Estimation of Spatially Continuous Near-Surface Relative Humidity Over Japan and South Korea

Haemi Park, Junghee Lee, Cheolhee Yoo, Seongmun Sim, and Jungho Im, *Member, IEEE*

***Abstract*—Near-surface relative humidity ($\mathrm{RH_{ns}}$) is an essential meteorological parameter for water, carbon, and climate studies. However, spatially continuous $\mathrm{RH_{ns}}$ estimation is difficult due to the spatial discontinuity of *in situ* observations and the cloud contamination of satellite-based data. This article proposed machine learning-based models to estimate spatially continuous daily $\mathrm{RH_{ns}}$ at 1 km resolution over Japan and South Korea under all sky conditions and examined the spatiotemporal patterns of $\mathrm{RH_{ns}}$. All sky estimation of $\mathrm{RH_{ns}}$ using machine learning has been rarely conducted, and it can be an alternative to the currently available $\mathrm{RH_{ns}}$ data mostly from numerical models, which have relatively low spatial resolution. We combined two schemes for clear sky conditions (scheme A, which uses satellite and reanalysis data) and cloudy sky conditions (scheme B, which uses reanalysis data solely). The relatively small numbers of data in extremely low and high $\mathrm{RH_{ns}}$ conditions (i.e., $<30\%$ or $>70\%$, respectively) were augmented by applying an oversampling method to avoid biased training. The machine learning models based on random forest (RF) and XGBoost were trained and validated using 94 *in situ* observation sites from meteorological administrations of both countries from 2012 to 2017. The results showed that XGBoost produced slightly better performance than RF, and the spatially continuous $\mathrm{RH_{ns}}$ model combined based on XGBoost yielded the coefficient of determination of 0.72 and a root-mean-square error of 10.61%. Spatiotemporal patterns of the estimated $\mathrm{RH_{ns}}$ agreed with *in situ* observations, reflecting the effect of topography on $\mathrm{RH_{ns}}$. We expect that the proposed $\mathrm{RH_{ns}}$ model could be used in various environmental studies that require $\mathrm{RH_{ns}}$ under all sky conditions as input data.**

***Index Terms*—East Asia, extreme gradient boosting, spatially continuous near-surface relative humidity.**

I. Introduction

NEAR-SURFACE relative humidity ($\mathrm{RH_{ns}}$), a ratio of existing humidity to the potential capacity of saturation vapor pressures under various air temperature ($\mathrm{T_a}$) [1], directly influences terrestrial ecosystems. $\mathrm{RH_{ns}}$ has been used as a crucial factor for a range of terrestrial and atmospheric research fields, including climatic modeling [2], meteorological factor estimations [3], aerosol estimation [4], energy balance modeling [5]–[7], drought monitoring [8], [9], vegetation productivity estimations [9], leaf wetness estimations [11], [12], climatic zone classifications [13], [14], and heat index calculations [15]. $\mathrm{RH_{ns}}$ data are generally obtained using radiosonde [16] and automated surface observing systems, which are point-based, and thus spatially discontinuous. Several alternative methods for generating spatially continuous $\mathrm{RH_{ns}}$ include 1) numerical modeling such as reanalysis data [17], 2) spatial interpolation based on *in situ* observations [18], and 3) satellite-based retrieval methods [19], [20].

Reanalysis data provide the spatial distributions and vertical profiles of RH. However, the relatively coarse spatial resolution of the reanalysis data is a major uncertainty for grasping the fine scale variation of $\mathrm{RH_{ns}}$ on the terrestrial ecosystems [21], [22]. In particular, fine resolution meteorological data (e.g., solar radiation, temperature, and humidity) are required in carbon cycle modeling because they work as factors for the water and temperature stress. For an example, Zhao *et al.* [22] found that the biases in the shortwave solar radiation and RH of reanalysis data could affect the model performance of gross and net primary production [22]. Although the biases inherent in estimating $\mathrm{RH_{ns}}$ are well known [23], the spatiotemporal information of $\mathrm{RH_{ns}}$ reanalysis data has been widely used in various environmental studies at regional to global scales, since its global coverage and the gridded data are useful to analyze the target variables qualitatively.

Spatial interpolation of *in situ* observations has been frequently used for spatiotemporal analysis of meteorological data [18], [24]. However, in complex terrain or in highlands, data sparsity is still problematic [25], [26]. The spatial distribution patterns in several datasets are often dissimilar when the datasets are generated using different *in situ* measurements and interpolation algorithms [27].

