機器學習在日射量預報之運用

趙俊傑[1]　張靖亞[1]　蔡政達[1]　張育承[2]
[1]資拓宏宇股份有限公司　[2]中央氣象局衛星中心

摘　要

政府的綠能政策規劃於2025年再生能源發電量將佔有全發電量中的百分之二十，其中太陽能光電裝置之容量目標達到200億瓦，將對供電系統之穩定及電力調度作業帶來相當大的挑戰。依據台電公司的統計資料顯示，太陽能光電系統目前最大發電量只達30億瓦。因此，如何有效掌握再生能源發電情況及維持系統穩定運轉，已為現階段應即早納入考量之課題。本文使用向日葵8號氣象衛星可見光雲圖計算不透明雲之反射率，帶入晴空日射量模式，計算即時之日射量，再透過雲導風推估不透明雲未來1-3小時的移動路徑，用以作日射量1至3小時之短時預測，用此方法雖可以估算日射量，但與日射量測站比較，仍不夠精準。本文希望透過機器學習的方式，先找出衛星資料估算數據與測站觀測日射量數據間的關係，再將衛星雲圖資料的預測結果應用在全台所有的地點，供太陽能發電之參考。

關鍵字：機器學習、日射量

一、 前言

日射量預報對於太陽能發電量之控制及農業栽培助益甚多，因此發展日射量預報之能力及準確性是有所需求的。目前應用衛星雲圖估算日射量的方法有許多人研究，例如胥立南(2015)及鄭光浩(2017)已發展同步衛星雲圖資料估算日射量，但甚少有用以預報日射量。本文應用向日葵 8 號同步衛星可見光雲圖資料，結合晴空日射量模式，發展日射量估計方法，再利用可見光雲動量之推估，以雲圖外延法做日射量短期預報，最後用機器學習方法與農業測站訓練的結果做日射量預報的修正，希望能作較為準確的日射量預報。

二、 資料及方法說明

本文使用的衛星資料為 Himawari-8 的 AHI 儀器所觀測的雲圖，AHI 有 16 個頻道，目前只使用第三頻道可見光頻道，其星下點解析度為 500 公尺，中心頻率是 0.64um。本文所參考使用的晴空日射量模擬程式為 REST2 套件(Gueymard, 2008)，REST2 為 REST2 為 Reference Evaluation of Solar Transmittance, 2 bands 的縮寫，是現今模擬軟體中可輸入參數較多、較為準確的一種(Sun 等 2019)，REST2 程式輸入之參數包含有雲圖背景值、氣膠光學厚度、二氧化氮、臭氧、水氣量、大氣壓力、太陽天頂角及 Angstrom exponent 等，這些參數包含空氣中的氣體份子(氣膠光學厚度、二氧化氮、臭氧。晴空日射量要能準確模擬，有賴於 REST2 程式輸入之參數能接近於觀測值，找尋目前可供下載的大氣資料庫，共有 NASA 的 MERRA-2 大氣資料庫及 ECMWF 的 CAMS 大氣資料庫符合資格。

NASA 的 MERRA-2 大氣資料庫提供的數據資料開始於 1980 年。由於同化系統的發展使得能夠加入現代高光譜輻射和微波觀測以及 GPS 掩星數據資料。它還使用了開始於 2004 年末的 NASA 臭氧剖面觀測結果。MERRA-2 包括 GEOS 模型和 GSI 同化系統的改進。空間解析度與 MERRA 大致相同（緯度方向約 50 km），隨著氣象同化的進步，MERRA-2 向 GMAO 的地球系統重分析目標邁出了重要的一步。MERRA-2 是第一個長期的全球重分析資料庫，它包含氣溶膠的衛星觀測結果，並顯示了它們與氣候系統中其他物理過程的相互作用。

ECMWF 之 CAMS 重分析資料是由哥白尼大氣監測服務系統（ Copernicus Atmosphere Monitoring

Service）產生的最新的全球大氣成分（AC）重分析數據資料庫，由時間一致的三維 AC 場組成，包括氣溶膠，化學物質和溫室氣體。在早期 MACC 重分析和 CAMS 臨時重分析的製作過程中獲得的經驗。數據資料目前涵蓋 2003-2019 年期間，CAMS 重分析數據的水平解析度大約為 80 km。

CAMS 重分析資料是使用 ECMWF 的綜合預報系統（IFS）的 CY42R1 中的 4DVar 數據資料同化而得的，垂直方向有 60 層混合 sigma / pressure（模型），最高層為 0.1 hPa。在這些層上都可獲得大氣數據，並且還將它們內插到 25 層等壓面，10 層等位溫層和 1 個等位渦層。還提供“表面或單層”數據資料。本研究最後選用 ECMWF 的 CAMS 大氣資料庫，主要考量為下載之方便性，

日射量模擬及日射量預報計算方法相似，其公式如下：

日射量模擬 = 晴空日射量*(1-不透明雲圖反照率)

日射量預報 = 預報晴空日射量*(1-預報不透明雲圖反照率)

