ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Identification of a spatial distribution threshold for the development of a solar radiation model using deep neural networks

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Identification of a spatial distribution threshold for the development of a solar radiation model using deep neural networks

Dae Gyoon Kang[1], **Yean-Uk Kim**[2,3], **Shinwoo Hyun**[4], **Kwang Soo Kim**[3,4,*], **Junhwan Kim**[5], **Chung-Kuen Lee**[6], **Atsushi Maruyama**[7], **Robert M Beresford**[8] and **David H Fleisher**[9]

[1] National Center for Agro-Meteorology, Seoul, Republic of Korea
[2] Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany
[3] Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea
[4] Department of Agriculture, Forestry and Bioresources, Seoul National University, Seoul, Republic of Korea
[5] Korea National University of Agriculture and Fisheries, Jeonju-si, Jeollabuk-do, Republic of Korea
[6] Division of Crop Post-Harvest Technology Research, National Institute of Crop Science, Suwon-si, Kyeonggi-do, Republic of Korea
[7] Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan
[8] The New Zealand Institute for Plant and Food Research Limited, Private Bag 92169, Auckland 1142, New Zealand
[9] Adaptive Cropping Systems Laboratory, United States Department of Agriculture—Agricultural Research Service (USDA-ARS), Beltsville, MD, United States of America
* Author to whom any correspondence should be addressed.

E-mail: luxkwang@snu.ac.kr

Keywords: solar radiation, spatial portability, artificial intelligence, empirical model, crop model

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
13 April 2023

REVISED
8 August 2023

ACCEPTED FOR PUBLICATION
5 September 2023

PUBLISHED
28 September 2023

Abstract

We propose an approach to develop a solar radiation model with spatial portability based on deep neural networks (DNNs). Weather station networks in South Korea between 33.5–37.9° N latitude were used to collect data for development and internal testing of the DNNs, respectively. Multiple sets of weather station data were selected for cross-validation of the DNNs by standard distance deviation (SDD) among training sites. The DNNs tended to have greater spatial portability when a threshold of spatial dispersion among training sites, e.g. 190 km of SDD, was met. The final formulation of the deep solar radiation (DSR) model was obtained from training sites associated with the threshold of SDD. The DSR model had RMSE values <4 MJ m^{-2} d^{-1} at external test sites in Japan that were within $\pm 6°$ of the latitude boundary of the training sites. The relative difference between the outputs of crop yield simulations using observed versus estimated solar radiation inputs from the DSR model was about 4% at the test sites within the given boundary. These results indicate that the identification of the spatial dispersion threshold among training sites would aid the development of DNN models with reasonable spatial portability for estimation of solar radiation.

1. Introduction

Solar radiation is one of the key variables that determine the productivity of agricultural ecosystems (Yang *et al* 2019). It is often measured using electrical sensors, e.g. pyranometers, installed at weather stations. However, observed data are not readily available in most areas due to high costs and technical limitations for the installation and maintenance of sensors (Fan *et al* 2018). For example, Thornton and Running (1999) reported that solar radiation is measured at about 0.2% of weather stations across the globe. Weather stations even in developed countries are equipped with solar radiation sensors at a low rate, e.g. about 1% and 4% of weather stations in the United States (Wang *et al* 2016) and South Korea (Kang *et al* 2019), respectively.

Models have been developed to estimate solar radiation using observed weather data available at a site of interest. Empirical models have been commonly used for this purpose due to their simplicity, convenience, and availability (Zhang *et al* 2019).

These models can be classified into four groups by input variable (Besharat *et al* 2013): sunshine duration-based models (Ångström 1924, Hamon *et al* 1954), temperature-based models (Bristow and Campbell 1984), cloud cover-based model (Black 1956), and other meteorological parameters-based models (Reddy 1987).

Application of machine learning methods can aid the development of a solar radiation model using commonly available weather data such as temperature and precipitation. Artificial neural networks (ANNs), which consist of input layers, hidden layers, and output layers, have been used to predict solar radiation (Çelik *et al* 2016, Antonopoulos *et al* 2019). Ghimire *et al* (2019) reported that ANN outperformed other machine learning methods including support vector regression, Gaussian process, and genetic programming approaches. ANN models would have greater accuracy than the other types of empirical models because ANN can reflect the complex nonlinear relationships between target and input variables (Abdelkawy *et al* 2020).

The training data for solar radiation models are often gathered from weather stations in a specific region. As a result, these models tended to have small errors of solar radiation estimates in the regions where the training data were obtained. In contrast, the accuracy of the given model can decrease when applied to other regions. For example, Muneer *et al* (2007) reported that cloud-based models had relatively greater accuracy in estimation of solar radiation using locally fitted coefficients instead of coefficients generalized over a larger region.

To the best of our knowledge, the spatial boundary of empirical models is rarely assessed during the model development stage. Here we propose an approach to develop a solar radiation model using spatial properties among weather stations where training data are collected. It was hypothesized that a threshold of the spatial distribution among training sites would exist to ensure spatial portability of the model. The research questions in the present study include:

(1) What is the error of the empirical solar radiation model taking into account the spatial distribution of weather stations where training data were obtained,
(2) What is the spatial boundary of the given model,
(3) What is the impact of the model on preparation of inputs for a crop model used for the assessment of agricultural productivity?

These questions were examined using data collected from weather stations in South Korea and Japan as inputs to deep neural networks (DNNs), which is a type of ANN that has multiple hidden layers between input and output layers. This paper is organized as follows: The DNN model is introduced in section 2.1 through 2.6 with the procedures for evaluation of spatial portability and training of the DNN model. The model outputs are compared with observed solar radiation data from sites where no training data were obtained in section 3. In section 3, spatial portability of the DNN model was also examined comparing rice yield predictions using observed versus simulated solar radiation data. A discussion on our findings was provided in section 4. Finally, the conclusion in section 5 summarizes the key findings and highlights the significance of further studies.

2. Materials and methods

2.1. Collection of weather data

Daily data measured at synoptic weather stations were collected in South Korea and Japan (figure 1). These data included maximum and minimum temperatures, precipitation, and solar radiation. Weather data measured in these countries were used as internal and external data for model building and testing, respectively (table 1). The internal data were used for the cross-validation and internal test of solar radiation models whereas the external data were used for an external test. The internal data were downloaded from the Korea Meteorological Administration website (https://data.kma.go.kr). These data were obtained from 20 weather stations where long-term measurements of solar radiation (>26 years) were available in South Korea. The external data were obtained from 49 weather stations in Japan using the Japan Meteorological Agency website (www.jma.go.jp).

2.2. Preparation of inputs

DNNs were used to develop a solar radiation model that requires a minimum set of weather variables (figure 2). DNNs have advantages in learning complex relationships between variables because a large number of hidden layers would allow for accurate estimation connecting a small number of input variables in a dense network (Ryu *et al* 2017). In the present study, DNNs were developed to estimate solar radiation using only two meteorological inputs including temperature and precipitation (figure 3). Geographic coordinate and date were also used as additional inputs to DNNs. Furthermore, these inputs to DNNs were transformed to account for the impact of topographic and geographic properties at a given site on estimation of solar radiation. As a result, potential temperatures of minimum and maximum temperature, precipitation and extraterrestrial solar radiation were used as the actual inputs to the DNNs (see supplementary information 1).