Satellite-based $\mathrm{RH_{ns}}$ retrieval methods are a tangible alternative to mitigating the limitation of the coarse spatial resolution of reanalysis data and the uncertainties caused by the sparsity of *in situ* stations for spatial interpolation. Satellite-derived environmental variables (e.g., land surface temperature (LST) and precipitable water) were often used to estimate $\mathrm{RH_{ns}}$ [19], [20]. The usefulness of satellite retrieved $\mathrm{RH_{ns}}$, especially using Moderate Resolution Imaging Spectroradiometer (MODIS), has been widely studied in the literature. The MODIS LST can be used for the estimation of saturation vapor pressure and

Manuscript received December 10, 2020; revised March 18, 2021 and May 31, 2021; accepted August 2, 2021. Date of publication August 10, 2021; date of current version September 9, 2021. This work was supported in part by the National Research Foundation of Korea under Grant NRF-2021R1A2C2008561 and in part by the Institute for Information and Communications Technology Promotion supported by the Ministry of Science and ICT, South Korea under Grant IITP-2021-2018-0-01424. *(Haemi Park and Junghee Lee contributed equally to this work.) (Corresponding author: Jungho Im.)*

Haemi Park is with the Earth Observation Research Center, Japan Aerospace Exploration Agency, Tsukuba 305-8505, Japan (e-mail: park.haemi@jaxa.jp).

Junghee Lee, Cheolhee Yoo, Seongmun Sim, and Jungho Im are with the Department of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea (e-mail: olive7861@unist.ac.kr; yoclhe@unist.ac.kr; smsim@unist.ac.kr; ersgis@unist.ac.kr).

Digital Object Identifier 10.1109/JSTARS.2021.3103754

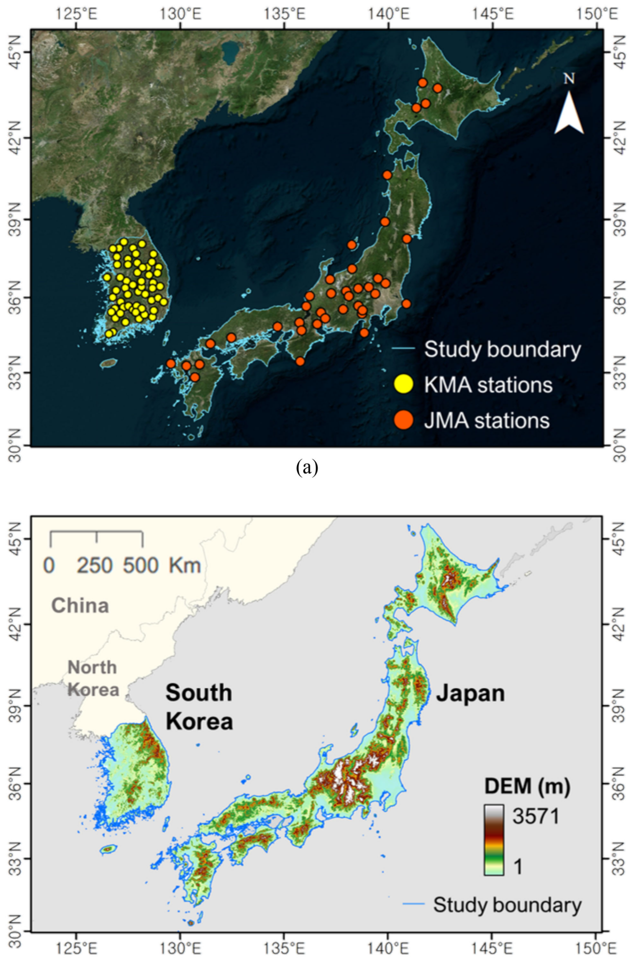
This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

vapor pressure deficit [28]. However, the MODIS LST has a limitation that is not available under cloudy sky conditions. Liao *et al.* [29] proposed a method to estimate RH_{ns} under clouds using Ta and LST, which were obtained under clear skies [29]. They also used a linear regression model to estimate the actual water vapor based on the precipitable water vapor product derived from MODIS. The proposed all-sky RH_{ns} model showed the coefficient of determination (R^2) ranging from 0.28 to 0.5 in the US. There is room for further improvement in estimating all-sky RH_{ns} through incorporating the effect of surface moisture conditions based on empirical modeling with additional input variables, which are related to evapotranspiration, including vegetation indices, surface thermal flux, and soil moisture.

