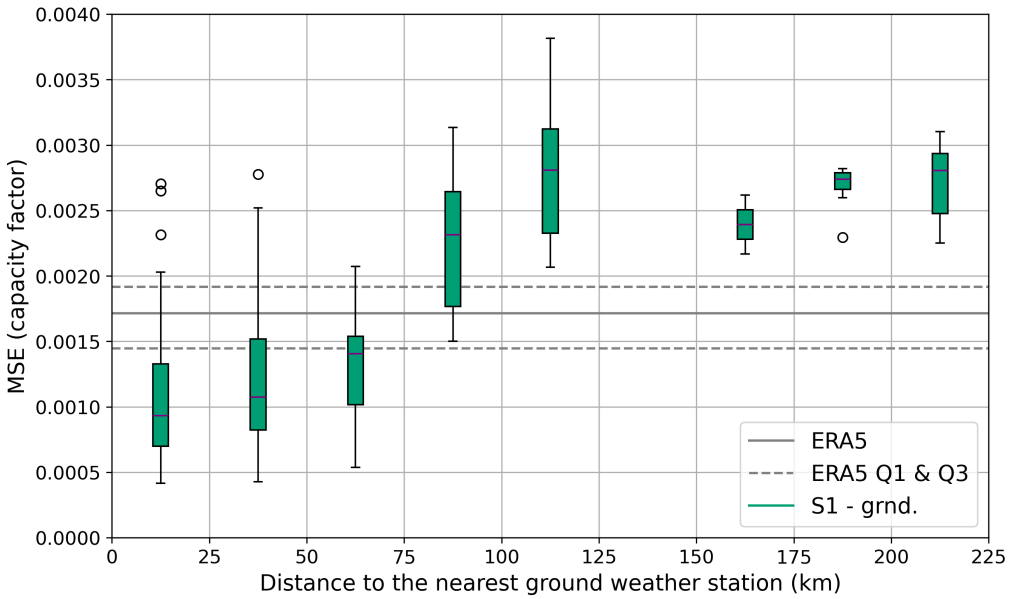


Figure 4.1: Comparison of MSE distributions for S1 (ground-based) and S2 (ERA5) across varying distances to the nearest ground weather station.

doi:10.6342/NTU202503010

4.3 Comparison of Dual-Source Weather Input Strategies (S3 vs. S4 vs. S5)

This section compares three dual-source input strategies: parallel input (S3), KED-based data merging (S4), and KRE-based data merging (S5). Figure 4.2 illustrates the MSE distributions of these strategies under varying nearest station distance scenarios.

Under short distance conditions, KED (S4) performs slightly better than the other two dual-source approaches. However, its advantage over using only ground-based data (S1) is marginal, suggesting limited benefit from merging ERA5 and ground data when observations are abundant.

As the nearest observation distance increases, the advantage of dual-source strategies becomes more evident. In the distance group of 100–125 km to the nearest ground station, all three dual-source strategies (S3, S4, and S5) start to outperform the ground-only baseline (S1), with parallel input (S3) showing the smallest errors, followed by KED (S4) and KRE (S5).

In the longest-distance groups (≥150 km), prediction performance becomes less consistent across strategies. KED performs on par with, or slightly worse than, the ground-only approach. The parallel input strategy exhibits inconsistent behavior. It ranks among the best in some distance intervals but among the worst in other. Notably, KRE (S5) remains the most robust, consistently achieving the lowest prediction errors in the longest-distance groups.

Overall, in the long distance range (≥75 km), parallel input (S3) reduces prediction error compared to the ground-only case in most intervals, except for the 150–175 km group.