Potential temperatures (Θ) at a reference atmospheric pressure level of 100 kPa (P_{1K}) were used to

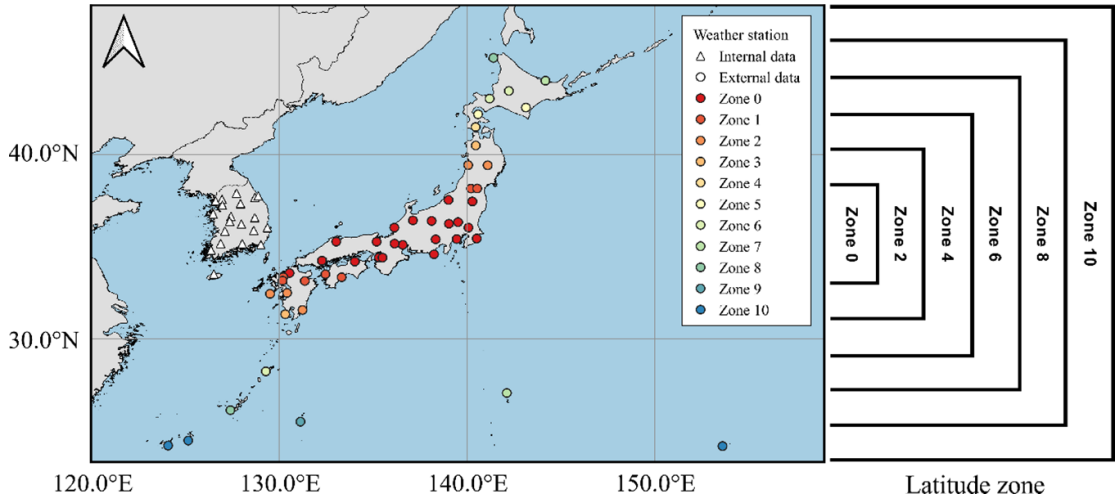

Figure 1. Spatial distribution of weather stations used to collect weather data for the development and the external test of the solar radiation estimation model. Internal data was used to train and validate deep neural network (DNN) and external data was used to external test for DNN. The weather stations were allocated to the zones identified by the latitude of weather stations. For example, Zone 0 and Zone 6 include the weather stations within 33.5–37.9° N and 27.5–43.9° N, respectively.

represent a reference temperature condition (Aybar-Ruiz *et al* 2016). Θ_{max} and Θ_{min} were determined for daily maximum and minimum temperatures as follows:

$$\Theta = T_o\left(\frac{P_{1K}}{P}\right)^{R/C_P} \quad (1)$$

where R and C_p indicate the gas constant of air (8.314 J K^{-1} mol^{-1}) and the specific heat capacity (1.005 kJ kg^{-1} K^{-1}), respectively. T_o is maximum (max) or minimum (min) temperature (K) measured at a weather station. The barometric pressure P (kPa) at the weather station with altitude alt (m) was estimated as follows (Azevedo and Crisóstomo 2016):

$$P = p_0\left(1 - \frac{L\cdot \text{alt}}{T_0}\right)^{\frac{gM}{RL}} \quad (2)$$

where p_0 and T_0 indicate standard atmospheric pressure (101.3 kPa) and standard temperature (288.16 K) at sea level, respectively. L, g, and M indicate the temperature lapse rate for dry air (0.00976 K m^{-1}), gravitational acceleration (9.81 m s^{-2}) and the molar mass of dry air (0.0290 kg mol^{-1}), respectively.

The extraterrestrial solar radiation (R_a; W m^{-2}), which is solar radiation at the top of the atmosphere, was determined using date and geographic coordinates at a given site as follows (King *et al* 2015):

$$R_a = \frac{86400}{\pi} I_{SC}\left[1 + 0.033\cos\left(2\pi\frac{\text{doy}}{365}\right)\right] \times \left[\cos\phi\,\cos\delta\,\sin\omega_s + \omega_s\,\sin\phi\,\sin\delta\right] \quad (3)$$

where I_{sc} and ϕ indicate the solar constant (1367 W m^{-2}) and latitude, respectively. The solar declination δ and the mean sunrise hour angle ω_s were calculated as follows (Teke *et al* 2015):

$$\delta = 0.409\,\sin\left(2\pi\frac{\text{doy}}{365} - 1.39\right) \quad (4)$$

and

$$\omega_s = \frac{\pi}{180}\arccos\left(-\tan\phi\,\tan\delta\right). \quad (5)$$

2.3. Identification of training site sets by spatial distribution of weather stations

The weather stations included in the internal data were grouped to identify the minimum level of spatial distribution among training sites for improvement in spatial portability of solar radiation models. The standard distance deviation (SDD) among weather stations was determined to represent the spatial distribution of training sites. The SDD has been used to quantify the degree of spatial dispersion across localities (Christodoulakis *et al* 2018). The SDD was calculated as follows (Hu *et al* 2014):

$$\text{SDD} = \sqrt{\frac{\sum(\lambda_j - \bar{\lambda})^2 + \sum(\phi_j - \bar{\phi})^2}{n_{ws}}} \quad (7)$$

where λ_j and ϕ_j are longitude and latitude of an individual weather station j, respectively, and $\bar{\lambda}$ and $\bar{\phi}$ represent the average longitude and latitude for all weather stations, respectively. The total number of weather stations, n_{ws}, was 15 in the present study. The SDD value was determined for every combination of these 15 weather stations included in the 20 synoptic weather stations. In total, 11 sets of weather stations were selected to have the SDD at each decile as well as at the minimum (143 km) and maximum (216 km) values (figure 4).

Table 1. Summary of the weather data collected from the weather stations in figure 1.

Site	Start[a]	End[a]	n[b]	Zone[c]	Site	Start	End	n	Zone
South Korea (Used for cross-validation and internal test)									
Jeonju	1982	2017	13 024	0	Incheon	1982	2017	12 535	0
Jeju	1982	2017	12 930	0	Andong	1983	2017	12 505	0
Mokpo	1982	2017	12 929	0	Daegu	1982	2017	12 387	0
Gwangju	1982	2017	12 923	0	Chuncheon	1982	2016	12 153	0
Pohang	1982	2017	12 773	0	Daejeon	1984	2017	12 074	0
Suwon	1982	2017	12 681	0	Jinju	1982	2015	11 978	0
Busan	1982	2017	12 656	0	Daegwallyeong	1982	2015	11 674	0
Seoul	1982	2017	12 616	0	Wonju	1982	2015	11 648	0
Seosan	1982	2017	12 562	0	Chupungnyeong	1982	2015	10 467	0
Cheongju	1982	2017	12 560	0	Gangneung	1982	2008	9304	0
Japan (Used for external test)									
Watsukanai	2011	2020	3652	8	Matsue	2011	2020	3648	0
Asahikawa	2011	2020	3647	6	Maizuru	2011	2013	818	0
Abashiri	2011	2020	3651	7	Hikone	2011	2020	3647	0
Sapporo	2011	2020	3651	6	Shimonoseki	2011	2020	3634	0
Obihiro	2011	2020	3643	5	Hiroshima	2011	2020	3650	0
Muroran	2011	2020	3650	5	Osaka	2011	2020	3647	0
Hakodate	2011	2020	3652	4	Nara	2011	2020	3578	0
Aomori	2011	2020	3648	3	Fukuoka	2011	2020	3650	1
Akita	2011	2020	3649	2	Saga	2011	2020	3647	1
Morioka	2011	2020	3646	2	Oita	2011	2020	3651	1
Yamagata	2011	2020	3652	1	Nagasaki	2011	2020	3649	2
Sendai	2011	2020	3651	1	Kumamoto	2011	2020	3638	2
Fukushima	2011	2020	3651	0	Kagoshima	2011	2020	3646	3
Niigata	2011	2020	3471	0	Miyazaki	2011	2020	3636	2
Toyama	2011	2020	3650	0	Matsuyama	2011	2020	3649	1
Nagano	2011	2020	3652	0	Takamatsu	2011	2020	3649	0
Utsunomiya	2011	2020	3652	0	Kochi	2011	2020	3650	1
Fukui	2011	2020	3650	0	Naze	2011	2020	3649	6
Maebashi	2011	2020	3641	0	Ishigakijima	2011	2020	3651	10
Nagoya	2011	2020	3652	0	Miyakojima	2011	2020	3652	10
Kofu	2011	2020	3650	0	Naha	2011	2020	3652	8
Tsukuba	2011	2020	3649	0	Minami-Daito	2011	2020	3640	9
Choshi	2011	2020	3643	0	Chichijima	2011	2020	3651	7
Shizuoka	2011	2020	3649	0	Marcus island	2011	2020	3651	10
Tokyo	2011	2020	3306	0	—	—	—	—	—