Machine learning approaches have been used to estimate not only RH_{ns} but also its vertical profile [30]–[32]. Shank *et al.* [30] predicted hourly near-surface dew point temperature (T_{dew}) with artificial neural networks under various conditions from freezing to very high temperatures. Li and Zha [31] estimated RH_{ns} using MODIS products and geographic information with random forest (RF) in China during June to September in 2009, resulting in R^2 of 0.7 and root-mean-square error (RMSE) of 7.4%. The monthly product of MODIS enhanced vegetation index was chosen to avoid the influence of cloud cover, and it showed the highest importance among the whole input variables [31]. The studies pointed out that there was a tradeoff between high accuracy for cloud-free areas and completely no information under clouds when using optical sensors [31], [32].

To overcome the spatial discontinuity caused by clouds, approaches for estimation of environmental parameters using remote sensing data under all-sky conditions has been recently proposed for estimating LST [33], soil moisture [34], daily actual evapotranspiration [35], and aerosol optical depth [4]. Microwave sensor images, less sensitive to clouds, are often combined with optical images using machine learning approaches for estimating spatially continuous environmental parameters. For example, passive microwave radiometer data have been used to estimate near-surface humidity in terrestrial and ocean surfaces [30], [36]. In another way, Park *et al.* [4] modeled aerosol optical depth under cloudy sky conditions using spatially continuous data composed of Goddard Earth Observing System Chemical Model (GEOS-chem) AOD and meteorological variables from a numerical model.

In this research, we attempted to estimate spatially continuous RH_{ns} regardless of weather conditions on land surface. The objectives of this article were to 1) propose machine learning-based models to estimate spatially continuous daily RH_{ns} over Japan and South Korea under all sky conditions, 2) examine the temporal patterns of RH_{ns} to verify the universality of the proposed model in the study area where seasonal fluctuations were large, and 3) explore the spatial patterns of RH_{ns} to confirm the spatial transferability of our proposed model, analyzing topographic effects to the model performance.



(a)

(b)

Fig. 1. Study area of this research. (a) Study boundary is enclosed with a blue line. Ground stations in the mainland of Japan and South Korea are plotted with yellow and red dots, respectively. The aerial background images are derived from Bing map. (b) Elevation in the study area using SRTM DEM.

II. Study Area and Materials

A. Study Area

Japan and Republic of Korea (South Korea) were selected as the study area for modeling spatially continuous daily RH_{ns} on land surface, which are frequently covered by clouds, especially in the summer monsoon season. Many recent studies reported that people in the region are exposed to agricultural damages by decreasing crop yields [37], [38], heat waves [39], and air pollutions [40], [41], which are related to RH_{ns}. The mainland of Japan and South Korea are located in the northeast Asia with the spatial extent of 33 °N125 °E–39 °N131 °E, and 30°N128 °E–46°N153 °E (see Fig. 1). Democratic People's Republic of Korea (North Korea) was excluded due to the lack of *in situ* data.

Climate zones of the study area consist of *temperate* and *snow* climates according to the scheme of Köppen-Geiger from 1980 to 2016 [14], [42]. The annual averages of temperature and precipitation in Japan (referred to JMA) and South Korea (referred to KMA) from 1981 to 2010 were 15.21 °C and