doi:10.6342/NTU202503010

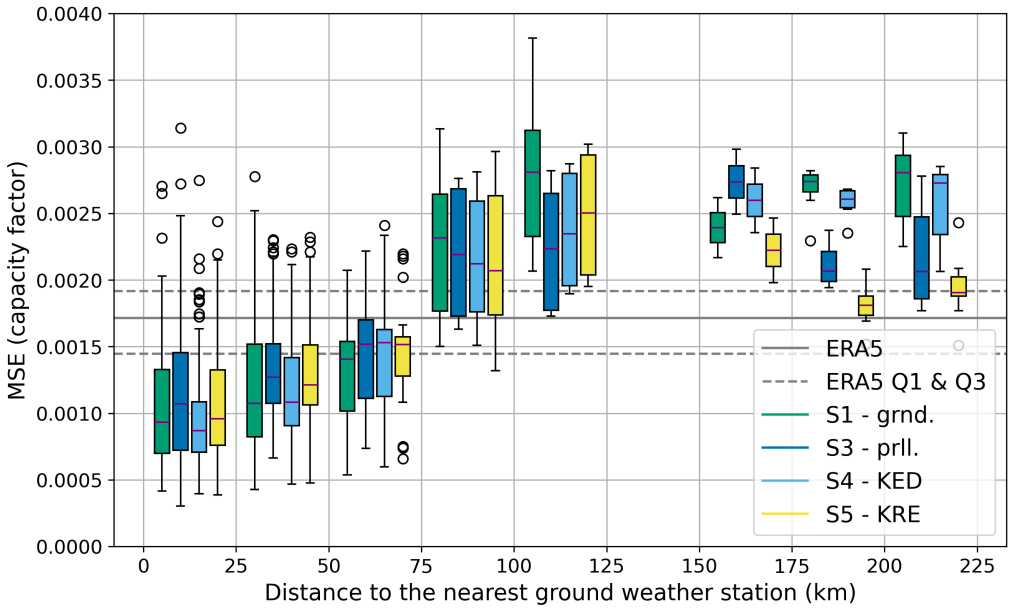


Figure 4.2: Comparison of MSE distributions for S3 (Parallel Input), S4 (KED Data Merging), and S5 (KRE Data Merging) across varying distances to the nearest ground weather station.

KRE (S5) consistently outperforms the other strategies, except for a slight underperformance in the 100–125 km interval. These results underscore the particular effectiveness of the KRE approach in maintaining predictive accuracy when observational coverage is sparse. Parallel input (S3) also shows relatively stable performance across most distance intervals, indicating its potential as a simple yet robust dual-source strategy under long-distance conditions.

However, it is worth noting that none of the three dual-source strategies outperform the ERA5-only baseline (S2). Despite their advantages over ground-only data (S1) under sparse observation conditions, these approaches still fall short of the overall performance provided by the gridded ERA5 data across long-distance scenarios.

doi:10.6342/NTU202503010

4.4 Impact of Key Ground-Only Variables (S3 vs. S6, S4 vs. S7, S5 vs. S8)

Following the correlation analysis proposed by Feng (2023, p.23)[11], five variables were found to be highly correlated with the solar power capacity factor (Spearman's absolute value > 0.6), listed in descending order of correlation strength: GR (global solar radiation), SP (sunshine percentage), SD (sunshine duration), TC (total cloud cover), and UVmax (daily maximum UV index). Among them, SP, SD, and UVmax are only available from ground-based observations (see Table 2.1 for data source summary) and are therefore excluded from the dual-source input combinations S3, S4, and S5 (see Table 3.2 for these input design).

This section investigates whether incorporating the three ground-only but highly correlated variables (SP, SD, and UVmax) into each dual-source input strategy can further enhance the accuracy of solar power potential prediction.

Figure 4.3 compares S3 vs. S6 (Figure 4.3a), S4 vs. S7 (Figure 4.3b), and S5 vs.S8 (Figure 4.3c), illustrating the impact of adding these three variables into strategies S3–S5 (detailed input design can be found in Table 3.3).

In most cases, the inclusion of these variables leads to improved prediction accuracy. One exception occurs in the parallel input strategy (S3 vs. S6; Figure 4.3a), where the addition of these variables unexpectedly increases prediction error. A review of the training logs for this case do not reveal clear anomalies compared to other input combinations, implying that the added ground variables may have inadvertently introduced noise or redundancy under this configuration.

doi:10.6342/NTU202503010

Another important observation is that for the KRE strategy (S5 vs. S8; Figure 4.3c), the inclusion of these variables almost improves model performance across all short- and long-distance scenarios. Notably, in the 175–200 km distance range, the prediction accuracy (of S8) approaches the median performance of the ERA5-only case (S2), showing the value of these ground-only variables in enhancing solar power potential prediction.

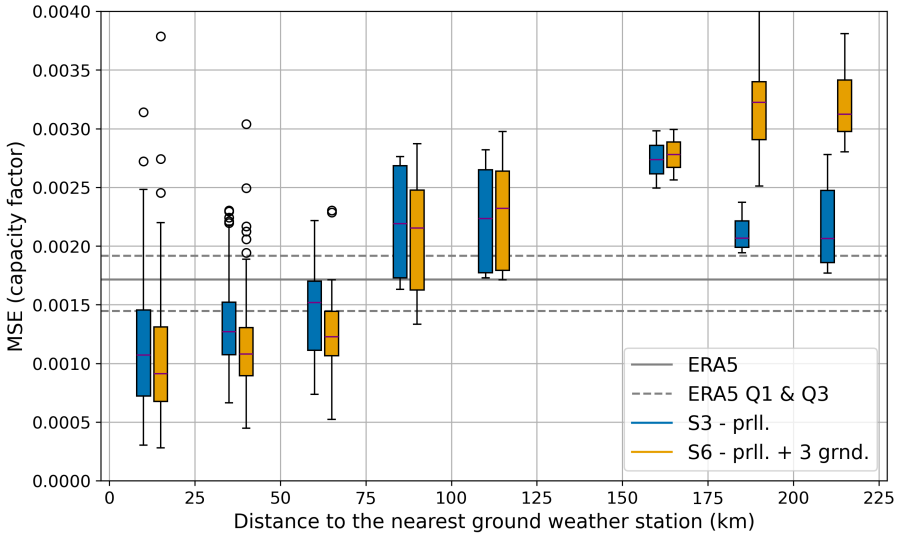
4.5 Final Comparison: S1 Ground-Only vs. S6–S8 Dual-Source Inputs with Key Ground-Only Variables

