The Journal of Engineering Geology
pISSN : 1226-5268 eISSN : 2287-7169

Vol. 33, No. 3, September, 2023, pp. 403-411
https://doi.org/10.9720/kseg.2023.3.403

RESEARCH ARTICLE

XGBoost를 이용한 타지키스탄 일사량 예측 모델

노정두[1] · 나태유[2] · 강성승[3*]

[1]전남연구원 전라남도탄소중립지원센터 전문연구위원, [2]한국지질자원연구원 심층처분환경연구센터 연구원,
[3]조선대학교 에너지자원공학과 교수

Modeling Solar Irradiance in Tajikistan with XGBoost Algorithm

Jeongdu Noh[1] · Taeyoo Na[2] · Seong-Seung Kang[3*]

[1]Researcher, Jeollanamdo Carbon Neutral Center, Jeonnam Research Institute
[2]Researcher, Geology and Space Division, Korea Institute of Geoscience and Mineral Resources
[3]Professor, Department of Energy and Resources Engineering, Chosun University

Abstract

The possibility of utilizing radiant solar energy as a renewable energy resource in Tajikistan was investigated by assessing solar irradiance using XGBoost algorithm. Through training, validation, and testing, the seasonality of solar irradiance was clear in both actual and predicted values. Calculation of hourly values of solar irradiance on 1 July 2016, 2017, 2018, and 2019 indicated maximum actual and predicted values of 1,005 and 1,009 W/m^2, 939 and 997 W/m^2, 1,022 and 1,012 W/m^2, 1,055 and 1,019 W/m^2, respectively, with actual and predicted values being within 0.4~5.8%. XGBoost is thus a useful tool in predicting solar irradiance in Tajikistan and evaluating the possibility of utilizing radiant solar energy.

Keywords: Tajikistan, XGBoost, solar energy, solar irradiance, renewable energy

초 록

본 연구는 XGBoost를 이용하여 타지키스탄의 일사량을 예측하여 타지키스탄의 재생에너지 자원으로서 복사 태양에너지의 활용 가능성을 평가하기 위함이다. 첫째, 타지키스탄의 일사량을 훈련모델, 검증모델, 시험모델을 통해 예측한 결과, 시간과 계절에 따른 일사량의 계절성이 실제값과 예측값 모두에서 뚜렷하게 구분되는 것을 확인하였다. 둘째, 타지키스탄의 2016, 2017, 2018, 2019년 등 각 연도의 7월 1일 시간당 일사량의 실제값과 예측값을 계산한 결과, 2016년 일사량의 최대 실제값과 예측값은 약 1,005 W/m^{2}과 1,009 W/m^2, 2017년에는 939 W/m^{2}과 997 W/m^2, 2018년에는 1,022 W/m^{2}과 1,012 W/m^2, 2019년에는 1,055 W/m^{2}과 1,019 W/m^{2}으로 나타났으며, 실제값과 예측값의 오차가 약 0.4~5.8%로 매우 비슷한 결과를 보였다. 결과적으로 타지키스탄의 일사량을 예측하여 복사 태양에너지의 활용 가능성을 평가하는 데 있어 XGBoost가 매우 유용한 도구로 활용될 수 있을 것으로 판단된다.