其中不透明雲圖反照率為衛星雲圖可見光反照率減掉當時之背景值，預報不透明雲圖反照率為不透明雲圖外延法之反照率，晴空日射量及預報晴空日射量為晴空日射量模式輸出，惟模式輸入參數分別為初始場及預報場資料。

三、 日射量模擬

(一)製作不透明雲圖

統計向日葵八號衛星一個月同時間的可見光資料，找出最小值，當成此時間之背景值，運用向日葵八號衛星即時雲圖，減掉當時之背景值，計算而成不透明度雲圖，雲圖範例、背景圖範例及不透明度雲圖範例如下圖所示。

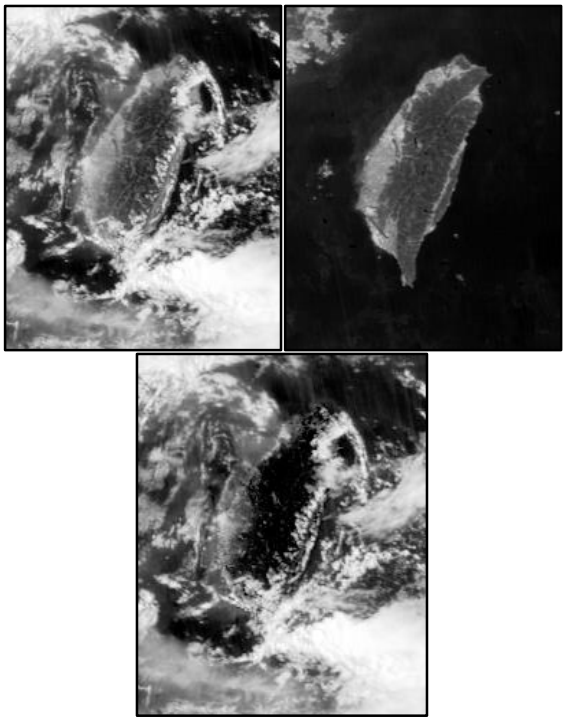


圖 1 圖左為雲圖範例，圖右為背景圖範例，圖下為不透明度雲圖範例，其關係為圖下=圖左-圖右

(二)日射量晴空模式

本文設定的範圍為東經 119 至 123 度，北緯 21 至 26 度，網格點數為 200x250，空間解析度約為 2 公里，將日本向日葵 8 號衛星可見光及 CAMS 模式資料皆內插至本文設定的範圍及解析度。另 CAMS 模式資料時間間隔為 3 小時一筆，日本向日葵 8 號衛星可見光資料時間間隔為 10 分鐘一筆，因此以衛星資料時間找最近 CAMS 模式資料時間。

晴空日射量計算需要氣熔膠光學厚度、二氧化氮、臭氧、水氣量、大氣壓力、太陽天頂角及 Angstrom 指數等資料輸入，另外亦需要背景值的輸入，背景值的計算採用一個月內同一時間的最低反照率，作業上則此採用前一年同月份的背景值，以 2018 年 07 月 14 日 01 時 UTC 為例，輸入參數如圖 2 所示，最接近的 CAMS 模式資料時間為 2018 年 07 月 14 日 00 時 UTC。

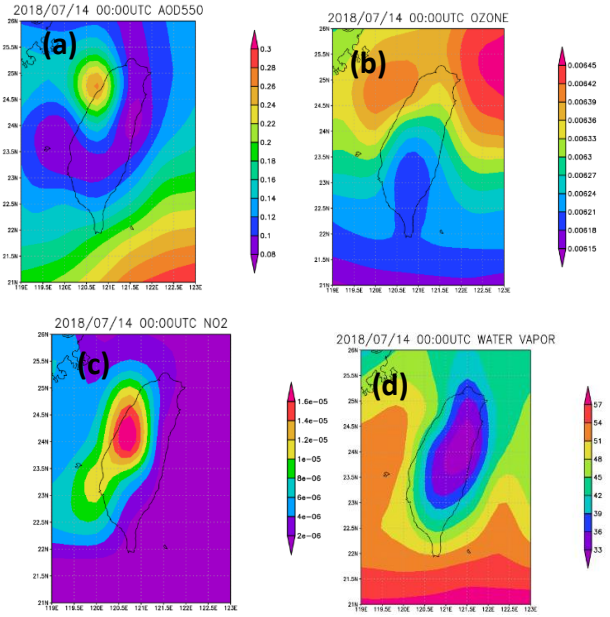


圖 2 ECMWF 的 CAMS 大氣資料庫 2018 年 7 月 14 日 0 時 UTC(a)氣熔膠光學厚度(b)臭氧(c)二氧化氮(d)水氣量。

2018 年 7 月 14 日 01 時 UTC 模擬的晴空日射量如圖 3(a)所示，圖中顯示平地日射量較大，海面及山區日射量較小，這是由於背景值在平地的反射率較大，如圖 3(b)所示，圖中色調越白反射率越大。