[a] Start and End are the first and the last years in which data were collected.

[b] n is the number of quality-checked data.

[c] Zone indicates the boundary identified by the latitude of weather stations. Zone 0 indicates the upper and lower boundary of South Korea.

2.4. Cross-validation of DNNs by training site set

Eleven preliminary models were obtained using the cross-validation procedure for the given training site data sets identified by the SDD value. Cross-validation has been used to minimize the selection bias for the development of DNNs (Renno *et al* 2015). In N-fold cross-validation, a whole dataset is randomly split into N groups for training and validation. The training and validation sets are prepared choosing N-1 subsets and the remaining data, respectively.

Cross-validation was performed as an alternative method to compare the central tendency of error statistics among the preliminary models (figure 5). The internal data was split into two groups for cross-validation and the internal test. Sixty percent of the weather data from each site were allocated randomly for cross-validation. The remaining data were reserved for the internal test. This approach ensures that a fraction of weather data at every site is left unknown or independent to the models.

In each cross-validation procedure, multiple candidate DNNs were obtained using training data (figure 5(a)). Training data were split into subgroups to obtain five DNNs. Each DNN was obtained using the gradient descent method, which searches the optimum set of parameter values through the back-propagation algorithm (Quej *et al* 2017). The learning rate, which determines the range of weight adjustment for the neural network, was set to be 0.000 05.

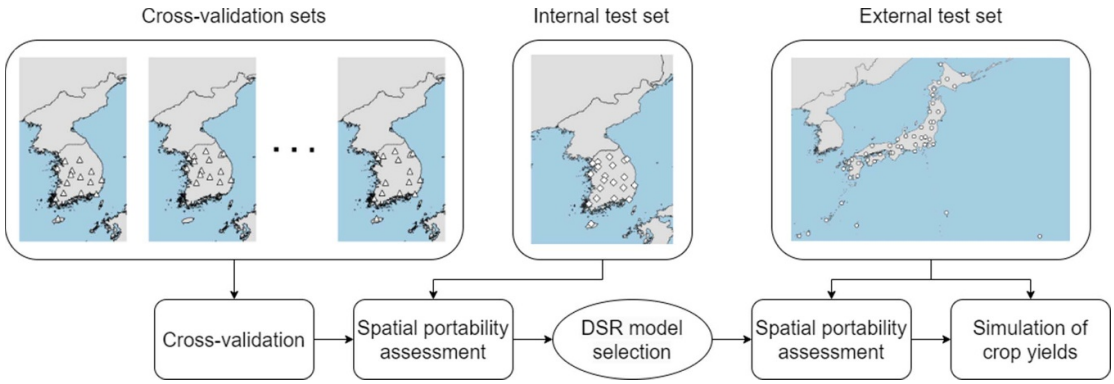

Figure 2. The process for development and validation of the deep neural network (DNN) model for estimation of solar radiation. Data sets for cross-validation and Internal test were obtained from weather stations in South Korea. Candidate models obtained from cross-validation were evaluated in terms of spatial portability using the internal test set. The model that had the greater spatial portability than other candidate models was chosen to be the deep solar radiation (DSR) model. Spatial portability of DSR model was assessed using the external test set, which consist of weather observation data in Japan.

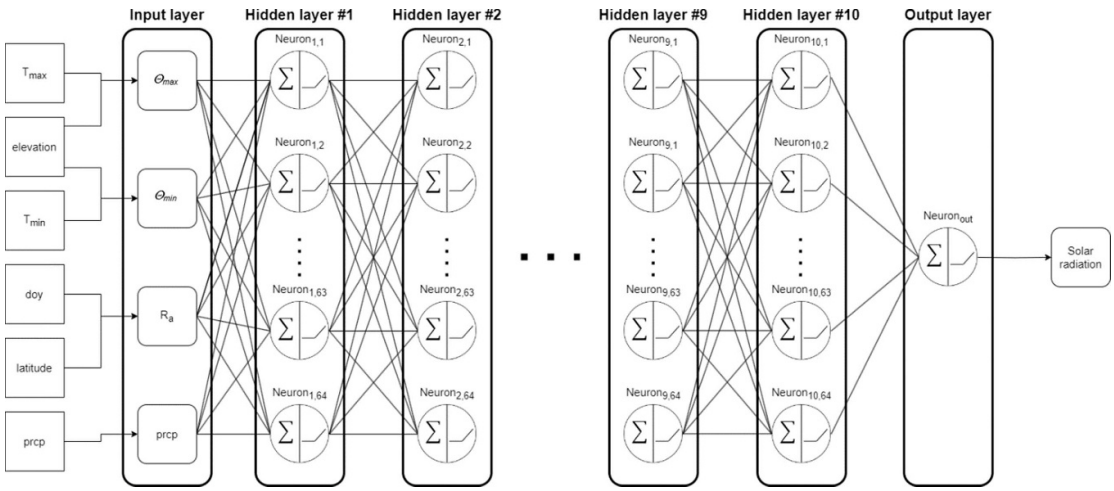

Figure 3. The overview of deep neural network architecture for estimation of daily solar radiation. The meteorological input data include T_{max}, T_{min}, and *prcp*, which represent daily maximum and minimum temperature and precipitation, respectively. These input data were converted into the input variables to the DNN including Θ_{max}, Θ_{min}, R_a, which indicate potential temperatures of daily maximum and minimum temperatures and extraterrestrial solar radiation, respectively. Geographic and topographic properties of a site including latitude, and elevation were used to determine these variables along with day of year (doy). Θ_{max} and Θ_{min} were calculated using T_{max} and T_{min} at a given elevation, respectively. The DNN model had 10 hidden layers and 64 nodes for each layer. ReLU was set to be the activation function. No bias was used in the present study.

The root mean square error was used to determine the loss for each training process as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum(\mathrm{est}_i - \mathrm{obs}_i)^2}{n}} \quad (8)$$

where obs_i and est_i indicate observed and estimated values of solar radiation (MJ m^{-2} d^{-1}) using the neural network for the total number of records n in the given training set. The Adam optimization method, which has been recommended for deep learning (Bock *et al* 2018), was used to update the parameter values.