주요어: 타지키스탄, XGBoost, 태양에너지, 일사량, 재생에너지

***Corresponding author:** Seong-Seung Kang
E-mail: kangss@chosun.ac.kr

Received: 17 July, 2023
Revised: 20 August, 2023
Accepted: 31 August, 2023

서 론

타지키스탄의 2030 국가발전전략에서 에너지 개발수요에 의하면, 국가의 에너지 수급 문제 해결과 국내 생산 전력의 효율적인 관리를 주요 목표로 제시하고 있다(Joint Ministries, 2022). 이것은 겨울철 전력 부족으로 인한 전기 공급 제한과 연간 약 2천억 달러 규모의 에너지 손실 등의 문제를 해결하기 위하여 신재생에너지 중심의 전력생산 방식의 다변화와 함께 에너지 자원의 효율적 활용 등 에너지 협력을 추구하고 있다. 그것의 일환으로 2018~2021년까지 KOICA(Korea International Cooperation Agency) 타지키스탄 전력소외지역 전력망 구축 사업, 2021년 카티키스탄, 수드그 주, 고르노바타흐샨 자치주 태양광 발전 및 ESS 구축 사업 등 한국-타지키스탄 간에 협력 사업이 수행되었다. 타지키스탄(Tajikistan)의 에너지 및 수자원부(Ministry of Energy and Water Resources)의 자료에 따르면 연간 에너지 생산량은 2015년 17.3백만 kWh에서 2020년에는 19.6백만 kWh로 약 13% 증가, 에너지 소비량은 2015년 15.6백만 kWh에서 2020년에는 17.9백만 kWh로 약 15% 증가한 것으로 나타났다(Kudusov et al., 2021, Table 1). 현재 타지키스탄에서 에너지 생산의 대부분은 수력 발전(hydro-electric power plants, HPPs)과 열병합 발전(combined heat and power plants, CHPs)이 차지하고 있으며, 수력 발전이 91.6%, 열병합 발전이 8.4% 비율을 차지하는 것으로 보고되었다(IEA, 2022). 하지만 타지키스탄의 수력 발전에 의한 전력 생산은 매년 건조한 날씨의 기후 조건과 계절적 특성에 따른 강의 흐름, 제한된 수력 발전 능력 한계 등의 원인으로 타지키스탄 전체 전력 수요를 충족시키기 어려운 상황이다. 따라서 현재의 이러한 전력 수급 문제를 해결하기 위하여 타지키스탄에서는 전력 생산의 새로운 대안으로 재생에너지 자원을 적극 고려하고 있다. 재생에너지 자원 중 타지키스탄에 가장 적합한 것으로는 연간 일광 시간이 2,100~3,635시간에 이르는 복사 태양에너지로서 타지키스탄의 지형 조건과도 잘 일치한다(Kudusov et al., 2021). 본 연구의 목적은 머신러닝 기법을 이용하여 타지키스탄에서 재생에너지 자원으로써의 복사 태양에너지의 활용 가능성을 평가하기 위함이다. 이를 위하여 i) 입력 자료로는 타지키스탄 위성 관측 자료와 Renewables.ninja 사이트에 공개된 자료를 사용하였으며, ii) 일사량 예측을 위해서는 머신러닝(machine learning)의 앙상블 모델 중 하나인 XGBoot를 이용하였다.

Table 1. Electricity generation and consumption in Tajikistan during 2014~2020 (Kudusov et al., 2021)

Year	2014	2015	2016	2017	2018	2019	2020
Generation (kWh)	16.4	17.1	17.3	18.1	19.7	20.7	19.6
Consumption (kWh)	15.0	15.6	15.7	16.6	17.1	17.6	17.9

연구 지역

타지키스탄의 행정구역은 수도 및 공화국 직할구, 2개의 주(수그드 주, 하틀론 주), 그리고 1개 자치주(고르노바다흐샨 자치주)로 나뉜다(Fig. 1). 수도는 두샨베(Dushanbe), 수그드(Sughd) 주의 주도는 후잔드(Khujand), 하틀론(Khatlon) 주 주도는 보흐타르(Bokhtar), 고르노바다흐샨(Gorno-Badakhshan) 자치주의 주도는 호로그(Khorog)이다. 수도 및 공화국 직할구, 수그드 주, 하틀론 주를 포함한 타지키스탄 서부의 주요 도시로는 Karakum, Kanibadam, Leninabad, Khujand, Isfara, Ura-Tyube, Penjakent, Jirgatol, Garm, Dushanbe, Tursunzade, Dzhauz, Pakhtaboy, Yovon, Nurek, Dangara, Kulob, Bokhtar, Kamsomolobod, Shartuz, 동부의 고르노바다흐샨 주 주요 도시로는 Fedchenko Glacier, Karakul, Khaburabad, Nau, Murgab, Khorog 등이다(Na et al., 2023). 이들 지역의 일사량에 대한 위성 관측 자료와 Renewables.ninja 사이트의 공개 자료를 사용하여 타지키스탄 일사량 실측값과 예측값을 분석하였다.

Fig. 1. Map showing administrative divisions of Tajikistan, including the Dushanbe: districts of republican suboidination, Sugd and khatlon provinces, and Gorno-Badakhshan autonomous region.

연구방법

입력 데이터

타지키스탄 일사량 예측을 위한 모델 입력 자료는 Pfenninger and Staffell(2016)이 태양광 및 풍력 발전을 시뮬레이션하기 위하여 사용된 위성 관측자료와 국제재분석모델(global reanalysis models)로 분석하여 Renewables.ninja 사이트에 공개한 자료를 이용하였다(Renewables.ninja, 2023). 이 사이트는 1980년 1월 1일부터 2019년 12월 31일까지의 기온, 강수량, 강설량 등 8가지 변수들에 대한 시간당 분석 자료를 무료로 제공하고 있다(Fig. 2). 예측 모델을 위한 각 변수에 대한 상관관계 분석 결과는 Fig. 3과 같다. 분석 결과에서 알 수 있듯이 상관계수는 대기 밀도와 기온, 지면 일사량과 대기 일사량이 0.97로 가장 높게 나타냈다. 타지키스탄의 일사량 예측은 최근 20년(2000년 1월~2019년 12월) 동안의 기상자료를 사용하였으며, 예측 모델은 XGBoost를 이용하여 수행하였다. 이들 자료 중 2000~2010년 자료는 훈련데이터, 2011~2015년 자료는 검증데이터, 2016~2019년 자료는 시험데이터로 사용되었다. 8개 변수 중 지표면 일사량은 출력 변수, 나머지 7개 변수인 기온, 강수량, 대기 일사량 등은 입력 변수로 사용하였다.