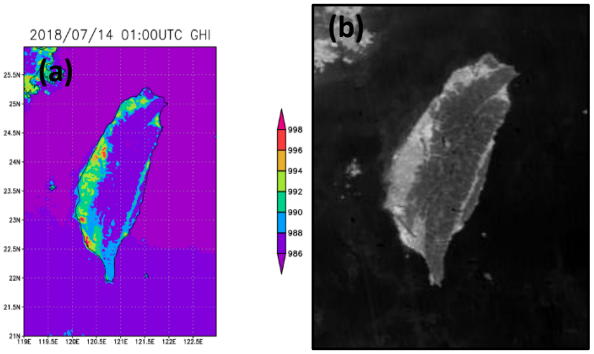


圖 3 2018 年 7 月 14 日 01 時 UTC(a) 晴空日射量模擬(b)7 月 01 時 UTC 背景值。

上一節已經詳述背景值的計算方法，亦介紹晴空日射量的模擬方法。2018 年 7 月 14 日 01 時 UTC 模擬的日射量如圖 4(a)所示，日射量在雲區較小，晴空較大，如圖 4(b)所示，圖中色調越白反射率越大。

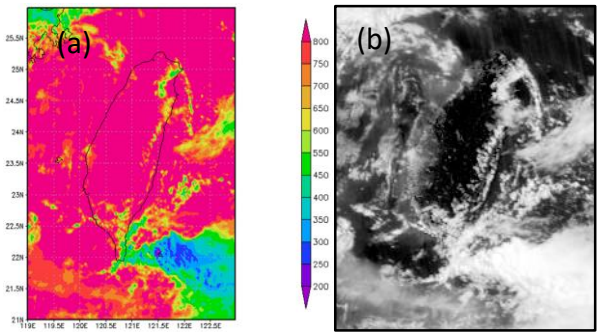


圖 4 2018 年 7 月 14 日 01 時 UTC(a)日射量模擬(b)不透明雲。

四、 機器學習方法導入

2016 年 5 月份日射量估算結果與茶業改良場觀測站比較之時序圖如圖所示，兩者還有許多不一致之處，包括低值高估及高值低估，估算量需要加以調整，方能接近觀測值，本文選用機器學習方法調校日射量估算值，希望能準確計算日射量。訓練的對象是農業站中的 11 個測站，測站名稱及位置如表 1 所示。

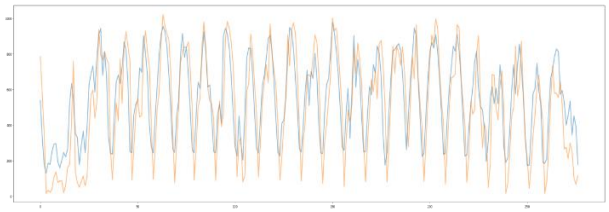


圖 5 茶業改良場 2016 年 5 月份觀測與估計值時序圖

表 1 機器學習的農業站

測站名稱	測站代號	測站高度	測站經度	測站緯度	備考
茶業改良場	82C16	195	121.2	24.9	
五峰工作站	72D08	1208	121.15	24.62	
農業試驗所	G2F82	85	122.067	24.03	
臺中農改場	72G60	19	120.51	24.02	
雲林分場	72K22	35	120.48	22.63	
義竹工	72M36	6	120.28	23.36	

作站					
高雄農改場	72Q01	24	120.53	22.71	
畜試恆春分所	B2Q81	20	120.78	21.97	
蘭陽分場	72U48	27	121.70	24.69	
花蓮農改場	72T25	36	121.56	23.97	
臺東斑鳩分場	72S20	240	121.07	22.03	

利用機器學習方法校驗預報日射量，及使用地面站觀測值與預測值做偏差訂正，須先將資料依序進行前處理、特徵抽取、建立模型、訓練、測試、驗證結果、模型選擇等程序。

資料前處理將各點位依照時間進行排序與檢視，觀察數據情況，進行特徵工程，調整之預測值中的資料項目(X)為：日射量(Irradiance)、反照度(albedo)、高度(height)、雲覆蓋狀況、年、月、日、小時、天頂角(zendth_angle)。

欲使用統計及資料轉換方法以增加可使用的特徵數量，來增加模型訓練的效率。使用的方法如下：

獨熱編碼(one-hot-encode)又稱為一位有效編碼，是將類別項目作為新的欄位並以數字 1 或 0 來表示類別狀態，可有效使用非數字類的數據來做為特徵使用，此處將資料的小時數以此方法表示，轉換範例如下表中左表至右表。。

表 2 、獨熱編碼示意表

序號	數據範圍		序號	數據範圍 -0-10	數據範圍 -10-15	數據範圍 -15-20
1	0-10	→	1	1	0	0
2	10-15		2	0	1	0
3	15-20		3	0	0	1

1.特徵四則運算

特徵間的加減乘除等四則運算對於特徵的萃取有部分的效果，這邊使用開源套件 feature tool 來進行特徵生產與篩選。

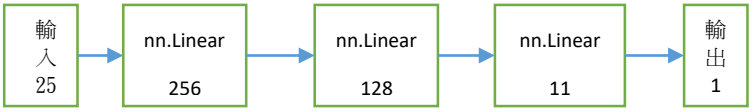
2.偏度(skewness)矯正測試

數據的份部並不會每次都趨近常態分佈，所以可先經由偏度的調整來使數據的分布更加平衡，這邊使用 Box-Cox 的變換法做調整並測試。

3.對數變換(log transformation)