The preliminary model was selected from five DNNs using the validation set split into five subsets (figure 5(b)). The RMSE values were calculated for each subset. One of the DNNs was chosen to have the minimum of the RMSE value on average for further analysis of spatial portability.

The DNNs for solar radiation estimation were implemented using TensorFlow (Google Inc., Mountain View, CA, USA). TensorFlow has a high degree of flexibility in organizing the structure of neural networks, which makes it suitable for the development of a solar radiation estimation model (Zang *et al* 2022). For example, Kaba *et al* (2018) used TensorFlow to develop a model to estimate daily solar radiation using sunshine duration, cloud cover, and other weather variables.

2.5. Spatial portability assessment of a deep solar radiation model

The internal test set was used to evaluate spatial portability of 11 preliminary models obtained from the

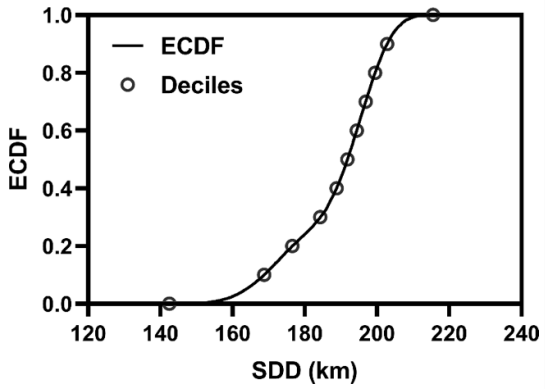

Figure 4. Empirical cumulative density function (ECDF) of the standard distance deviation (SDD) for the given set of weather stations, which is a combination of 15 weather stations selected from the 20 synoptic weather stations in South Korea.

cross-validation procedures. The RMSE values were initially determined for each weather station. These RMSE values were subsequently averaged across all weather stations and denoted by m_{RMSE} to represent the overall error across sites for the individual preliminary models. The coefficient of variation was also determined to indicate the spatial variation of RMSE as follows:

$$\mathrm{CV}_{\mathrm{RMSE}} = \frac{\mathrm{sd}_{\mathrm{RMSE}}}{m_{\mathrm{RMSE}}} \tag{9}$$

where $\mathrm{sd}_{\mathrm{RMSE}}$ indicates the standard deviation of the RMSE by site. A lower value of m_{RMSE} *and* $\mathrm{CV}_{\mathrm{RMSE}}$ indicates higher spatial portability of model across sites.

The final solar radiation model, referred as the deep solar radiation (DSR) model, was chosen from the 11 preliminary models for further analysis. The relationships between the errors of the preliminary models and SDD were obtained using the internal test set. A segmented regression analysis was performed between m_{RMSE} and $\mathrm{CV}_{\mathrm{RMSE}}$, and SDD to examine if a threshold value of SDD exists for improvement in spatial portability. Under such an assumption, the model with the value of SDD near the breaking point where the slope of the regression model changes, e.g. from negative to positive, was chosen to be the DSR model. Such regression analysis was carried out using the *segmented* R package (Muggeo 2017).

Spatial portability of the DSR model was evaluated using the external test set. The error statistics were grouped by weather station zones to examine the spatial boundary for the DSR model (figure 1). Zone 0 was set to include weather stations in South Korea in which latitude ranged from 33.5 to 37.9° N. The boundary of the following zones increased by 1° of latitude. For example, Zone 1 and Zone 6 were defined within 32.5–38.9° N and 27.5–43.9° N, respectively. In addition, the errors of the DSR models were analyzed in terms of distance from the sea (DFS) of each weather station included in the external test set (see supplementary information 2).

2.6. Simulation of crop yield using observed and estimated solar radiation

Crop growth simulations were performed to evaluate the applicability of the DSR model for assessment of agricultural ecosystem productivity. In the present study, the ORYZA2000 model (Bouman and van Lear 2006) was used to simulate rice yield using observed and estimated solar radiation data at each of the weather stations (see supplementary information 3). Crop growth was simulated under common crop management options for japonica rice in Japan. Transplanting dates for rice grown near each weather station were obtained from the Ministry of Agriculture, Forestry and Fisheries (MAFF) where yearly statistics of the cultivation schedule of rice are provided (www.maff.go.jp/j/tokei/kouhyou/sakumotu/sakkyou_kome/index.html). Seedbed duration was set to be 30 d. The number of hills per m^2 and the number of plants per hill were set to be 3 and 16.67, respectively. Nitrogen fertilizer was set to be 90 kg N ha^{-1}, which was split as a 50:20:30 application ratio at basal, tillering, and panicle initiation stage, respectively. It was assumed that no water stress occurred because rice was fully irrigated in Japan. The leading rice cultivar data by prefecture was collected from the statistical yearbooks available on the MAFF website (www.maff.go.jp/j/tokei/kouhyou/syokuryo_nenkan/). The rice cultivars were classified into three maturity groups: early, medium, and mid-late. The cultivar parameters of the three maturity groups were obtained from Lee *et al* (2015).

The outputs from the ORYZA2000 model were grouped by weather input data for comparison purposes. The input data for the reference crop yield simulations were prepared using observed solar radiation at the weather stations in Japan. Another set of weather input data was generated using estimates of solar radiation obtained from the DSR model. The values of RMSE and normalized RMSE (NRMSE) were determined by comparing rice yield values obtained from these sets of input data. The value of NRMSE was calculated as follows:

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\bar{Y}_{\mathrm{Obs}}} \tag{10}$$

where $\bar{Y}_{\mathrm{Obs}}$ indicates the mean value of rice yield simulated using the solar radiation observation.