XGBoost

타지키스탄 일사량 예측 모델은 회귀와 분류 예측에 유용한 XGBoost(eXtreme Gradient Boosting)를 활용하여 실시하였다. XGBoost는 기존의 GBM(Gradient Boost Machine)을 개선한 방식으로 병렬처리가 가능하며 잔차를 이용해 모델을 학습하고 과적합을 방지하기 위하여 몇 개의 인자들이 추가된 알고리즘이다. 다른 트리(tree) 기반의 학습 방식과는 달리 CART(Classification and regression tree) 모델을 기반으로 학습하며 식 (1)을 사용한다. 여기서, 데이터가 입력변수 x와 출력변수 y 구성되었을 때, $\hat{y}$는 데이터 x의 예측값, K는 사용된 CART의 개수, f는 CART의 모델을 각각 나타낸다(Chen and Guestrin, 2016; Yoon, 2020; An, 2021).

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

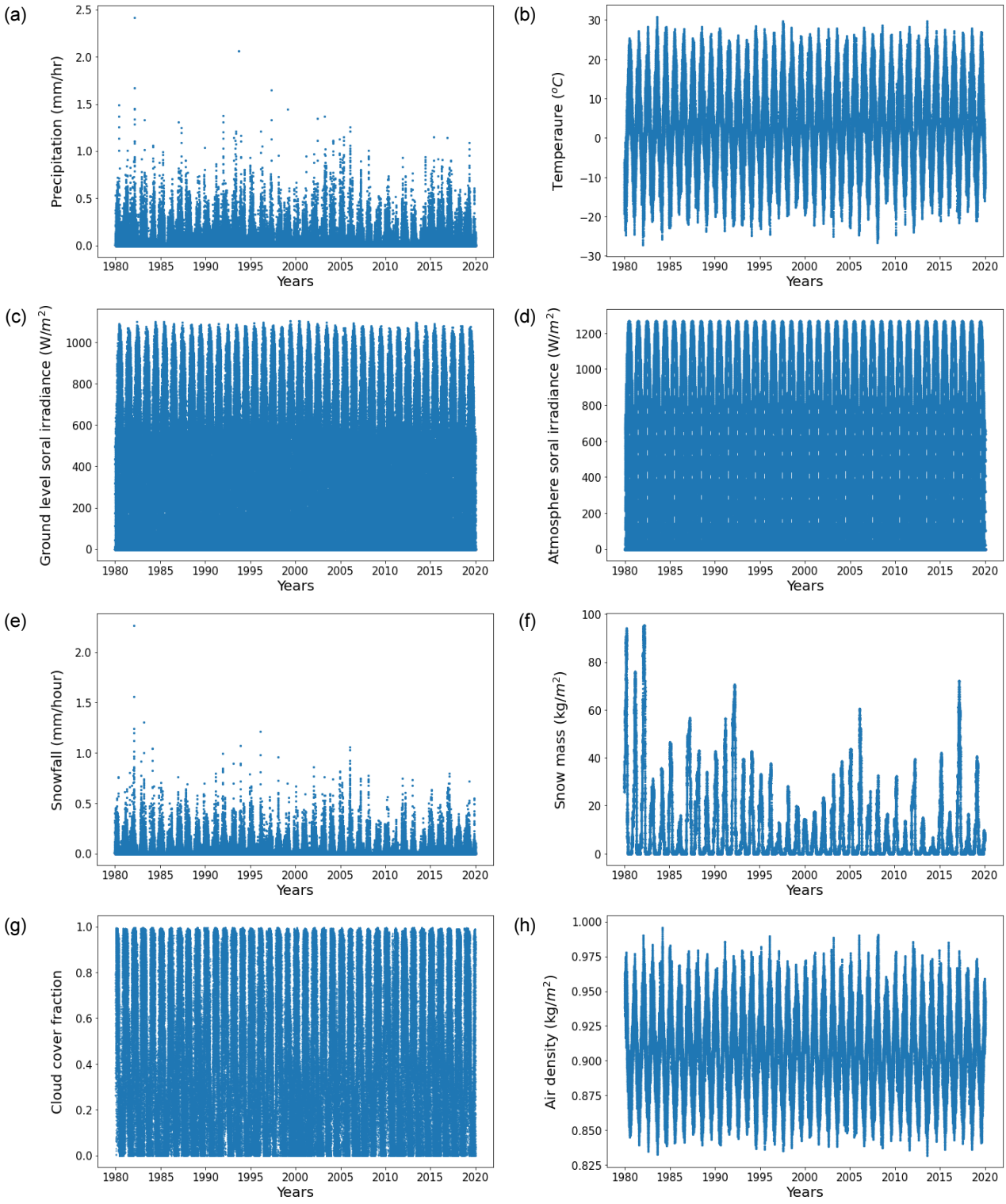

Fig. 2. Temperal trends in meteorological data for Tajikistan: (a) precipitation; (b) temperature; (c) ground solar irradiance; (d) atmosphere solar irradiance; (e) snowfall; (f) snow mass; (g) cloud cover fraction; and (h) air density (from Renewables.ninja, 2023).

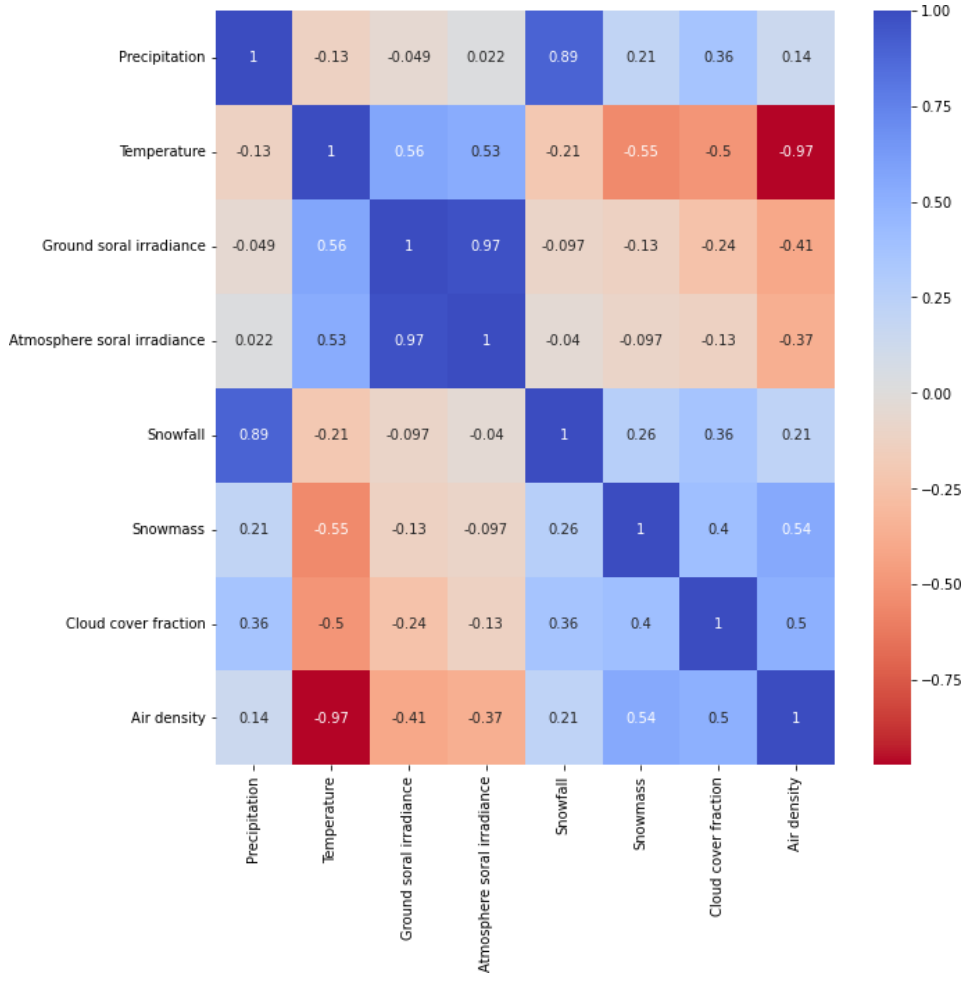

Fig. 3. Correlation coefficients among meteorological values for the prediction model based on the XGBoost algorithm.