將方差穩定，使資料平衡，判斷資料的分布是否偏向線性分布，來做為模型的挑選依據。

4. 資料區間化，給予資料以固定比例的間隔給予分類指標，如：日照度介於 0 至 10 之間標示為 1，介於 10



至 20 標示為 2，以此類推。

5. 統計數據增量。

6. 標準化(Preprocessing)將日射量、反照度、高度、天頂角分別進行以下二種標準化的方法。

(1) 最大最小值標準化(MinMaxScaler)將屬性縮放到一個指定的最大和最小值(0-1 之間) ，公式如下：

$$x_* = \frac{(x - x_{min})}{x_{max} - x_{min}} \times (max - min) + min$$

(2) 標準差標準化(StandardScaler)，公式如下，其中 μ 為所有樣本數據的均值，σ 為所有樣本數據的標準差：

$$x^* = \frac{x - \mu}{\sigma}$$

7. 前後值資料特徵

由於在後續進行測試時，使用最近鄰居法(KNN, k-nearest neighbors)時訓練狀況較佳，測試時表現不佳，出現過度擬合(overfitting)的狀況，此模型原理為鄰近資料的投票機制，意指目標數值的前後數值將影響該數值的生成，故推測此模型的原理可能對於這次的時序資料會有較好的適應性，所以新增了前後值為特徵。數值生成邏輯如下：

表 3 、前後值資料示意表

Data	前 1 筆	前 2 筆	後 1 筆
100	100	100	200
200	100	200	300
300	200	100	400
400	300	200	400

(一) 模型測試

使用機器學習與深度學習建立模型，評估訓練狀況，模型如下：

1. 深度學習(Deep learning)

使用 pytorch 進行神經網路的建構，圖中數字代表數據的維度，會將輸入的特徵維度提升到一個自設的最大值，然後再依序降低，最後降成輸出所需的 1 維資料，模型中的維度設定使用自動化機器學習(AutoML)進行調整。架構如下：

圖 6、神經網路結構圖

2. 機器學習(Machine learning)

將使用下列模型進行測試，並搭配 TPOT 進行模型組合的配對。

(1) 支援向量迴歸(SVR, Support Vector Machine Regression)

(2) 線性迴歸(Linear Regression)

(3) 隨機森林(RandomForest)

(4) 極限梯度提升(XGBoost, eXtreme Gradient Boosting)

(5) 最近鄰居法(KNN, k-nearest neighbors)

(6) LightGBM(LGBM,Light Gradient Boosting Machine)

(二) 模型訓練特徵測試

將特徵工程取得之資料匯入模型中，進行重要性評估，於每次訓練後從模型內部的參數調出各項特徵的重要性分數，來評估需保留及剔除的特徵，增進模型學習的效率。下列為特徵測試中各階段測試具指標性的結果。

1. 使用特徵：對數變換、標準化、日射量、天頂角資料。

相關性分數(correlation score)：0.77，可以發現使用最大最小值標準化的數據以及日射量預測值的重要性較高，對數轉換的項目分數較低，由此推論可先不考慮此特徵。