In Section 4.3, we initially compares three dual-source input strategies, and in Section 4.4, we demonstrates that incorporating three key ground-only variables: sunshine percentage (SP), sunshine duration (SD), and daily maximum UV index (UVmax) can provide positive benefits for solar power potential prediction. This section presents a final comparison between the two single-source baseline combinations (S1 and S2) and the dual-source combinations (S6, S7, and S8) that include both data types along with the three key ground variables.

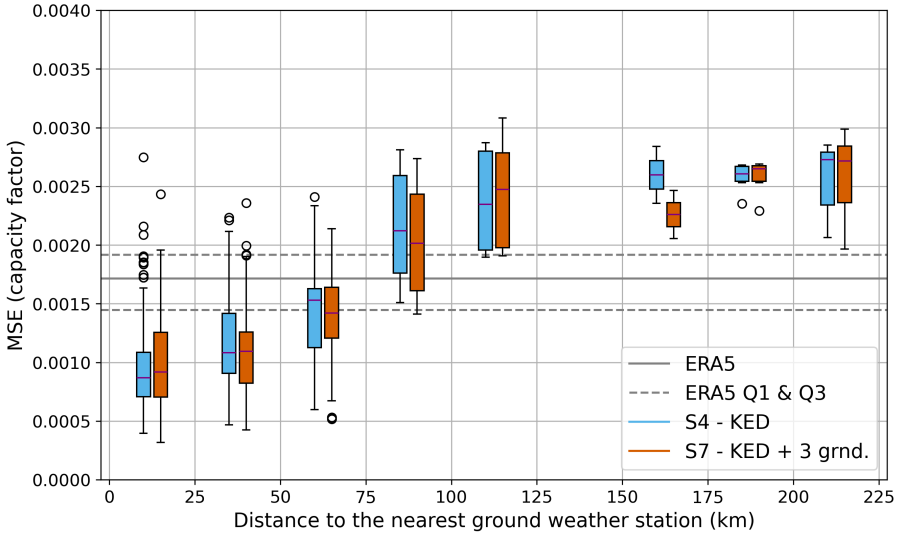
The comparison results are shown in Figure 4.4 and the results show:

- **Short-distance scenarios** (0–75 km to the nearest ground station): The three strategies: parallel input, KED, and KRE perform similarly. However, none of them demonstrate a clear improvement over the ground-only approach (S1).
- **Medium-distance scenarios** (75–125 km): The parallel input strategy achieves the best performance overall, followed by KED.
- **Longest-distance scenarios** (150–225 km): KRE shows a clear advantage in pre-

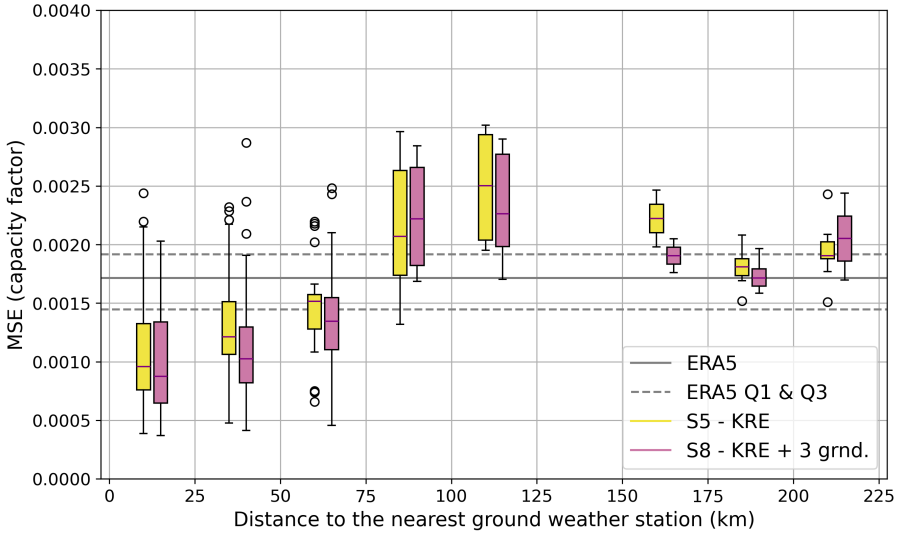
doi:10.6342/NTU202503010



(a) MSE distributions for S3 vs. S6.