CART 모델을 훈련시키기 위한 목적함수는 식 (2)와 같으며, $l(y_i, \hat{y_i})$는 실제값과 예측값의 차, Ω는 과적합을 방지하기 위한 모델의 정규화 함수(regularization)를 나타낸다. XGBoost의 첨가(additive) 방식과 테일러 확장(Taylor expansion)을 사용하여 t 단계의 목적함수 식을 표현하면 식 (3)과 같다. 여기서, $f_t(x_i)$는 t번째 반복 트리에 i번째 샘플에 대한 예측값, $\Omega(f_i)$는 과적합 방지를 위한 인자, 그리고 g_i와 h_i는 $\hat{y_i}^{(t-1)}$에 대한 1차 및 2차 편미분 값으로써, g함수는 $g_i = \delta_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$, h함수는 $h_i = \delta^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$로 정의할 수 있다. 식 (3)을 이용하여 모델을 훈련과 검증에 사용하고, 모델의 최적화 및 최적의 분할 지점을 찾기 위해 greedy 알고리즘이나 approximate 알고리즘 등을 사용한다(Chen and Guestrin, 2016; Yoon, 2020; An, 2021).

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

$$obj(t) = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f^2(x_i) \right] + \Omega(f_i) \tag{3}$$

일사량 예측 모델

일반 예측 모델에서는 Fig. 4a와 같이 입력 변수와 출력 변수 관계에 시간이 고려되지 않기 때문에 동일 시점에서 예측 모델의 훈련과 검증이 수행된다. 시계열 분석의 경우도 이와 마찬가지로 동일 시점에서 예측 모델을 수행하기도 하지만 Fig. 4b와 같이 이전 시점의 자료를 활용하여 그 이후를 예측하는 방법도 많이 사용하기도 한다. 예를 들면, 기상 관측과 같은 시계열 자료는 입력 변수나 시간 간격에 따라 여러 가지 방법으로 예측 모델을 구축하기도 한다. 본 연구에서는 일사량을 예측하기 위하여 입력 변수는 고정하고 출력 변수를 이전 시간으로 이동시켜 학습하여 과거의 값들로부터 미래의 값을 예측하는 방법이 사용되었다.

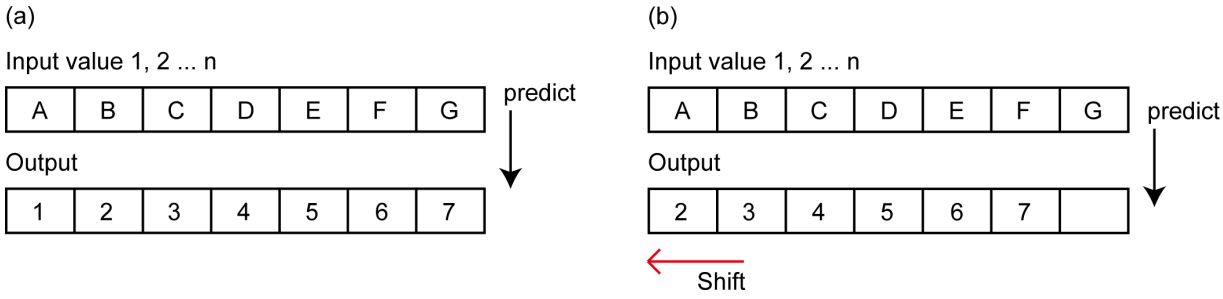

Fig. 4. Examples of the relationship between input and output variables in the prediction model for (a) the general model and (b) the time series model.

예측 모델 성능을 높이기 위해서는 초매개변수를 최적화해야 하는데, 이를 위하여 그리드 탐색 방법과 시계열 교차검증을 사용하였다. 그리드 탐색 방법은 초매개변수를 일정 구간으로 구분하여 최적의 값을 산정하는 방식이다(Fig. 5a). 전후 데이터 사이의 상관관계가 존재하는 시계열 데이터는 초매개변수 최적화 시 기존에 사용하는 교차검증 방법을 사용하기에 무리가 있다. 따라서 이러한 경우에는 시간순으로 나열된 데이터를 보존하면서 훈련용 데이터는 시험용 데이터보다 앞선 시간에 연속적으로 할당되어 초매개변수를 검증하는 방법이 사용되며, 이러한 방법이 Fig. 5b에서 보여주는 시계열 교차검증이다.