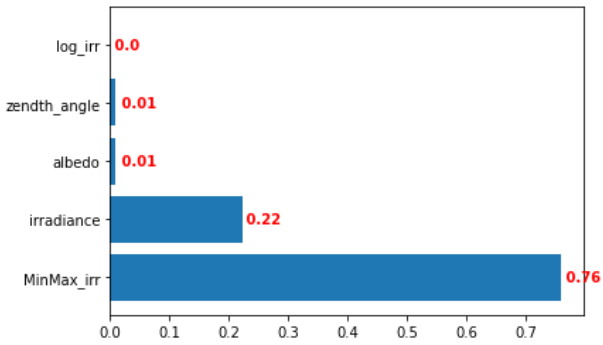


圖 7、特徵重要性排名

2. 使用特徵：獨熱編碼、區間化、標準化、日射量。

相關性分數(correlation score)：0.874，可以發現使用前述三種標準化的數據、日射量預測值、小時類別重要性高。

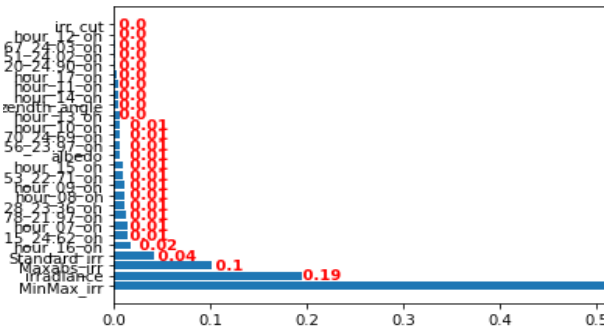


圖 8、特徵重要性

於此階段數據校正狀況如下(藍線為預測值，橘線為觀測值)，可以發現圖中低值在修正後都有明顯調整，但紅圈處仍未貼近完全：

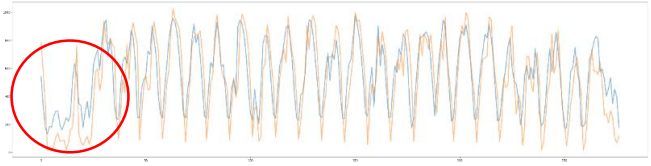


圖 9、修正前預測值與觀測值比較

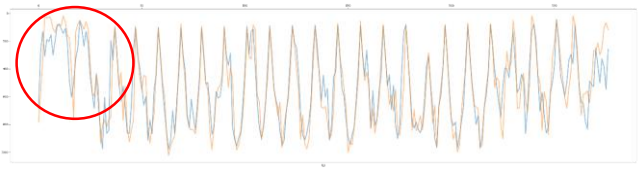


圖 10、修正後預測值與觀測值比較

3. 使用特徵：前後數值、獨熱編碼、區間化、標準化、日射量。

相關性分數(correlation score)：0.935，可以發現藉由前後值資料的特徵加入後可增加此時間序列模型的校正表現。

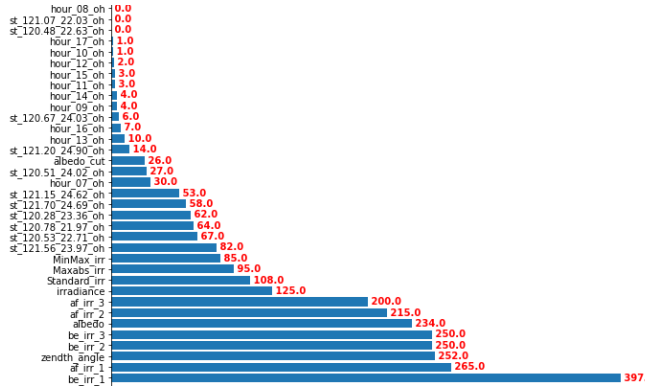


圖 11、特徵重要性圖-3

於此階段數據校正前後比較狀況如下(藍線為預測值，橘線為觀測值)：

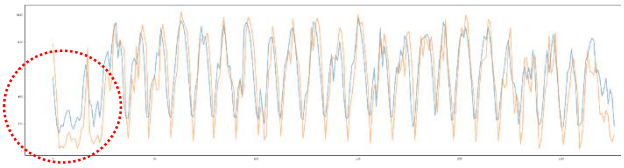


圖 12、修正前預測值與觀測值比較圖-2

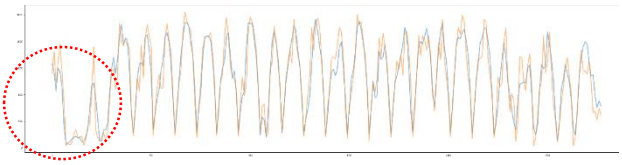


圖 13、修正後預測值與觀測值比較圖-2

(三) 模型超參數調整(Hyperparameter tuning)

使用 optuna、GridSearchCV、RandomizedSearchCV 等開源套件執行。下列圖為 optuna 於 LGBMR 的參數調整結果，參數調整過程依照下列資訊進行微調與設定：

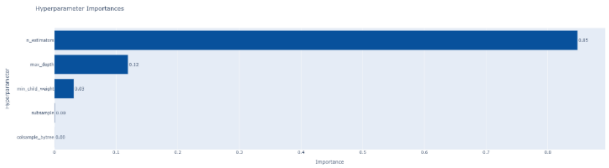


圖 14、超參數重要性比較

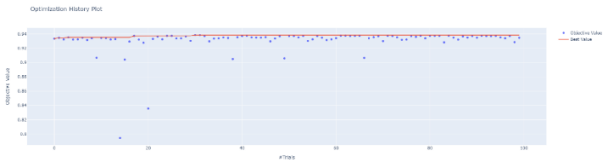


圖 15、相關性分數分布圖

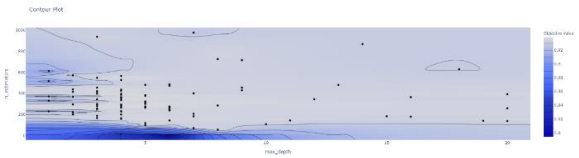


圖 16、超參數相關性分數對應圖

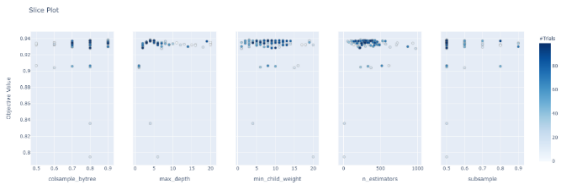


圖 17、超參數交互對照選擇優先狀況圖

(四) 初步結果驗證

經過特徵選定、超參數調節後，最終進行結果驗證，使用 k 折交叉驗證 (k-fold cross-validation)，此方法將資料照比例切割，並於每次訓練時使用不同的訓練資料、測試資料，循環檢驗模型在各種資料的組合下可以達成的準確度，避免單一資料檢驗的不確定性。表現最佳的模型驗證結果如下表，依照相關性指標平均值的最高者作為最終使用的模型：