(b) MSE distributions for S4 vs. S7.



(c) MSE distributions for S5 vs. S8.

Figure 4.3: Comparison of MSE distributions for key ground-only variable impact scenarios across varying distances to the nearest ground weather station.

doi:10.6342/NTU202503010

dictive accuracy.

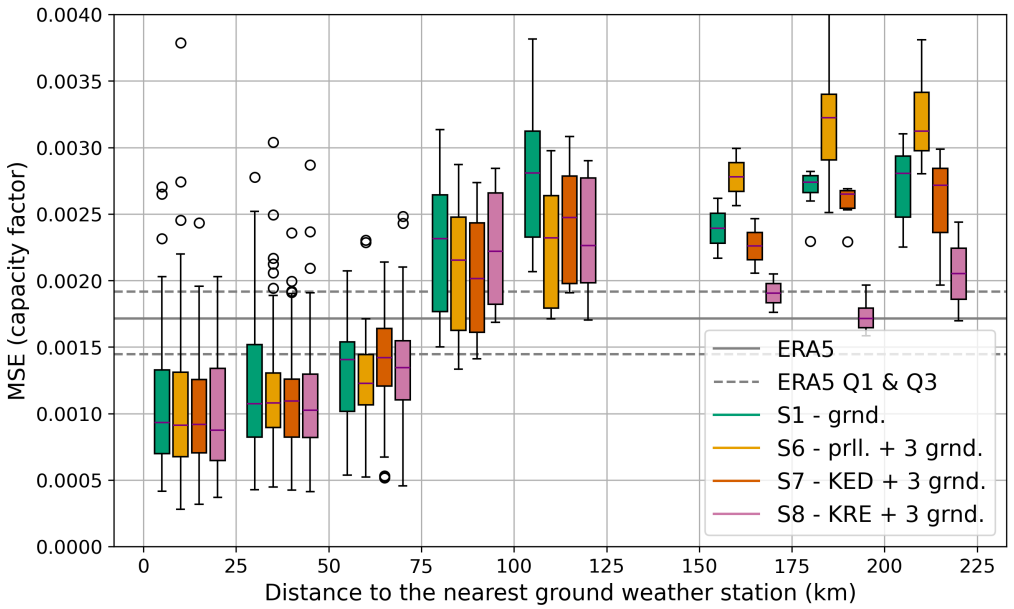


Figure 4.4: Comparison of MSE distributions for S1 (Ground Observations), S6 (Parallel Input + 3 Key Ground-Only Variables), S7 (KED Data Merging + 3 Key Ground-Only Variables), and S8 (KRE Data Merging + 3 Key Ground-Only Variables) across varying distances to the nearest ground weather station.

4.6 Consideration of Model Uncertainty and Data Bias (Extended in Appendix)

The performance comparison in this study is based on a Long Short-Term Memory (LSTM) model. While this provides a flexible and powerful framework for solar potential prediction, it is important to acknowledge the inherent limitations of AI-based, data-driven models. These include potential biases introduced by limited or imbalanced training data, as well as model uncertainty stemming from the stochastic nature of neural network training.

In particular, data-driven models are highly sensitive to the quality and quantity of the input data. This study uses a total of just over 30,000 data points of solar power generation

doi:10.6342/NTU202503010

paired with corresponding weather input for model training. Although this represents a valuable dataset, it remains relatively limited in size, especially considering the variability and complexity of both solar output and weather conditions.

To illustrate the possible effects of these limitations, the study employs an exploratory approach to provide preliminary insights. For clarity and focus, the discussion is presented separately in Appendix A.

doi:10.6342/NTU202503010

Chapter 5 Conclusion

This study aims to evaluate the performance of different weather inputs for solar potential prediction and to examine their applicability under varying ground station filtering scenarios. The key conclusions of this study are summarized as follows.

- **Single data source: Performance reversal occurs beyond 75 km.** When using only one type of data (either ground-based or gridded), ground-based observations yield higher prediction accuracy when the nearest ground weather station is within 75 km. However, as this distance increases, their performance deteriorates rapidly. This trend aligns with the general understanding that ground-based data are more accurate but spatially limited, while gridded data are less precise but offer broader and more uniform spatial coverage.