TABLE I
DESCRIPTION OF DATASETS FOR SPATIALLY CONTINUOUS RH_{NS} ESTIMATION

	Data	Variable (abbreviation)	Unit	Spatial resolution	Temporal resolution	Source
In-situ	KMA	Relative Humidity (RH_{ns})	%	In-situ	1 min	KMA
	JMA				10 min	JMA
Satellite-based data	MYD04_L2	Aerosol Optical Depth (AOD)	Column-integral	10 km	~1 day	LAADS DAAC
	MYD05_L2	Total Precipitable Water (TPW)	cm	1 km	~1 day	LAADS DAAC
	MYD07_L2	Surface Pressure (SP)	hPa	5 km	~1 day	LAADS DAAC
		Water Vapor (WV)	cm			
	MYD13A2	Normalized Difference Vegetation Index (NDVI)	-	1 km	16 days	LAADS DAAC
	MYD16A2	Evapotranspiration (ET)	kg/m^2/8day	0.5 km	8 days	LAADS DAAC
		Latent Heat Flux (LE_S)	J/m^2/day			
		Potential Evapotranspiration (PET)	kg/m^2/8day			
		Potential Latent Heat Flux (PLE_S)	J/m^2/day			
	GPM 3IMERGHH	Precipitation (P)	mm	0.1 °	30 min	GES DISC
	SRTM	Digital Elevation Model (DEM)	m	~90 m	-	Earth Explorer
Reanalysis data	GLDAS-Noah	Rootzone Soil Moisture (RootSM)	kg·m^{-2}	0.25 °	3 hr	GES DISC
		Soil Moisture—4 layers (SM)	kg·m^{-2}			
		Soil Temperature—4 layers (ST)	K			
	ERA5-Land	Air temperature above 2m (T2m)	K	~9 km	3 hr	ECMWF
		Dew point temperature above 2m (D2m)	K			
		Skin Temperature (T_{skin})	K			
		Leaf Area Index, high vegetation (LAI_{high})	m^2·m^{-2}			
		Leaf Area Index, low vegetation (LAI_{low})	m^2·m^{-2}			
		Surface Latent Heat Flux (LE_R)	J·m^{-2}			
		Surface Sensible Heat Flux (H_R)	J·m^{-2}			
		Surface net Solar Radiation (R_{sr})	J·m^{-2}			
		Surface Solar Radiation downwards ($R_{\text{sr,down}}$)	J·m^{-2}			
		Surface net Thermal Radiation (R_{th})	J·m^{-2}			
		Surface Thermal Radiation downwards ($R_{\text{th,down}}$)	J·m^{-2}			
		10 metre U wind component (U10)	m·s^{-1}			
		10 metre V wind component (V10)	m·s^{-1}			

1610.61 mm, and 12.44 °C and 1347.11 mm, respectively. The average altitudes in Japan and South Korea according to the SRTM digital elevation model (DEM) are 382 m and 255 m, respectively [43] (see Fig. 1).

B. Station-Based RH_{ns} Data

In situ RH_{ns} data measured at 1-2 m above the ground were downloaded from JMA[1] and KMA[2] (see Fig. 1). The RH_{ns} data measured at 13:30 (1:30 P.M. Japan Standard Time (JST); corresponding to the satellite data collection time) of each day in the period from 2012 to 2017 at 39 and 55 stations in Japan and South Korea, respectively, were used for training and validation of our spatially continuous RH_{ns} model. The averaged RH_{ns} during the period were found as 72.54% and 53.55% for Japan and South Korea, respectively.

C. Satellite-Based Data

The atmospheric and land products from Aqua MODIS were used in this article. Aqua MODIS passes at 1:30 P.M. equatorial crossing time, which might be more suitable to estimate dry humidity conditions than the data from Terra (10:30 A.M.). It coincides with the atmospheric mixing ratio, which generally starts to increase from the early morning along with the increasing solar radiation, peaking at noon [44]. The atmospheric variables from the level-2 MODIS atmospheric products (i.e., MYD04_L2, MYD05_L2, and MYD07_L2) with version 6.1—water vapor, total precipitable water, surface pressure, and aerosol optical depth—were used as input variables (see Table I). Aerosol optical depth was considered for RH_{ns} modeling since it can be indirectly related to relative humidity from the hygroscopic growth of aerosols [45]. Land surface characteristics also directly or indirectly affect RH_{ns}. The normalized difference vegetation index, evapotranspiration, and latent heat flux from the MODIS land products (i.e., MYD13A2 and MYD16A2) were used as land input variables (see Table I). The process of evapotranspiration releases moisture into the air from vegetation, resulting in increasing RH_{ns}, and the surface energy balance can directly change surface temperature, affecting relative humidity.

Precipitation and elevation were additionally considered in developing machine learning-based RH_{ns} models. The Global Precipitation Measurement (GPM) Integrated MultisatellitE Retrievals (IMERG) product combines precipitation data retrieved from multisatellite sensors in the GPM constellation. The 30-m GPM precipitation data (3IMERGHH, V06B) were accumulated into the 1-h, 3-h, and 1-day, and denoted as precip_1 h, precip_3 h, and precip_1day (see Table S1). The SRTM 3 arc-second void-filled DEM was used to reflect the inverse relationship between relative humidity and temperature relative to altitude.

D. Reanalysis Data

ERA5-Land reanalysis data, the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis, is the replayed result of a single stand-alone reintegration of the near-surface atmospheric fields from ERA5 atmospheric

[1] [Online]. Available: https://www.jma.go.jp/

[2] [Online]. Available: https://data.kma.go.kr/

reanalysis with precipitation and a lapse-rate adjustment but without direct assimilation of observations [46]. ERA5-Land has an enhanced spatial resolution of about 9 km on land, when compared to the ocean-covered ERA5 of about 32 km. ERA5-Land variables used in this article are presented in Table I. Directly or indirectly related variables with $\mathrm{RH_{ns}}$ were extracted including temperatures, solar radiations, winds, vegetations, and surface energy balance (i.e., latent heat flux and sensible heat flux). Solar radiation has a negative correlation with relative humidity [47], and wind speed is related to a transpiration rate that controls $\mathrm{RH_{ns}}$ around plants [48], [49]. The wind speed was calculated using the 10 m U wind component (U10) and 10 m V wind component (V10).