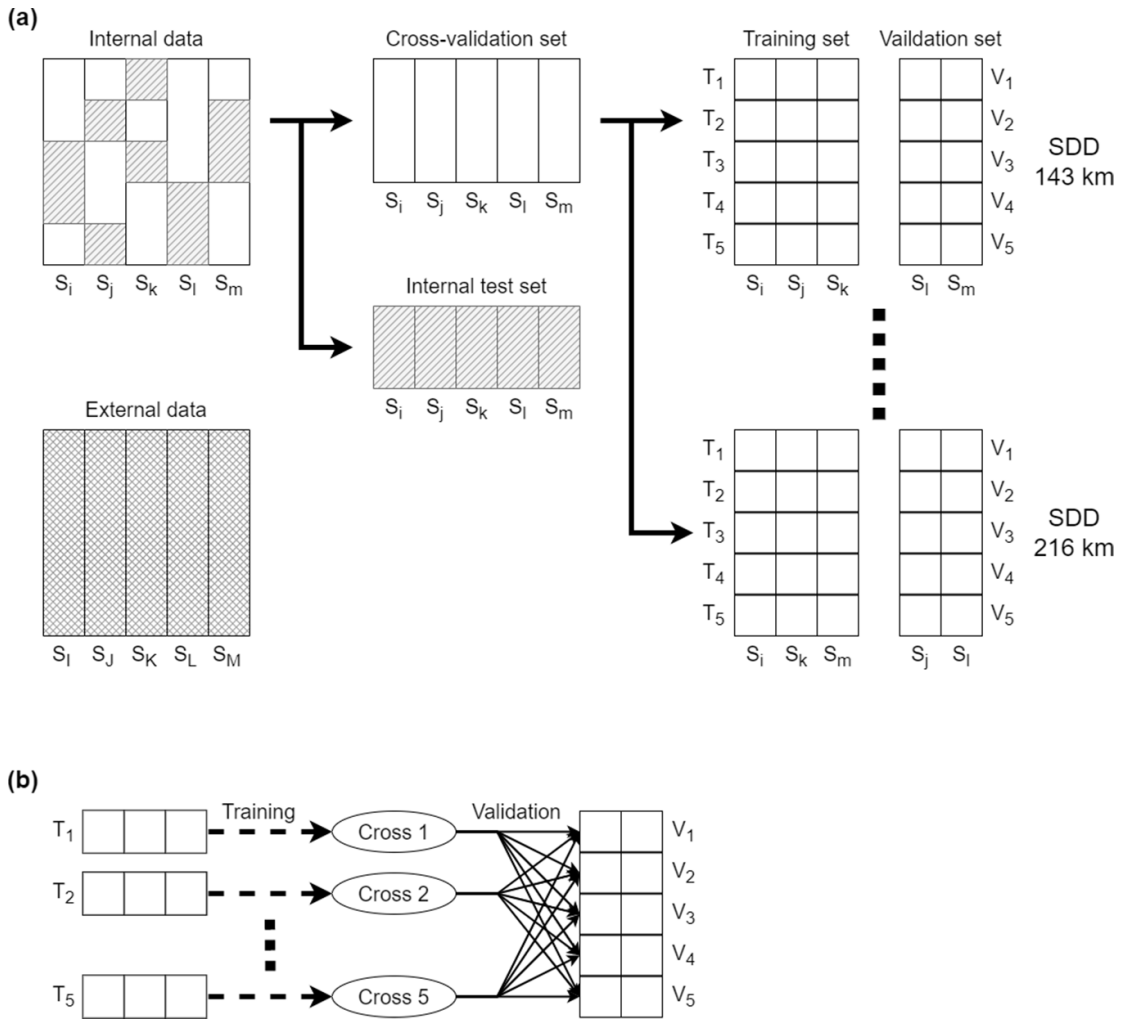

Figure 5. Simple representation of (a) data preparation and (b) cross-validation. Korean data were used as internal data whereas Japanese data were used for external testing. Internal data were classified into two groups for cross-validation and internal testing. The cross-validation dataset was further divided into training and validation sets by site. Training (S_i, S_j, S_k) and validation sites (S_l, S_m) were selected to be 60% and 40% of weather stations with specific standard distance deviation (SDD). These training (T) and validation (V) sets were split into five subgroups. As a result, five candidate models were obtained from cross-validation. Another combination of training sites with a different SDD (e.g. S_i, S_k, S_m) and validation sites (e.g. S_j, S_l) were selected to perform cross-validation and internal test, separately.

3. Results

3.1. Spatial portability of preliminary models for internal test set

The preliminary models had greater spatial portability for the internal test set when the given training sites had the values of SDD near the threshold value of 192 km (figure 6; see supplementary information 4). The value of m_{RMSE} decreased at the rate of 0.01 MJ m^{-2} d^{-1} with the increase of 10 km in SDD when the values of SDD were lower than 192 km. However, the value of m_{RMSE} increased at the rate of 0.005 MJ m^{-2} d^{-1} with the increase of 10 km in SDD when the values of SDD were higher than 192 km. The value of CV_{RMSE} also decreased at the rate of 0.018 with the increase of 10 km in SDD values up to 192 km. However, the rate decreased to 0.004 when the values of SDD were higher than 192 km. Under such relationships, the DSR model was chosen to be the solar radiation model with the value of SDD near 192 km for further analysis.

3.2. Spatial portability of the final model for the external test set

The error statistics of the DSR model differed by the latitude zone for the external test (figure 7). From the latitude zone of South Korea, i.e. Zone 0, (33.5–37.9° N), to Zone 6 (27.5–43.9° N), the DSR model had relatively smaller errors at the external test sites. For example, the m_{RMSE} and CV_{RMSE} of the DSR model were relatively similar. In contrast, these values of DSR tended to increase as the SDD value increased for the training sites. In particular, the DSR model had RMSE values <4 MJ m^{-2} d^{-1} at 73% of weather stations within Zone 6 (figure 8(a)). In contrast, the values of RMSE were >5 MJ m^{-2} d^{-1} at every weather station outside of Zone 6 (figure 8(a)).

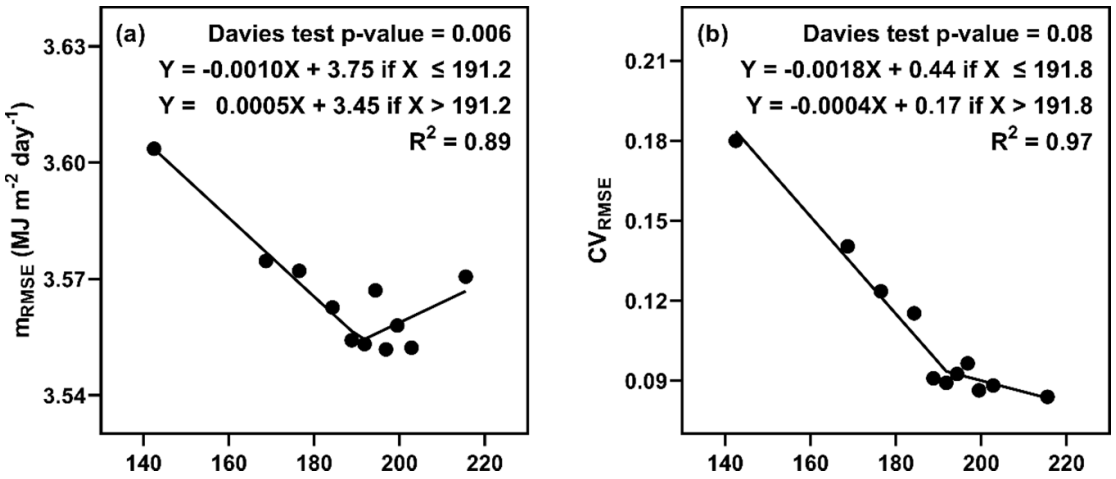

Figure 6. Relationships between the standard distance deviation (SDD) and the error mean (a) and variation (b) of the solar radiation models for the internal test set. m_{RMSE} and CV_{RMSE} are the mean and coefficient of variation of the root mean square error, respectively.

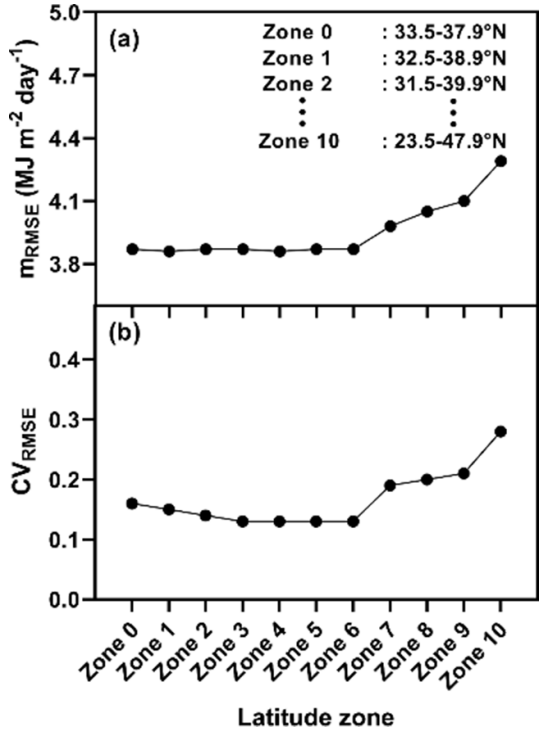

Figure 7. The error statistics of the deep solar radiation (DSR) model for the external test set within different latitude zones. (a) m_{RMSE} and (b) CV_{RMSE} indicate the mean and coefficient of variation of the root mean square error, respectively.