The performance reversal observed around the 50–75 km highlights the effective spatial range within which ground-based observations remain reliable for solar power potential prediction.

- **Key ground-only variables: Positive influence on prediction performance.** In this study, the three highly correlated ground-only variables (sunshine percentage, sunshine duration, and maximum UV index) is found to improve prediction accu-

doi:10.6342/NTU202503010

racy in most cases, as shown in Figure 4.3.

Building upon the above findings, discussion on dual data sources is narrowed to focus on the critical distance range of 50–75 km. Although two groups of dual-source input combinations are examined in this study (S3–S5 and S6–S8), the subsequent analysis focuses on S6–S8, which include the three key ground-only variables. As these variables are identified as effective predictors in the preceding section, their inclusion offers a more representative basis for evaluating the performance of dual-source strategies, as shown in Figure 4.4.

- **Dual data sources: Integrating ground and gridded data improves over gridded-only input, with performance comparable to ground-only.** Within the critical distance range of 50–75 km, integrating ground-based observations (CWA dataset) with gridded data (ERA5-reanalysis) enhances prediction accuracy compared to using gridded data alone (S2). All three dual-source strategies: parallel input (S6), KED (S7), and KRE (S8), achieve lower median MSE than S2, with respective reductions of approximately 0.0004, 0.0003, and 0.00035 in the MSE of solar capacity factor predictions. Parallel input exhibits the best performance, followed by KRE.

In summary, this study identifies that the effective range of ground-based observations for solar power potential prediction is approximately 75 km. Beyond this range, ERA5 gridded data show better stability and applicability, making them an advantageous strategy when ground observations are limited. Within this critical range (50–75 km), dual-source strategies outperforms gridded-only input; however, their performances do not show a clear advantage over ground-only input. Nevertheless, the results collectively

doi:10.6342/NTU202503010

confirm the positive impact of incorporating ground-based observations within their effective spatial range. In addition, three key ground-only variables: sunshine percentage, sunshine duration, and maximum UV index are found to have the potential to enhance prediction accuracy.

From a practical standpoint, this study addresses the concern about weather input raised in Section 1.1, namely the limited number and spatial coverage of ground-based weather stations in many regions. While some recent studies rely solely on gridded datasets or even attempt to incorporate multiple gridded sources, the findings here highlight the continued value of incorporating ground-based observations. The results demonstrate that ground-based data within a 50–75 km range can substantially enhance prediction accuracy, even without dense station coverage. Therefore, rather than relying solely on gridded datasets, practitioners are encouraged to incorporate available ground-based observations whenever possible.

While this study explores the use of both ground-based and gridded weather data, the results do not show a clear improvement over using single-source alone. Within the 75 km effective range, parallel input (S6) and data merging methods such as KED (S7) and KRE (S8) show no clear advantage over the ground-only input (S1). Beyond this range, the gridded-only input (S2) outperforms the dual-source methods (S6 – S8). In contrast, previous studies have reported performance gains from the ground-based and gridded data integration. For example, Journée et al. (2012) [26] showed that KED reduced the daily global horizontal irradiation (GHI) estimation error (MAE) by more than 20% compared to ground-based interpolation (KG). Similarly, Qin et al. (2022) [8] demonstrated that a deep learning model trained on both satellite images and ground observations achieved 10–20% lower MAE in one- to six-hour ahead GHI forecasts compared to a model trained

doi:10.6342/NTU202503010

solely on ground data.

These contrasting findings suggest that the effective integration of ground and gridded data, and the true potential of data merging, still remain open questions, requiring further exploration.

doi:10.6342/NTU202503010

References

[1] IEA. Renewables 2024. https://www.iea.org/reports/renewables-2024, 2024. IEA, Paris. Licence: CC BY 4.0.

[2] Nurul Jannah, Teddy Surya Gunawan, Siti Hajar Yusoff, Mohd Shahrin Abu Hanifah, and Siti Nadiah Mohd Sapihie. Recent advances and future challenges of solar power generation forecasting. IEEE Access, 12:168904–168924, 2024.

[3] Naylene Fraccanabbia and Viviana Cocco Mariani. Evaluating machine learning in short-term forecasting time series of solar power. Brazilian Journal of Applied Computing, 13(2):105–112, May 2021.

[4] Jae Heo, Kwonsik Song, SangUk Han, and Dong-Eun Lee. Multi-channel convolutional neural network for integration of meteorological and geographical features in solar power forecasting. Applied Energy, 295:117083, 2021.

[5] Tao Fan, Tao Sun, Hu Liu, Xiangying Xie, and Zhixiong Na. Spatial-temporal genetic-based attention networks for short-term photovoltaic power forecasting. IEEE Access, 9:138762–138774, 2021.