表 4、交叉驗證結果表

模型名稱	R^2_平均值	R^2_標準差	相關性_平均值	相關性_標準差
線性迴歸(Linear Regression)	0.80605	0.040504	0.897515	0.022746
極限梯度提升(XGBoost	0.827553	0.047569	0.909308	0.026694
隨機森林(RandomForest)	0.828381	0.049155	0.909736	0.027594
最近鄰居法(KNN)	0.782458	0.052372	0.884055	0.030058
LightGBM(LGBM)	0.838393	0.051174	0.915194	0.028508

五、日射量預報

(一)日射量預報方法

本節則介紹日射量預報計算方法，日射量預報為日射量三小時短時預報，並假設雲在 3 小時內移動不會增強及減弱，其公式如下：

日射量預報 = 預報晴空日射量*(1-不透明雲預報反射率)*機器學習修正量

其中不透明雲預報反射率為不透明雲依照雲導風外推移動後之反射率，第 1 至 3 小時雲的移動方向及速度為當時雲導風的方向及速度，雲導風的計算乃使用 Himawari-8 衛星第 3 頻道 500 公尺解析度、每 10 分鐘雲圖計算雲動量，一小時的雲動量為一小時內每十分鐘所計算的雲動量的平均。

雲動量計算方式為以圖中心 24x24 像素當成選取目標區，在下一時間雲圖以各方向移動比對，區域內搜索比對範圍為 80x80 像素，在比對目標區時計算交叉相關係數(cross-correlation coefficient)，其中最大值代表實際移動區域及此區域之雲動量，如圖 18 所示。由於目前 Himawari-8 衛星 10 分鐘觀測全球乙次，所以計算之風場為 10 分鐘間隔之移速，單位需轉換至每小時速度。

2019 年 08 月 04 日 01 時 UTC 之雲動量計算結果如圖 19 所示，用此風場預報第 0 及 1 小時不透明雲如圖 20 所示，其分別代表 2019 年 08 月 04 日 01 時 UTC 及 2019 年 08 月 04 日 02 時 UTC 之不透明雲預報反射率。預報第 2 及 3 小時不透明雲如圖 21 所示，其分別代表 2019 年 08 月 04 日 03 時 UTC 及 2019 年 08 月 04 日 04 時 UTC 之不透明雲預報反射率。因此便可以用來預報日射量。