For example, Marcus Island had the largest error with RMSE of >9 MJ m^{-2} d^{-1} (figure 8(b)).

Spatial portability of the DSR model tended to decrease as the DFS decreased (figure 9). For example, m_{RMSE} and CV_{RMSE} of the DSR model were 3.5 MJ m^{-2} d^{-1} and 0.03, respectively, at weather stations with DFS $\geqslant$ 50 km within Zone 6. Those for weather stations with DFS $<$ 50 km within Zone 6 were 4.0 MJ m^{-2} d^{-1} and 0.13, respectively. The smallest error of RMSE of 3.3 MJ m^{-2} d^{-1} occurred at Utsunomiya station where DFS is 76 km (figure 8(c)). In contrast, the greatest error, e.g. RMSE of 6.2 MJ m^{-2} d^{-1}, within Zone 6 occurred at Choshi station where DFS is 9 km.

3.3. Application of the DSR model to crop growth simulations

The outcomes of the rice yield simulation had a relatively large degree of agreement within Zone 6 when observed and estimated solar radiation data were used as inputs to the crop model (figure 10(a)). In contrast, the discrepancy between simulation outputs increased at sites outside Zone 6 when input data were prepared using the estimates of solar radiation

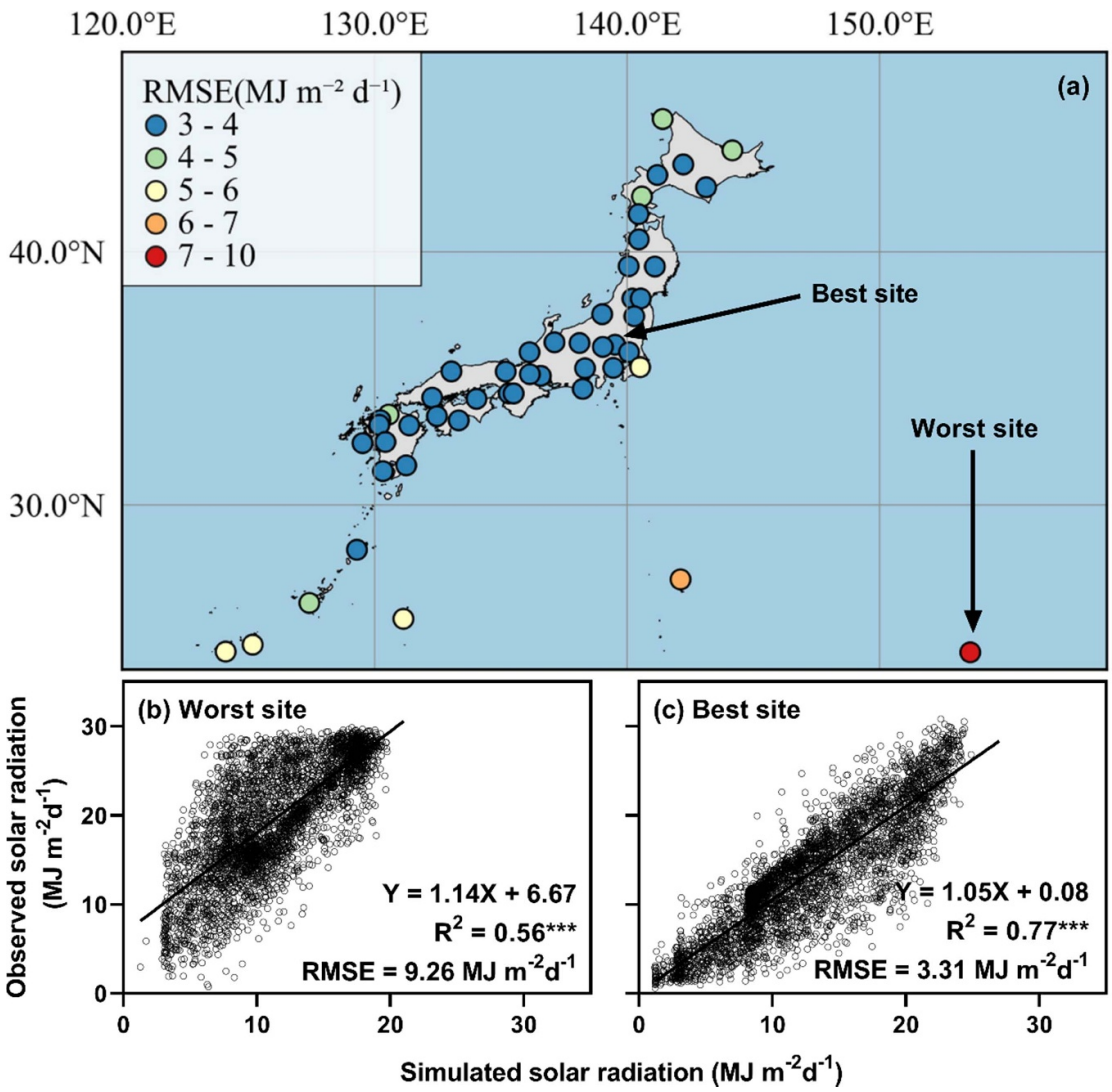

Figure 8. The errors of the deep solar radiation (DSR) model at (a) all the external test sites, (b) the worst sites and (c) the best sites. RMSE indicates root mean square error. *** represents $p < 0.001$.

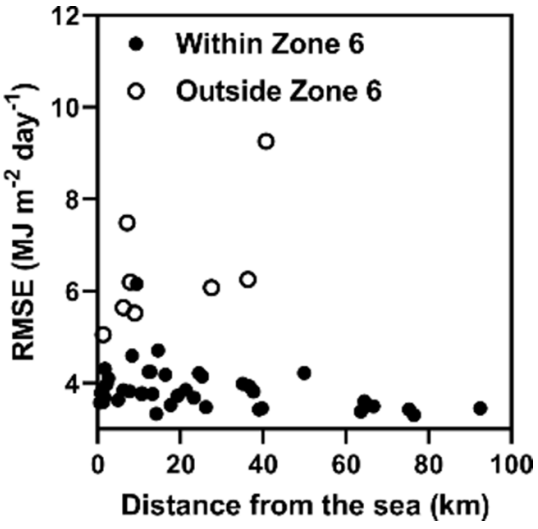

Figure 9. Relationship between the distance from the sea and the RMSE values of the DSR model for the external test set.