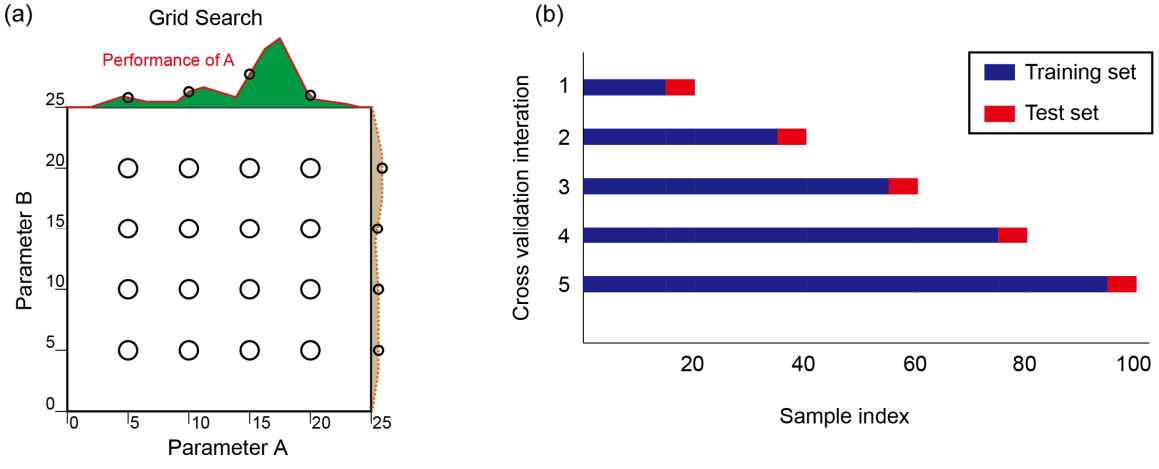

Fig. 5. Methods for hyperparameter optimization: (a) grid search and (b) time series cross-validation.

일사량 예측 모델 결과

타지키스탄의 일사량을 예측하기 위하여 예측 모델의 초매개변수 최적화는 데이터 수를 고려하여 학습에 필요한 4개의 초매개변수를 선정하였다(Table 2). 초매개변수 최적화를 수행한 결과, 최대깊이(max depth)는 9, 학습률(learning rate)은 0.05, 학습기 개수(n_estimators)는 3,000, 감마(gamma)는 0으로 나타났다. 타지키스탄 일사량을 예측한 결과를 나타내면 Fig. 6과 같다. Fig. 6a는 훈련모델(train model), Fig. 6b는 검증모델(validation model), 그리고 Fig. 6c는 시험모델(test model)의 결과를 각각 보여준다. 그림에서 파란색 점은 실제값, 주황색 점은 XGBoost로 일사량을 예측한 값을 나타낸다. 각 예측 모델의 결과로부터 알 수 있듯이, 시간과 계절에 따른 일사량의 계절성(seasonality)이 실제값과 예측값 모두에서 뚜렷하게 구분되는 것을 확인하였다. 또한 그 값이 매우 유사하게 나타나는 것을 확인할 수 있다. 대부분 타지키스탄 주요 도시들의 최대 일사량이 6~7월에 나타난 점을 고려하여 시험모델(test model)의 2016, 2017, 2018, 그리고 2019년 등 각 연도의 7월 1일 시간당 일사량의 실제값과 예측값을 Fig. 7로 도시하였다. 2016년 7월 1일 일사량의 최대 실제값은 약 1,005 W/m^2, 최대 예측값은 1,009 W/m^2, 2017년 7월 1일에는 최대 실제값이 939 W/m^2, 최대 예측값이 997 W/m^{2}으로 나타났다. 2018년 7월 1일 일사량의 최대 예측값은 1,022 W/m^2, 최대 예측값은 1,012 W/m^2, 2019년 7월 1일에는 최대 실제값이 1,055 W/m^2, 최대 예측값이 1,019 W/m^{2}으로 나타났다. 그래프에서 알 수 있듯이, 실제값과 예측값 사이의 차는 4~58 W/m^{2}의 범위며, 오차는 약 0.4~5.8%로 매우 유사한 결과를 나타냈다(Table 3).

Table 2. Results of hyperparameter optimization for predicting solar irradiance in Tajikistan