[6] Rakesh Mondal, Surajit Kr Roy, and Chandan Giri. Solar power forecasting using domain knowledge. Energy, 302:131709, 2024.

doi:10.6342/NTU202503010

[7] Dazhi Yang, Jan Kleissl, Christian A. Gueymard, Hugo T.C. Pedro, and Carlos F.M. Coimbra. History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining. Solar Energy, 168:60–101, 2018. Advances in Solar Resource Assessment and Forecasting.

[8] Jun Qin, Hou Jiang, Ning Lu, Ling Yao, and Chenghu Zhou. Enhancing solar pv output forecast by integrating ground and satellite observations with deep learning. Renewable and Sustainable Energy Reviews, 167:112680, 2022.

[9] Shuting Zhao, Lifeng Wu, Youzhen Xiang, Jianhua Dong, Zhen Li, Xiaoqiang Liu, Zijun Tang, Han Wang, Xin Wang, Jiaqi An, Fucang Zhang, and Zhijun Li. Coupling meteorological stations data and satellite data for prediction of global solar radiation with machine learning models. Renewable Energy, 198:1049–1064, 2022.

[10] P. G. Kosmopoulos, S. Kazadzis, M. Taylor, Alkiviadis F. Bais, K. Lagouvardos, V. Kotroni, I. Keramitsoglou, and C. Kiranoudis. Estimation of the solar energy potential in greece using satellite and ground-based observations. In Theodore Karacostas, Alkiviadis Bais, and Panagiotis T. Nastos, editors, Perspectives on Atmospheric Sciences, pages 1149–1156, Cham, 2017. Springer International Publishing.

[11] Yi-Fan Feng. Mapping solar power potential using artificial intelligence with ground observation and reanalysis dataset: A case study of taiwan. Master's thesis, National Taiwan University, Taipei, Taiwan, 2023.

[12] Central Weather Administration. Central weather administration observation data inquire system. https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp, 2022. Accessed on May 6, 2023; The dataset is no longer available.

doi:10.6342/NTU202503010

[13] Central Weather Administration. Central weather administration agricultural meteorological observation network monitoring system. https://agr.cwb.gov.tw/NAGR/history/station_day, 2022. Accessed on May 6, 2023; The dataset is no longer available.

[14] J. Muñoz Sabater. Era5-land hourly data from 1950 to present. https://doi.org/10.24381/cds.e2161bac, 2019. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), Accessed on October 29, 2024.

[15] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 hourly data on single levels from 1940 to present. https://doi.org/10.24381/cds.adbb2d47, 2023. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), Accessed on October 29, 2024.

[16] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 hourly data on pressure levels from 1940 to present. https://doi.org/10.24381/cds.bd0915c6, 2023. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), Accessed on October 29, 2024.

[17] European Centre for Medium-Range Weather Forecasts. Era5-land: data documentation. https://confluence.ecmwf.int/display/CKB/ERA5-Land%3A+data+documentation#heading-Accumulations, 2024. ECMWF, Accessed on October 8, 2024.

[18] Taiwan Power Company. Daily power generation of wind and solar energy. https:

doi:10.6342/NTU202503010

//data.gov.tw/dataset/17140, 2024. Accessed on August 2024; The dataset no longer available.

[19] Taiwan Power Company. Daily solar power generation and average unit capacity statistics. https://data.gov.tw/dataset/29938, 2024. Accessed on July 2024.

[20] Taiwan Power Company. Renewable energy site information. https://data.gov.tw/dataset/17141, 2024. Accessed on August 2024.

[21] Google Maps. Google maps coordinate lookup. https://www.google.com/maps, 2024. Accessed on September 2024.

[22] Georges Matheron. Principles of geostatistics. Economic Geology, 58(8):1246–1266, 1963.

[23] Clayton V. Deutsch. Correcting for negative weights in ordinary kriging. Computers Geosciences, 22(7):765–773, 1996.

[24] Gordon Hudson and Hans Wackernagel. Mapping temperature using kriging with external drift: Theory and an example from scotland. International Journal of Climatology, 14(1):77–91, 1994.

[25] Scott Sinclair and Geoff Pegram. Combining radar and rain gauge rainfall estimates using conditional merging. Atmospheric Science Letters, 6(1):19–22, 2005.

[26] Michel Journée, Richard Müller, and Cédric Bertrand. Solar resource assessment in the benelux by merging meteosat-derived climate data and ground measurements. Solar Energy, 86(12):3561–3574, 2012. Solar Resources.

doi:10.6342/NTU202503010

Appendix A — Data Bias and Model Uncertainty

A.1 Overview

This study explores a simple approach to approximate the potential influence of data-related bias and model uncertainty. Specifically, the same input combinations are evaluated under different training set shuffle seeds—that is, different random shuffling orders used before training, which may lead to variations in the resulting data distributions. This setup allows readers to observe how changes in data exposure to the model may affect prediction outcomes.