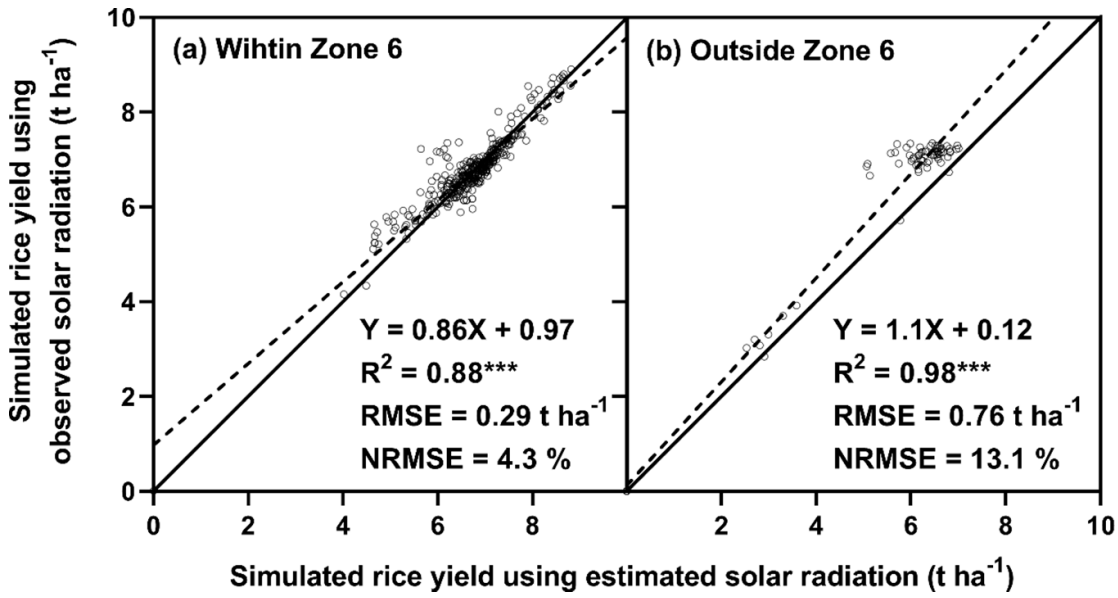

Figure 10. Scatter plots of simulated rice yield using observed solar radiation and simulated rice yield using the estimated solar radiation obtained from the deep solar radiation (DSR) model for the external test set (a) within Zone 6 (27.5–43.9° N) and (b) outside the range of Zone 6. RMSE and NRMSE indicate root mean square error and normalized root mean square error, respectively. *** represents significance at $p < 0.001$.

(figure 10(b)). For example, crop yield RMSE was 0.29 t ha^{-1} and 0.76 t ha^{-1} at the sites within and outside Zone 6, respectively. The values of NRMSE were about one-third (4.3%) at the sites within Zone 6 compared with that (13.1%) at sites outside Zone 6.

4. Discussion

Our results demonstrated that a DNN for estimation of solar radiation had greater spatial portability when the training sites were identified using the threshold value of spatial distribution (see supplementary information 5). It was found that spatial portability of the solar radiation models was significantly affected by SDD among training sites. Yet, there was a breakpoint where the impact of SDD on spatial portability of the model was positive. When the DSR model was chosen from multiple sets of DNNs using the threshold value, it had spatial portability within a large area, e.g. the latitude boundary between 27.5° N and 43.9° N in North East Asia. Furthermore, the outputs of the model resulted in a large degree of agreement between crop yield simulations using observed and estimated solar radiation as inputs. These results suggest that it is preferable to identify a reasonable set of training sites using SDD.

It is likely that an optimum spatial distribution range would exist for training sites where spatial portability can be improved for the solar radiation model. Knowledge on this range of SDD can be used to evaluate the suitability of training sites. It was found that a minimum level of spatial dispersion among training sites would be about 190 km in the study region because the m_{RMSE} and CV_{RMSE} tended to remain similar for SDD > 190 km (see supplementary information 6–7). However, the upper limit of the optimum range of SDD was yet to be found due to a small spatial extent where training sites were used for the cross-validation procedures in the present study. This merits further research to examine the impact of SDD values on spatial portability of the DNNs using more weather stations in both Korea and Japan, for example.

Our results indicate that spatial portability of solar radiation models would be affected by the topographic complexity of study regions. The magnitude of irradiation on a surface would be greater at a higher altitude due to a shorter path length through the atmosphere (Blumthaler *et al* 1997). In the present study, such an effect was taken into account using the potential temperatures (Θ) at the reference atmospheric pressure level of 100 kPa. The DFS was another topographic factor that affected spatial portability of the DSR model. For example, the RMSE values of the DSR model tended to decrease as the DFS decreased at sites of interest. The DFS can be added to the input variables of the DNNs once training data measured at the weather stations located in both inland and coastal areas become available.

It is likely that application of DNNs facilitated the representation of the complex relationship between solar radiation and weather variables commonly available from weather stations. Empirical models that use temperature and precipitation as inputs have been reported to have relatively low errors in the region where the parameters of the models were calibrated. For example, Fan *et al* (2018) reported that the temperature and precipitation based model had a

RMSE of about 3.6 MJ m^{-2} d^{-1} in southern China. Hunt *et al* (1998) also reported that the same type of model had a RMSE of about 4.1 MJ m^{-2} d^{-1} in Ontario, Canada. Still, these models resulted in large differences between crop yield simulated using observed and estimated solar radiation as inputs to the crop model in comparison with the DSR model at the weather stations in Japan within Zone 6 (see supplementary information 8). In particular, this outcome was obtained although parameters for those models were calibrated using the same training data for the DSR model. Because the identical weather input data, e.g. temperature and precipitation, were used to develop these models, the differences in spatial portability likely resulted from differences in the model algorithms. Çelik *et al* (2016) and Alsina *et al* (2016) reported that the error statistics for a DNN model differed by the number of nodes. In the present study, the solar radiation models had a relatively large number of nodes, e.g. 64 within each fully connected layer. Such complex structure of neural networks would have allowed for the DSR model to have relatively small errors in estimation of solar radiation.

A crop model has higher sensitivity to changes in solar radiation than other weather variables (Bert *et al* 2007). This indicates the importance of accurate estimation of solar radiation in crop growth simulations. In the present study, the rice yield values simulated using estimated and observed solar radiation as inputs had smaller differences when SDD was larger for the solar radiation models (see supplementary information 9). In particular, the DSR model had a NRMSE of 4% within the latitude range of $\pm 6°$, which extends the suitable areas up to 1200 km from the boundaries of training sites. This result suggests that the DSR model would be a useful tool for researchers to generate weather input files of crop models at local sites where solar radiation data are unavailable.

5. Conclusion

Our results demonstrated that the use of SDD among training sites improved spatial portability of the solar radiation models, which resulted in RMSE < 4 MJ m^{-2} d^{-1} at the majority (73%) of sites in Japan. The DSR model with larger SDD, e.g. >190 km, had accurate estimates of solar radiation within the latitude range of $\pm 6°$ from the areas where training data were obtained. Furthermore, the difference between simulated rice yield values using solar radiation observations versus estimates from the DSR model as inputs to a crop model were small, e.g. 4%, within the same latitude range. These results indicate that the assessment of the spatial distribution of training sites would aid the development of solar radiation models with reasonable spatial portability. Further studies using a weather station network with a wide range of SDD values would be merited to determine the upper limit of the optimum SDD range.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Data/Code availability

All of the material is owned by the authors and/or no permissions are required.

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Funding

This work was carried out with the support of Cooperative Research Program for Agriculture Science&Technology Development (Project No. RS-2022-RD010426), Rural Development Administration, Republic of Korea.

Conflict of interest

I declare that the authors have no competing interests as defined by IOP, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Consent for publication

I confirm that I understand journal Environmental Research Letters is a transformative journal. When research is accepted for publication, there is a choice to publish using either immediate gold open access or the traditional publishing route.

Ethics approval

The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher

Consent for participate

I have read the IOP journal policies on author responsibilities and submit this manuscript in accordance with those policies.