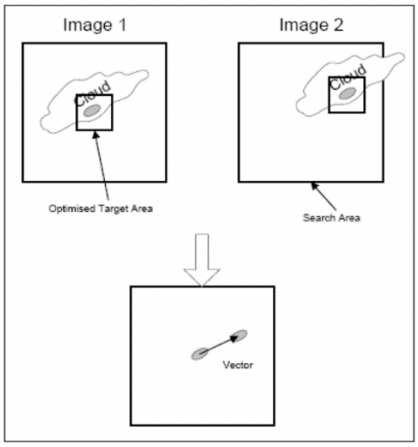


圖 18 雲動量計算示意圖。

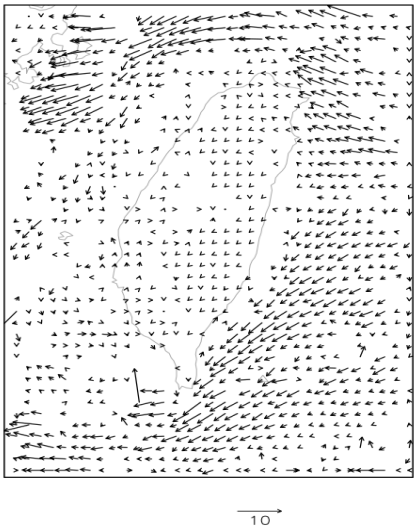


圖 19 2019 年 08 月 04 日 01 時 UTC 雲導風。

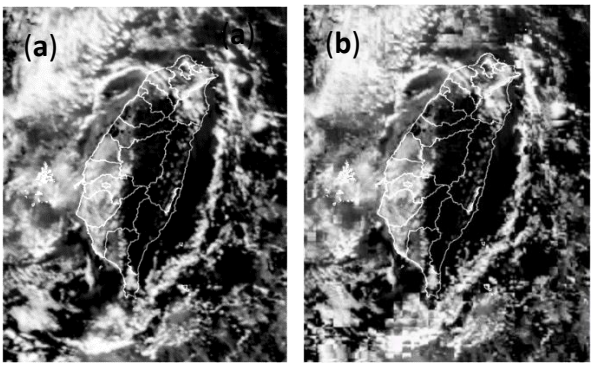


圖 20 2019 年 08 月 04 日 01 時 UTC 不透明雲第(a)0 及(b)1 小時預報圖

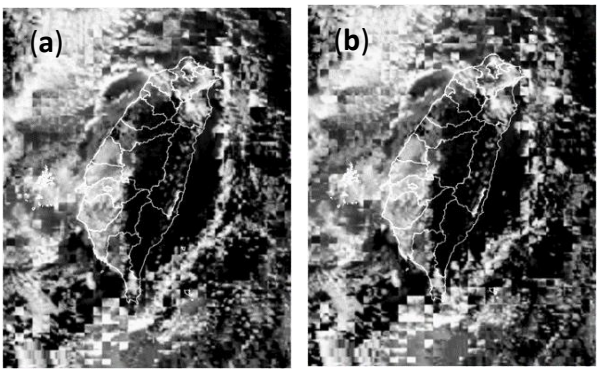


圖 21 2019 年 08 月 04 日 01 時 UTC 不透明雲第(a)2 及(b)3 小時預報圖

預報晴空日射量為 REST2 晴空日射模擬程式輸入預報時間的背景值及預報時間的 CAMS 模式大氣氣體參數值，ECMWF 之 CAMS 模式除了有重分析資料外，亦有 Near-real-time 預報資料可供下載(網址為 https://apps.ecmwf.int/datasets/data/cams-nrealtime/)。

日射量預報所需要的資料(包含不透明雲及晴空日射量)皆計算出來後，即可計算日射量預報及最後一步驟的 AI 修正，2019 年 08 月 04 日 01 時 UTC 第 0 至 3 小時預報日射量未經 AI 修正及已經 AI 修正結果如圖 22 及圖 23 所示，從兩圖中比較可以發現台灣山區經 AI 修正後，日射量值有降低的現象，推測是因為訓練時山區雲量較多，日射量較小，因此在預測日射量時就反映出此資料特性，但高山上因為沒有低層大氣的散射及吸收，實際上日射量較大，下一節將驗證其優缺點。

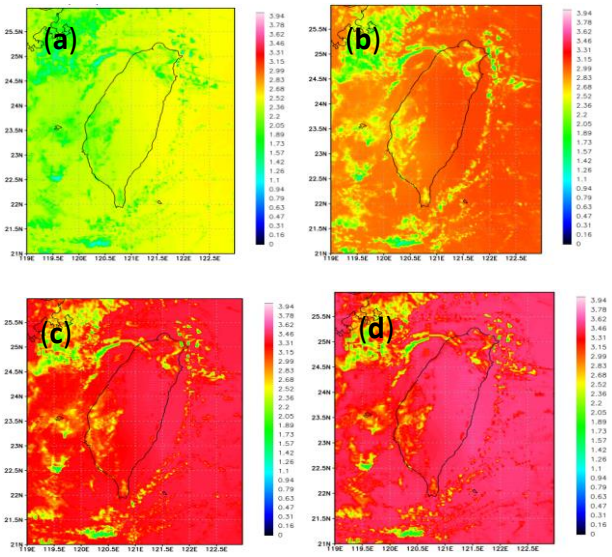


圖 22 2019 年 08 月 04 日 01 時 UTC 未經 AI 修正之預報日射量第(a)0、(b)1、(c)2 及(d)3 小時預報圖

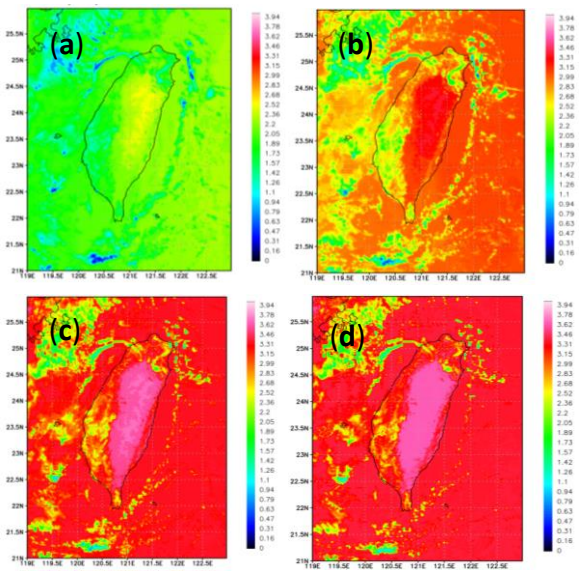


圖 23 2019 年 08 月 04 日 01 時 UTC 預報日射量第(a)0、(b)1、(c)2 及(d)3 小時預報圖

(二) 日射量預報驗證

本節主要比對日射量預報值與測站觀測值之差異，測站觀測值是農業站及局屬站地面觀測儀器所量測的數值，因為有定期的調校，其值亦相當的準確。這一小節驗證用"點"的方式檢驗，即找與測站同一時間圖中最近點當成比較對象，比較點的位置如圖 24 所示，圖 24(a)及(b)分別為農業站及局屬站。2019 年 08 月 04 日 0 至 7 時 UTC 第 0 小時預報日射量結果分別與當時所有農業站及局屬站日射量觀測值之比較如圖 25(a)及(b)所示，其與農業站相關係數 R2 值可達 0.61，但與局屬站相關係數 R2 值只有 0.35，這是因為當初機器學習校驗時用農業站來訓練之故，且只選擇農業站中的 11 個站當作訓練的目標，未來可以拿所有測站資料進入機器學習訓練，應該可以改善日射量估計值與觀測值相關性。