The soil temperature and soil moisture data were obtained from the Global Land Data Assimilation System (GLDAS) Version 2.1 coupled with the Noah Land Surface Model with the 3-h 0.25° × 0.25° spatial grid. GLDAS provides the root zone soil moisture, soil moisture, and soil temperature at four levels—0–10 cm, 10–40 cm, 40–100 cm, and 100–200 cm.

III. Methodology

A. Schemes for Spatially Continuous RH_{ns}

The spatially continuous $\mathrm{RH_{ns}}$ maps were produced through the combination of two schemes (i.e., schemes A and B) under different conditions. Satellite-derived atmospheric products have often no data in the low-quality pixels containing clouds and aerosols, which results in spatial discontinuity in modeling. In this article, we designed schemes A and B based on whether to use satellite-derived products as input parameters. Scheme A was applied for where the satellite-derived data were available and used both satellite-derived products and reanalysis data, as shown in Table S1. Scheme B used only spatially continuous input parameters, which are reanalysis data (i.e., GLDAS-Noah and ERA5-Land) and SRTM DEM (Table S1). Spatially continuous $\mathrm{RH_{ns}}$ maps were generated by filling gaps in scheme A with scheme B results.

B. Data Preprocessing and Oversampling

In the data preprocessing step, the collected data were bilinearly resampled into a 1 km spatial grid based on the MODIS sinusoidal tile grid. Then, linear temporal interpolation was performed on 8 days or 16 days MODIS land products to calculate daily values. The data for schemes A and B were constructed by extracting the preprocessed input variables (see Table S1) at the ground RH stations.

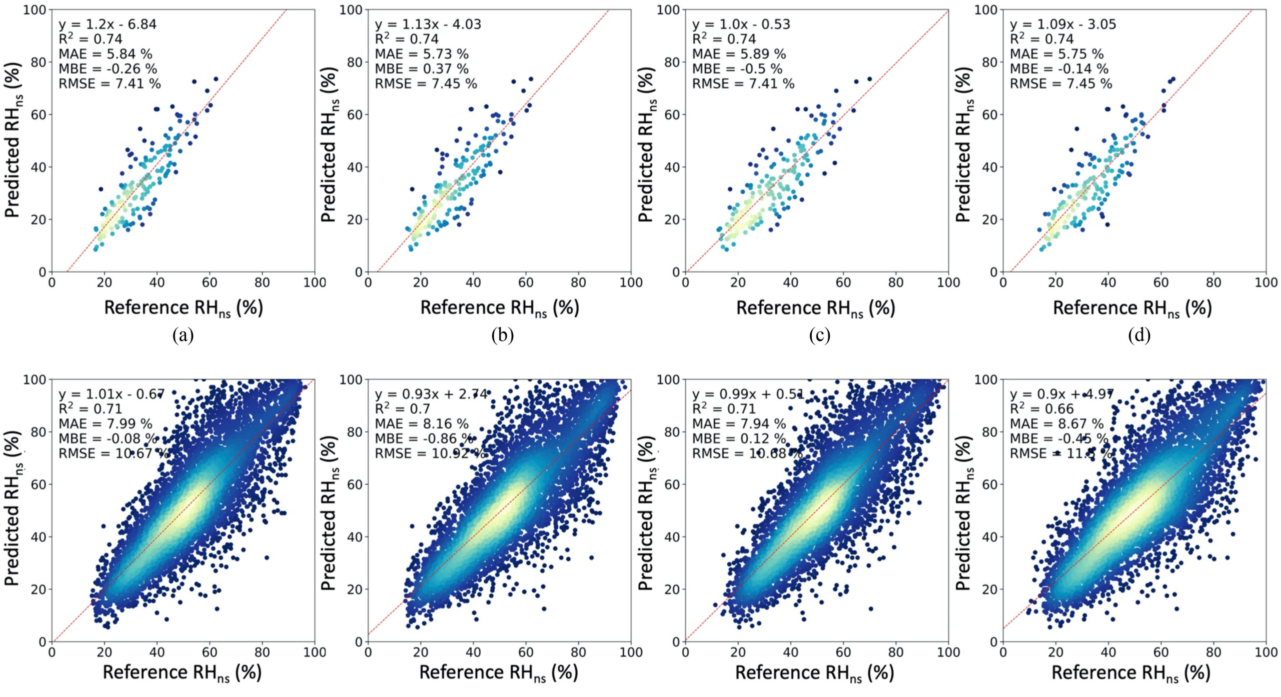
We divided the data into training, validation, and test sets. The test set consists of the data from the tenth day of every month during 2012–2017 for unbiased evaluation of the model. The remaining data were divided into training (80%) and validation (20%) sets through the stratified random sampling approach considering the distribution of the *in situ* $\mathrm{RH_{ns}}$ values.