The following sections present results grouped by input combinations: S1, S3, S4, and S5 first (see Figure A.1), followed by S1, S6, S7, and S8 (see Figure A.2). Here, S1 stands for ground-only input; S3, S4, and S5 denote dual-source strategies without additional ground variables; and S6, S7, and S8 represent the corresponding dual-source strategies including the three key ground-only variables. For a refresh on details of input design, please refer to Section 3.2.2.

doi:10.6342/NTU202503010

Note: The main results presented in this study are based on the first selected shuffle seed: 84408. Three additional shuffle seeds were arbitrarily selected and used in the figures shown in this appendix.

doi:10.6342/NTU202503010

A.2 Results from Other Shuffle Seeds for Inputs S1, S3, S4, and S5

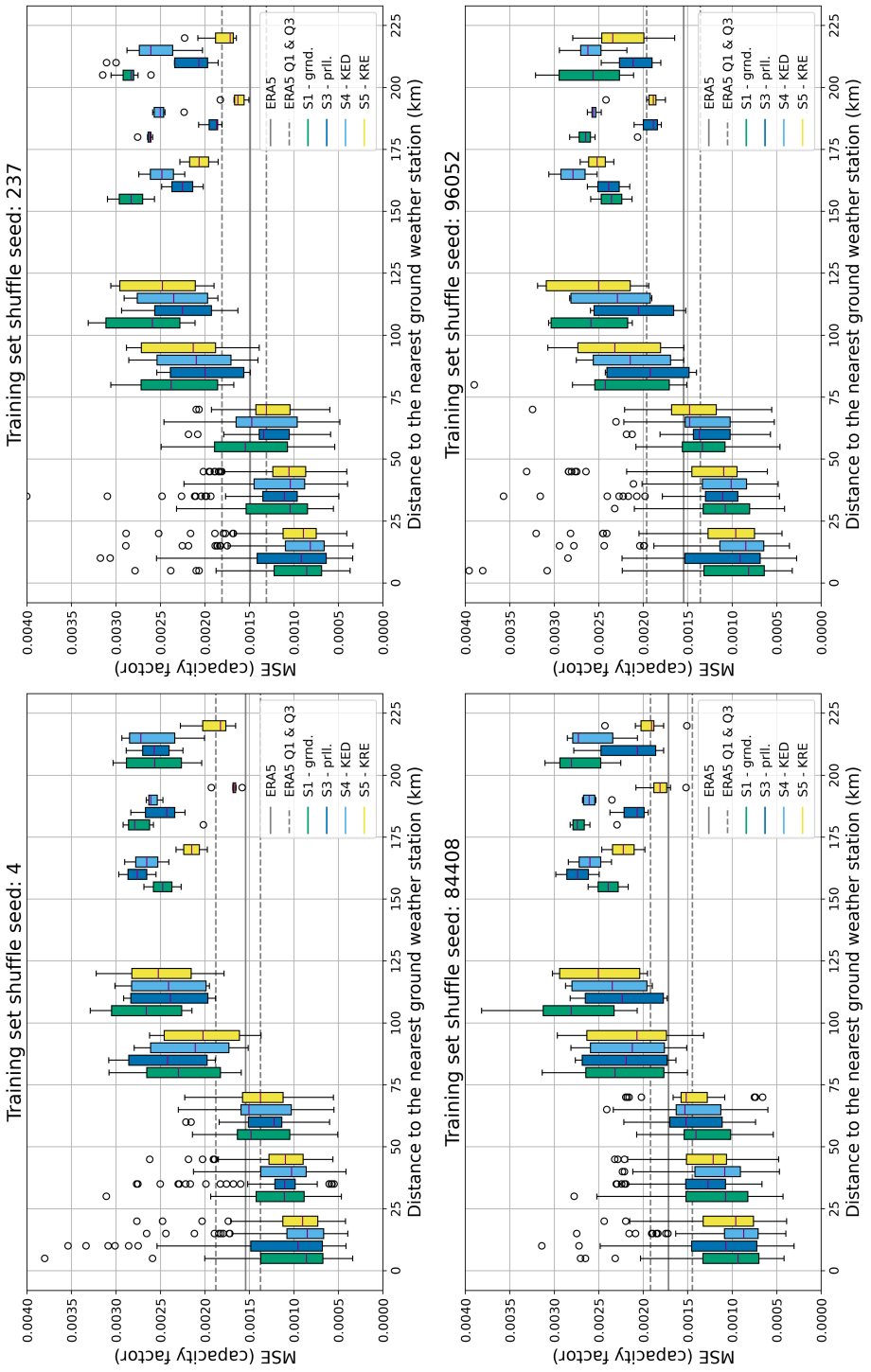


Figure A.1: Prediction results under four different training set shuffle seeds for input combinations S1, S3, S4, and S5. (The seed used in main study: 84408).