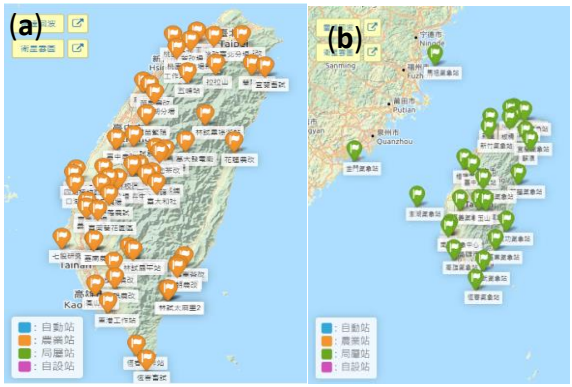


圖 24 (a)農業站位置示意圖，(b)局屬站位置示意圖。

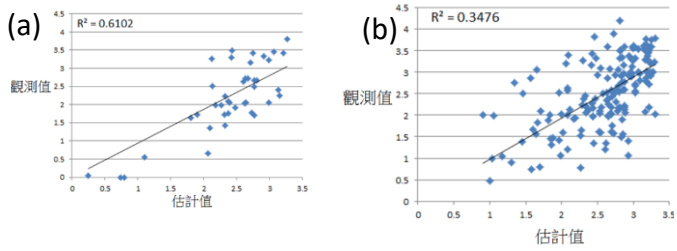


圖 25 2019 年 8 月 4 日預測 0 小時(a)農業站及(b)局屬站的日射量觀測值與估計值之散布圖。

2019 年 08 月 04 日 0 至 7 時 UTC 第 1 小時預報日射量結果分別與當時所有農業站及局屬站日射量觀測值之比較如圖 26(a)及(b)所示，其與農業站相關係數 R^2 值為 0.52，但與局屬站相關係數 R^2 值只有 0.23，由於山區雲量於 10 點後開始逐漸增加，預報與觀測的雲量有差異，故相關係數比 0 小時預報稍降低。

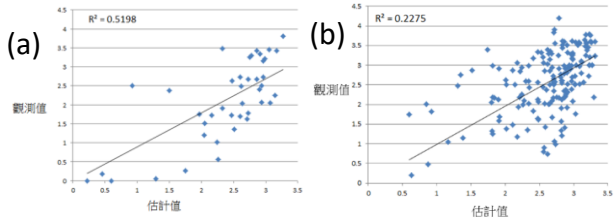


圖 26 2019 年 8 月 4 日預測 1 小時(a)農業站及(b)局屬站的日射量觀測值與估計值之散布圖。

2019 年 08 月 04 日 0 至 7 時 UTC 第 2 及 3 小時預報日射量結果與當時所有農業站日射量觀測值之比較如圖 27(a)及(b)所示，其相關係數 R^2 值分別為 0.43 及 0.48，由於山區雲量於 10 點後開始增加，第 2 及 3 小時預報與觀測的雲量有差異，故相關係數比 1 小時預報更降低，惟第 2 及 3 小時預報相關係數差異不大。

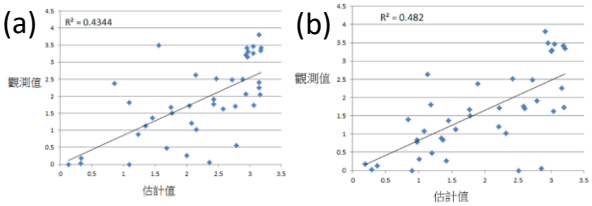


圖 27 2019 年 8 月 4 日預測(a)2 及(b)3 小時農業站的日射量觀測值與估計值之散布圖。

六、結論

經過前面描述日射量預報之驗證，下面列出幾點結論：

1.日射量短時預報由雲動量加上 REST2 模式及 AI 校驗初步完成，預測結果 3 小時還算可用，若 2~3 小時雲變化大，則預測結果較不準，可以利用 AI 進行山區雲成長之學習訓練，改善日射量短時預報。

2. 機器學習校驗可分為有雲及無雲，有雲時不取高度值(因為不相關)，無雲才取，並擴大校驗測站，包含所有農業站及局屬站，可以增加預報準確率。

3. 雲動量可以嘗試使用雲塊追蹤法計算，改善雲動量之準確率。

4. 此方法也可以應用於日射量中長期預報，即利用氣象預測模式之預報場帶入輻射傳遞方程式，模擬可見光反照率。

七、參考文獻

胥立南，2015：“應用 MTSAT2 衛星資料估算臺灣地表日射量”，104 年天氣分析與預報研討會論文全文彙編，中央氣象局，臺北市。

鄭光浩等人：“應用 Himawari-8 估計臺灣地表日射量之校驗及探討”， 2017 年天氣分析與預報研討會，中央氣象局，臺北市。

C. A. Gueymard, “REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation – Validation with a benchmark dataset”, Solar Energy, vol. 82, no. 3, pp. 272 - 285, 2008.

X. Sun, J. M. Bright, C. A. Gueymard, B. Acord, P. Wang, N. A. Engerer, Worldwide performance assessment of 75global clear-sky irradiance models using principal component analysis, Renewable Sustainable Energy Rev. 111 (2019)550 – 570