There are a few numbers of samples for the extreme $\mathrm{RH_{ns}}$ values (very low and high values; Table S2), which could cause the biased training of a model [50]. Thus, oversampling was performed on the training set where $\mathrm{RH_{ns}}$ values were very low and high: 0–30% and 70–100%. Under the assumption that the $\mathrm{RH_{ns}}$ values are not significantly different in the neighboring pixels, the surrounding pixels within the 5 × 5 window from a training sample (i.e., a center pixel) were perturbed within 3% of the centered $\mathrm{RH_{ns}}$ value [51]. Consequently, the RHns data simulated based on the stochastic way in the very low and high ranges were added to the original training set as oversampled data. Some oversampled data were additionally filtered out when the Mahalanobis distances [52], [53] of the input variables were greater than the mean plus one standard deviation for each section in the original training set (i.e., 0–10%, 10–20%, 20–30%, 70–80%, 80–90%, and 90–100%). The original and oversampled training sets for schemes A and B were constructed based on the abovementioned processes.

C. Machine Learning Approaches

In this article, extreme gradient boosting (XGBoost) and RF algorithms were tested for comparison. Extreme gradient boosting (XGBoost) has often resulted in better performance than other machine learning algorithms in various environmental problems [54]–[59]. RF also has been applied to solve various environmental problems in remote sensing fields [60]–[66]. Both RF and XGBoost are based on the classification and regression trees (CART). CART has a well-known problem related to the model instability in which the tree structure changes significantly by small change in training data [67]. RF and XGBoost adopt different approaches to mitigating the weakness of CART. RF [68] uses an ensemble approach based on a multitude of decision trees from the bootstrapped samples (i.e., bagging) with random subsets of training samples and input features at a node. Such bagging helps to control the overfitting of the model by reducing the variance [69], [70]. RF provides information on variable importance by calculating the change of mean square error in percentage using out-of-bag data not used in training [68]. Boosting is another ensemble way to reduce a model bias by sequentially updating weights of multiple weak learners. XGBoost [71] is a relatively recent improvement of gradient boosting, by focusing on the model performance and computational speed, with 1) the scalability and speed by regularization, 2) sparsity awareness, and 3) weighted quantile sketch. XGBoost provides information of variable importance as the number of appearance times an input variable in a tree. The hyperparameters of both machine learning models were optimized with the Bayesian optimization using the *bayes_opt* library in python based on the training and validation sets. The optimized parameters are described in Table S3.

For complement of relative variable importance embedded in each RF and XGBoost, Shapely Additive exPlanations (SHAP) values were used to analyze the contributions of each feature on the model prediction and the interaction of features [72]. The feature contribution of a data point is calculated as the difference between the predicted value on that data point and the average of the repeated predictions [73]. Compared to existing feature importance, SHAP values have the advantages of 1) consistency based on the solid theory (i.e., game theory), 2) contrastive explanations from positive and negative contributions



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2. Model performance of the $\mathrm{RH_{ns}}$ model using the test data for scheme A [i.e., (a)–(d)] and scheme B [i.e., (e)–(f)]. The test data were extracted the tenth day of every month during 2012–2017. The subscripts marked as "ori" and "ovr" indicate that the models used the original and oversampled datasets, respectively. The color scheme from blue to yellow indicates the density of points from low to high. (a) Scheme A, $\mathrm{RF_{ori}}$. (b) Scheme A, $\mathrm{RF_{ovr}}$. (c) Scheme A, $\mathrm{XGBoost_{ori}}$. (d) Scheme A, $\mathrm{XGBoost_{ovr}}$. (e) Scheme B, $\mathrm{RF_{ori}}$. (f) Scheme B, $\mathrm{RF_{ovr}}$. (g) Scheme B, $\mathrm{XGBoost_{ori}}$. (h) Scheme B, $\mathrm{XGBoost_{ovr}}$.

of each variable to the target variable, and 3) understanding the interactions between input variables. The interpretation of SHAP values should be used with care since the contribution of input features is calculated under the given set of input features [74].