Author contributions

D G K and K S K conceptualized the approach. D G K, Y-U K and K S K wrote the main manuscript. S H,

J K, C-K L, and A M prepared data sets. R M B and D H F revised the manuscript. All authors reviewed the manuscript.

ORCID iDs

Dae Gyoon Kang https://orcid.org/0000-0001-9056-5272
Yean-Uk Kim https://orcid.org/0000-0002-9431-8575
Shinwoo Hyun https://orcid.org/0000-0002-8321-7648
Kwang Soo Kim https://orcid.org/0000-0003-2284-4389
Junhwan Kim https://orcid.org/0000-0003-2155-5294
Chung-Kuen Lee https://orcid.org/0000-0001-8699-4576
Atsushi Maruyama https://orcid.org/0000-0002-5901-9529
Robert M Beresford https://orcid.org/0000-0003-1854-4236
David H Fleisher https://orcid.org/0000-0002-0631-3986

References

Abdelkawy M A, Sabir Z, Guirao J L and Saeed T 2020 Numerical investigations of a new singular second-order nonlinear coupled functional Lane–Emden model *Open Phys.* **18** 770–8

Alsina E F, Bortolini M, Gamberi M and Regattieri A 2016 Artificial neural network optimisation for monthly average daily global solar radiation prediction *Energy Convers. Manage.* **120** 320–9

Angstrom A 1924 Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation *Q. J. R. Meteorol. Soc.* **50** 121–6

Antonopoulos V Z, Papamichail D M, Aschonitis V G and Antonopoulos A V 2019 Solar radiation estimation methods using ANN and empirical models *Comput. Electron. Agric.* **160** 160–7

Aybar-Ruiz A, Jiménez-Fernández S, Cornejo-Bueno L, Casanova-Mateo C, Sanz-Justo J, Salvador-González P and Salcedo-Sanz S 2016 A novel grouping genetic algorithm–extreme learning machine approach for global solar radiation prediction from numerical weather models inputs *Sol. Energy* **132** 129–42

Azevedo J and Crisóstomo S 2016 Weather stations-assisted barometric altimeter for Android: interpolation techniques for improved accuracy *2016 IEEE Sensors Applications Symp. (SAS) (20–22 April 2016)* (IEEE) pp 1–6

Bert F E, Laciana C E, Podestá G P, Satorre E H and Menéndez A N 2007 Sensitivity of CERES-maize simulated yields to uncertainty in soil properties and daily solar radiation *Agric. Syst.* **94** 141–50

Besharat F, Dehghan A A and Faghih A R 2013 Empirical models for estimating global solar radiation: a review and case study *Renew. Sustain. Energy Rev.* **21** 798–821

Black J N 1956 The distribution of solar radiation over the earth's surface *Arch. Meteorol. Geophys. Bioklimatol.* B **7** 165–89

Blumthaler M, Ambach W and Ellinger R 1997 Increase in solar UV radiation with altitude *J. Photochem. Photobiol.* B **39** 130–4

Bock S, Goppold J and Weiß M 2018 An improvement of the convergence proof of the ADAM-optimizer (arXiv:1804.10587)

Bouman B and van Laar H 2006 Description and evaluation of the rice growth model ORYZA2000 under nitrogen-limited conditions *Agric. Syst.* **87** 249–73

Bristow K L and Campbell G S 1984 On the relationship between incoming solar radiation and daily maximum and minimum temperature *Agric. For. Meteorol.* **31** 159–66

Çelik Ö, Teke A and Yıldırım H B 2016 The optimized artificial neural network model with Levenberg–Marquardt algorithm for global solar radiation estimation in Eastern Mediterranean Region of Turkey *J. Clean. Prod.* **116** 1–12

Christodoulakis J, Varotsos C A, Cracknell A P and Kouremadas G A 2018 The deterioration of materials as a result of air pollution as derived from satellite and ground based observations *Atmos. Environ.* **185** 91–99

Fan J, Chen B, Wu L, Zhang F, Lu X and Xiang Y 2018 Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions *Energy* **144** 903–14

Ghimire S, Deo R C, Downs N J and Raj N 2019 Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia *J. Clean. Prod.* **216** 288–310

Hamon R W, Weiss L L and Wilson W T 1954 Insolation as an empirical function of daily sunshine duration *Mon. Wea. Rev.* **82** 141–6

Hu X, An S and Wang J 2014 Exploring urban taxi drivers' activity distribution based on GPS data *Math. Probl. Eng.* **2014** 1–13

Hunt L A, Kuchar L and Swanton C J 1998 Estimation of solar radiation for use in crop modelling *Agric. For. Meteorol.* **91** 293–300

Kaba K, Sarıgül M, Avcı M and Kandırmaz H M 2018 Estimation of daily global solar radiation using deep learning model *Energy* **162** 126–35

Kang D, Hyun S and Kim K S 2019 Development of a deep neural network model to estimate solar radiation using temperature and precipitation *Korean J. Agric. For. Meteorol.* **21** 85–96

King D A, Bachelet D M, Symstad A J, Ferschweiler K and Hobbins M 2015 Estimation of potential evapotranspiration from extraterrestrial radiation, air temperature and humidity to assess future climate change effects on the vegetation of the Northern Great Plains, USA *Ecol. Modell.* **297** 86–97

Lee C K, Kim J and Kim K S 2015 Development and application of a weather data service client for preparation of weather input files to a crop model *Comput. Electron. Agric.* **114** 237–46

Muggeo V M R 2017 segmented: regression models with break-points/change-points estimation (available at: https://CRAN.R-project.org/package=segmented)

Muneer T, Younes S and Munawwar S 2007 Discourses on solar radiation modeling *Renew. Sustain. Energy Rev.* **11** 551–602

Quej V H, Almorox J, Arnaldo J A and Saito L 2017 ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment *J. Atmos. Sol.-Terr. Phys.* **155** 62–70

Reddy S J 1987 Estimation of global solar radiation and evaporation through precipitation-a note *Sol. Energy* **38** 97–104

Renno C, Petito F and Gatto A 2015 Artificial neural network models for predicting the solar radiation as input of a concentrating photovoltaic system *Energy Convers. Manage.* **106** 999–1012

Ryu S, Noh J and Kim H 2017 Deep neural network based demand side short term load forecasting *Energies* **10** 3

Teke A, Yıldırım H B and Çelik Ö 2015 Evaluation and performance comparison of different models for the estimation of solar radiation *Renew. Sustain. Energy Rev.* **50** 1097–107

Thornton P E and Running S W 1999 An improved algorithm for estimating incident daily solar radiation from measurements

of temperature, humidity, and precipitation *Agric. For. Meteorol.* **93** 211–28

Wang L, Kisi O, Zounemat-Kermani M, Salazar G A, Zhu Z and Gong W 2016 Solar radiation prediction using different techniques: model evaluation and comparison *Renew. Sustain. Energy Rev.* **61** 384–97

Yang Y, Xu W, Hou P, Liu G, Liu W, Wang Y and Li S 2019 Improving maize grain yield by matching maize growth and solar radiation *Sci. Rep.* **9** 3635

Zang H, Jiang X, Cheng L, Zhang F, Wei Z and Sun G 2022 Combined empirical and machine learning modeling method for estimation of daily global solar radiation for general meteorological observation stations *Renew. Energ.* **195** 795–808

Zhang Y, Cui N, Feng Y, Gong D and Hu X 2019 Comparison of BP, PSO-BP and statistical models for predicting daily global solar radiation in arid Northwest China *Comput. Electron. Agric.* **164** 104905