doi:10.6342/NTU202503010

A.3 Results from Other Shuffle Seeds for Inputs S1, S6, S7, and S8

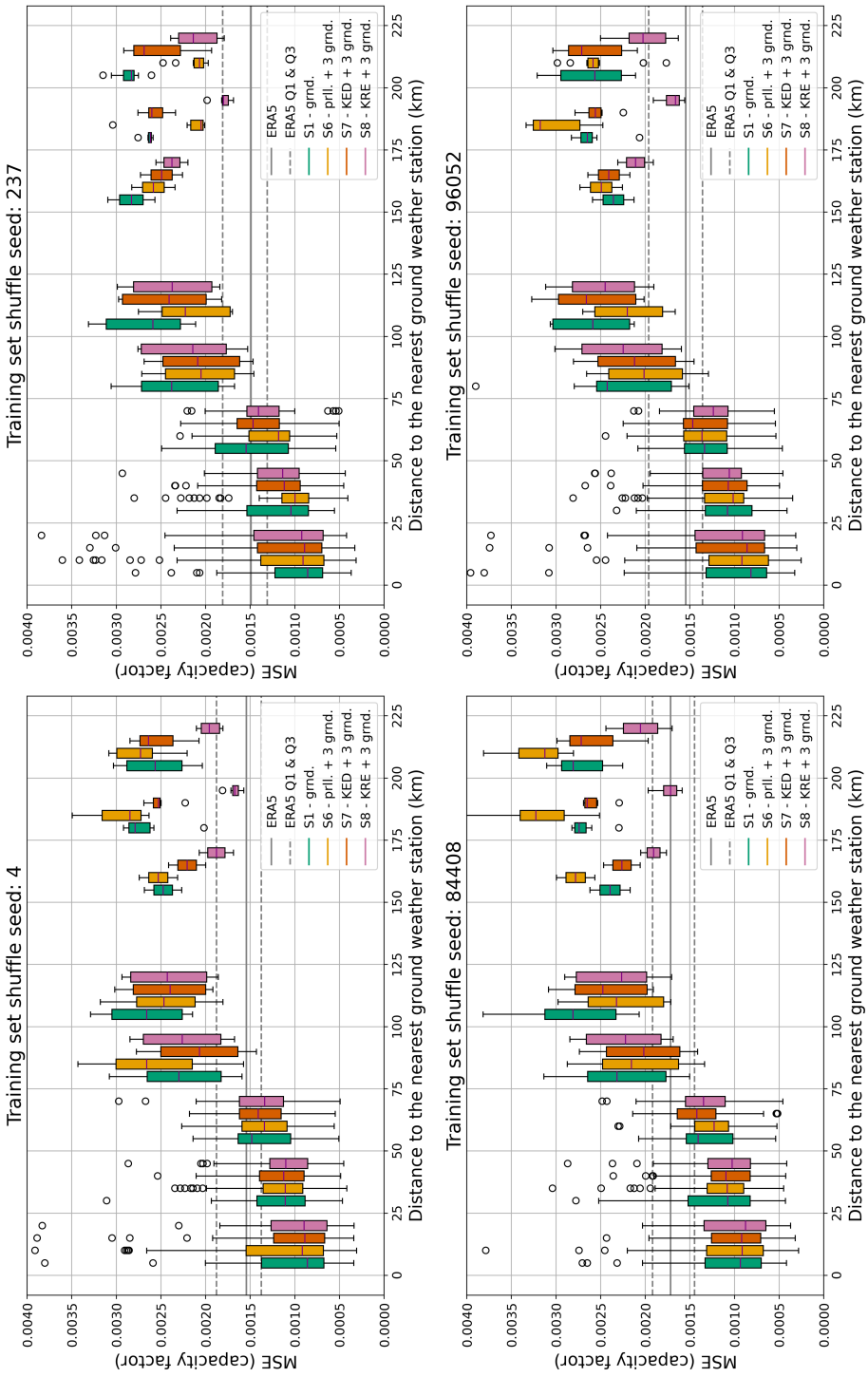


Figure A.2: Prediction results under four different training set shuffle seeds for input combinations S1, S6, S7, and S8. (The seed used in main study: 84408).

doi:10.6342/NTU202503010