Maximum depth	Learning rate	n_estimators	Gamma
9	0.05	3,000	0

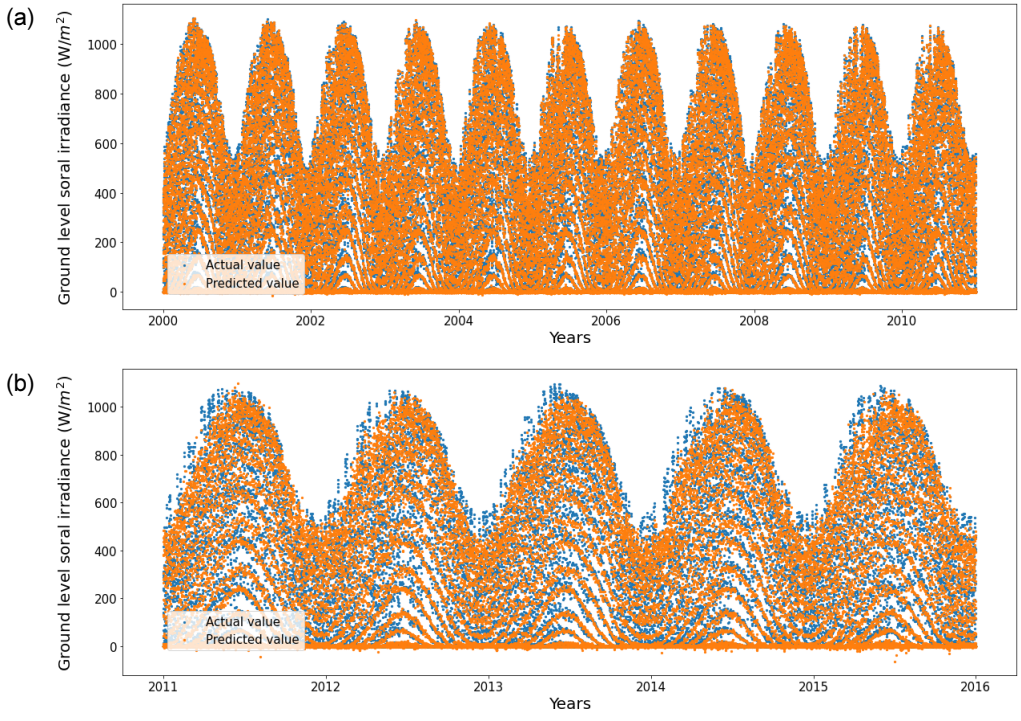

Fig. 6. Results of the prediction model for ground-level solar irradiance in Tajikistan for the (a) training model, (b) validation model, and (c) test model.

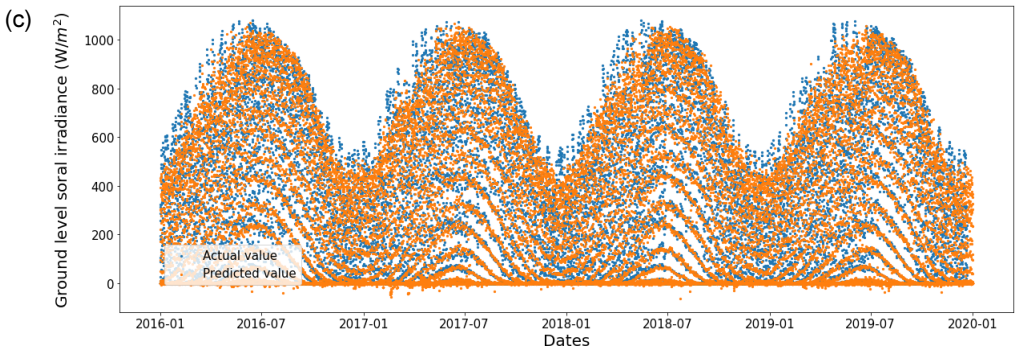

Fig. 6. Continued.