D. Assessment of Model Performances

The model performances were evaluated using four accuracy metrics—coefficient of determination ($\mathrm{R^2}$), mean absolute error (MAE), mean bias error (MBE), and RMSE. The relative RMSE (rRMSE, %) was calculated by dividing the RMSE by the mean observed data. The spatially continuous $\mathrm{RH_{ns}}$ map was produced with the best models of each scheme A and B with higher $\mathrm{R^2}$ and lower MAE, MBE, and RMSE. The leave-one-out cross-validations on temporal (i.e., year) and spatial (i.e., ground stations) domains were conducted to examine the stability of the models.

E. Comparison With Reanalysis and Numerical Weather Prediction Data

The $\mathrm{RH_{ns}}$ estimated in this article was compared to the manually calculated $\mathrm{RH_{ns}}$ from ERA5-land products. As a calculation method of $\mathrm{RH_{ns}}$ using $\mathrm{T_a}$ and $\mathrm{T_{dew}}$, (1) was used [75]

$$\mathrm{RHn}s = \left(\frac{112 - 0.1 + T_a + T_{\mathrm{dew}}}{112 + 0.9T_a}\right)^8 \times 100 \qquad (1)$$

where $\mathrm{RH_{ns}}$ is relative humidity, $\mathrm{T_a}$ air temperature above 2 m of ERA5-land, and $\mathrm{T_{dew}}$ dew point temperature of ERA5-land. Both Ta and $\mathrm{T_{dew}}$ are in °C.

The $\mathrm{RH_{ns}}$ of the meso-scale model (MSM) of JMA was additionally compared to the $\mathrm{RH_{ns}}$ estimated in this article [76]. The MSM model has a 5 km spatial resolution and the forecast frequency is 8 times per day (00, 03, 06, 09, 12, 15, and 18 UTC). In this article, the 03 UTC and 25 h forecast data were chosen to refer 13:00 JST (where UTC+09), and the ninth date was used for matching with our test set (every tenth date for each month). The coverage of MSM (22.4 N, 120 E–47.6 N, 150 E) includes the whole area of Japan and South Korea.

IV. Results and Discussion

A. Performance of the Spatially Continuous RH_{ns} Models

The model performances of schemes A and B for training/validation and test datasets are presented in Table S4 and Fig. 2, respectively. While RF and XGBoost with training and validation data showed similar $\mathrm{R^2}$, XGBoost resulted in slightly better performance (i.e., 0–2% smaller MAE, MBE, and RMSE) than RF. The model performance with the test set showed similar results with the validation set. The performance on the test set between RF and XGBoost was also similar (see Fig. 2), but RF showed a less detailed spatial distribution of $\mathrm{RH_{ns}}$ when compared to the XGBoost-derived $\mathrm{RH_{ns}}$ distribution (see Fig. S1). DEM, a spatially static variable, showed the highest rank in the relative variable importance by RF (not shown). Notably,