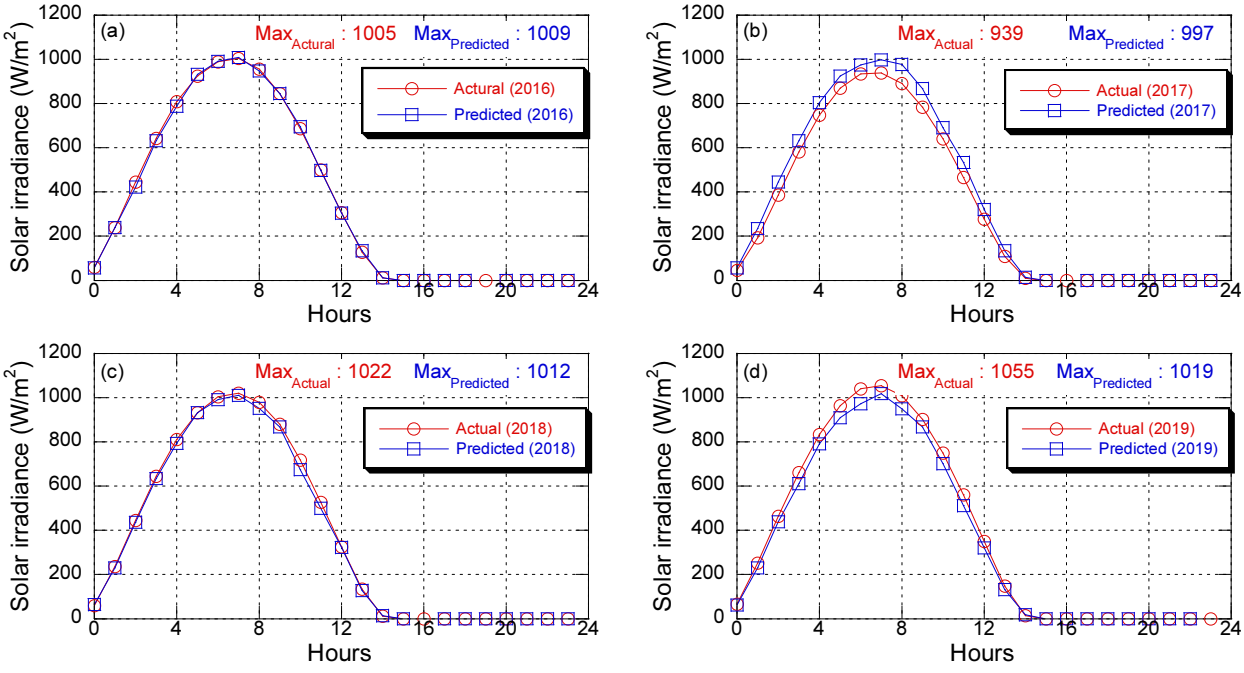

Fig. 7. Observed and predicted solar irradiance values for 1 July in (a) 2016, (b) 2017, (c) 2018, and (d) 2019.

Table 3. Observed and predicted solar irradiance values, and associated errors

2016			2017			2018			2019		
Actual (W/m^2)	Predicted (W/m^2)	Error (%)	Actual (W/m^2)	Predicted (W/m^2)	Error (%)	Actual (W/m^2)	Predicted (W/m^2)	Error (%)	Actual (W/m^2)	Predicted (W/m^2)	Error (%)
1,005	1,009	0.40	939	997	5.81	1,022	1,012	0.98	1,055	1,019	3.41

결 론

타지키스탄에서 재생에너지 자원으로써 복사 태양에너지의 활용 가능성을 평가하기 위하여 머신러닝의 앙상블 모델 중 하나인 XGBoost를 이용하여 타지키스탄의 일사량을 예측하였다. 그 결과를 요약하면 다음과 같다.

타지키스탄의 일사량을 훈련모델, 검증모델, 시험모델을 통해 예측한 결과, 시간과 계절에 따른 일사량의 계절성이 실

제값과 예측값 모두에서 확인 가능하였으며, 그 값도 매우 유사하게 나타났다. 타지키스탄 주요 도시들의 최대 일사량이 6~7월인 점을 고려하여 시험모델을 이용하여 2016, 2017, 2018, 그리고 2019년 등 각 연도의 7월 1일 시간당 일사량의 실제값과 예측값을 계산한 결과, 2016년 일사량의 최대 실제값과 예측값은 약 1,005 W/m^{2}과 1,009 W/m^2, 2017년에는 939 W/m^{2}과 997 W/m^2, 2018년에는 1,022 W/m^{2}과 1,012 W/m^2, 2019년에는 1,055 W/m^{2}과 1,019 W/m^{2}으로 실제값과 예측값의 오차가 약 0.4~5.8%로 매우 비슷한 결과를 나타냈다. 이상의 결과를 종합해볼 때, 타지키스탄의 일사량을 예측하는 데 있어 향후 XGBoost가 매우 유용한 도구로 활용될 수 있을 것이며, 더 많은 자료가 확보된다면 결과의 정확도는 더 높아질 것으로 판단된다.

References

An, K.M., 2021, Developing a prediction model for firm innovation and performance using statistical matching and machine learning ensemble techniques, Doctoral Dissertation, Dongguk University, 98-105 (in Korean with English abstract).

Chen, T., Guestrin, C., 2016, XGBoost: A scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17 August 2016, San Francisco, CA, USA, 785-794.

IEA (International Energy Agency), 2022, Tajikistan 2022: Energy sector review, 134p.

Joint Ministries, 2022, Tajikistan country partnership strategy.

Kudusov, M.A., Madvaliev, U., Elistratov, V.V., 2021, Evaluation of the efficiency of already existing network solar photovoltaic plants operating 24/7/365 in low-voltage power supply systems of social facilities in the city of Dushanbe, Applied Solar Energy, 57(4), 323-332.

Na, T., Noh, J., Kim, H., Kang, S.S., 2023, Analysis of climate, weather, solar radiation and solar energy in major cities of Tajikistan, The Journal of Engineering Geology, Submitted (in Korean with English abstract).

Pfenninger, S., Staffell, I., 2016, Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data, Energy, 114, 1251-1265.

Renewables.ninja, 2023, Meteorological data for Tajikistan, Retrieved from https://renewables.ninja/.

Yoon, H.R., 2020, A empirical study on the financial stability prediction model of South Korea's public enterprises with machine learning techniques, Doctoral Dissertation, Hansung University, 36-43 (in Korean with English abstract).