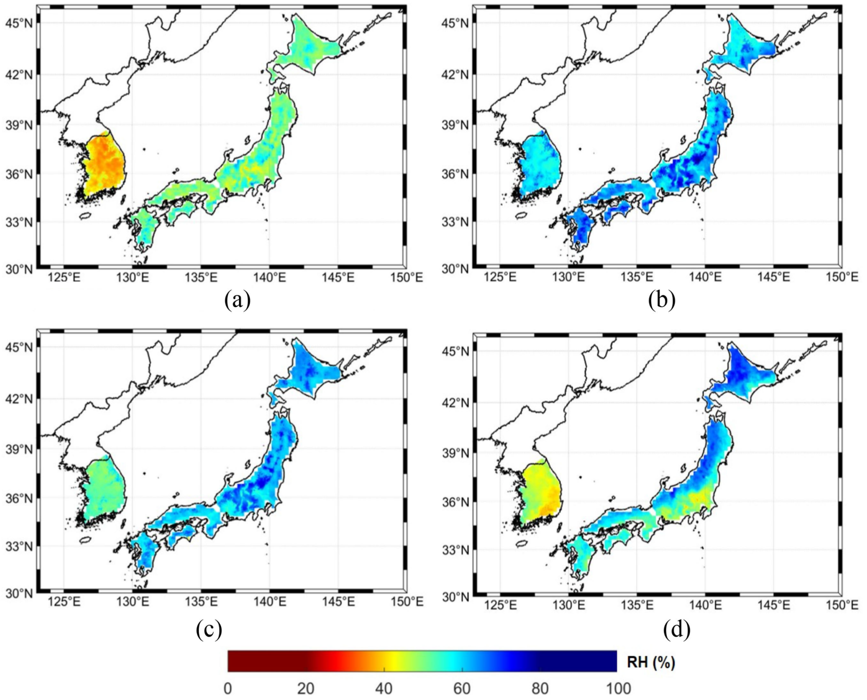


Fig. 3. Spatial distribution of the seasonal average of RH_{ns}. (a) Spring (March to May). (b) Summer (June to August). (c) Fall (September to November). (d) Winter (December to February).

the R^2 between DEM and the daily averaged RH_{ns} of each station was low around 0.09. This could lead to overfitting on the RF model because of the spatial autocorrelation of geographic input variables [77]. Li *et al.* [78] also suggested comparing the spatial pattern between models even though when the model performances were high. Thus, the RH_{ns} model of XGBoost was selected for further analysis.

The oversampling was expected to complement the underrepresented data (i.e., typically for extreme ranges) by augmenting sample size for the ranges. The performance on the test set (see Fig. 2) was similar between the models developed using the original and oversampled sets. Fig. S2 shows the spatial and temporal cross-validation results of the XGBoost model with the original and oversampled sets. The cross-validation results of scheme A were slightly enhanced after oversampling as the slope and R^2 increased towards 1 and the MAE, MBE, and RMSE decreased. Scheme B showed a different tendency when compared to the scheme A. It should be noted that the slope somewhat increased after oversampling. However, the decrease of R^2 and increase of MAE, MBE, and RMSE were found implying that the performance of scheme B was slightly degraded. The oversampling slightly enhanced the model performance where the number of samples in the extreme ranges (i.e., 0–30% and 70–100%) is small enough such as scheme A. On the other hand, the oversampling seems not useful for scheme B, which only consists of reanalysis data with coarse spatial resolution. In scheme B, a central pixel and its corresponding neighboring pixels (within 5 × 5 km) could belong to the same pixel in the data before resampling [79]. Based on the abovementioned results, the final models of schemes A and B were selected: the XGBoost models with the oversampled dataset and original dataset, respectively.

The spatially continuous RH_{ns} maps were generated by combining the final models of schemes A and B (scheme AB afterwards): The predicted RH_{ns} values of scheme B was filled where scheme A results did not exist (see Fig. 3). The performance of scheme AB resulted in R^2, MAE, and RMSE as 0.72, 7.88%, and 10.61%, respectively. When the same test data (i.e., clear sky test samples) were used, scheme A resulted in slightly better estimation of RH_{ns} than scheme B (see Fig. S3). The performance on the test set is comparable with the previous studies considering the research period and location (see Table II). It is notable that the proposed spatially continuous RH_{ns} model showed similar model performance with previous studies that developed under clear sky conditions.

B. *Contribution of the Input Variables on RH_{ns} Models*

The contribution of the input variables for schemes A and B were analyzed with the variable importance of XGBoost (see Fig. 4) and the SHAP values (see Fig. 5). The common contributing input variables for both schemes were solar and thermal radiation, latent heat, and sensible heat, which are closely related to the factors of the surface energy balance, implying the close relationship between the relative humidity and evaporation [80]. The solar and thermal radiation from ERA5 were the most contributing features for scheme B [see Figs. 4(b) and 5(b)]. The summary plot of scheme B [see Fig. 5(b)] shows the negative and positive relationships for R_{th} (or $R_{th,down}$) and R_{sr} (or $R_{sr,down}$), respectively. Both R_{th} and R_{sr} are closely related to the surface

TABLE II
COMPARISON OF THE PERFORMANCE OF THE PROPOSED MODEL WITH THOSE OF THE PREVIOUS STUDIES

	Study area	**Study period**	**R^2**	**MAE (%)**	**RMSE (%)**
Proposed model (scheme AB)	South Korea and Japan	2012-2017	0.72	7.88	10.61
		Summer of 2012-2017	0.67	7.42	9.94
Li and Zha [31]	China	Summer of 2009	0.7	5.6	7.4
Ramírez-Beltrán et al. [20]	Mesoamerica and Caribbean countries (0-30° N, 60-100° W)	2010-2011	0.42-0.62	6.4-7.7	8.5-10.2

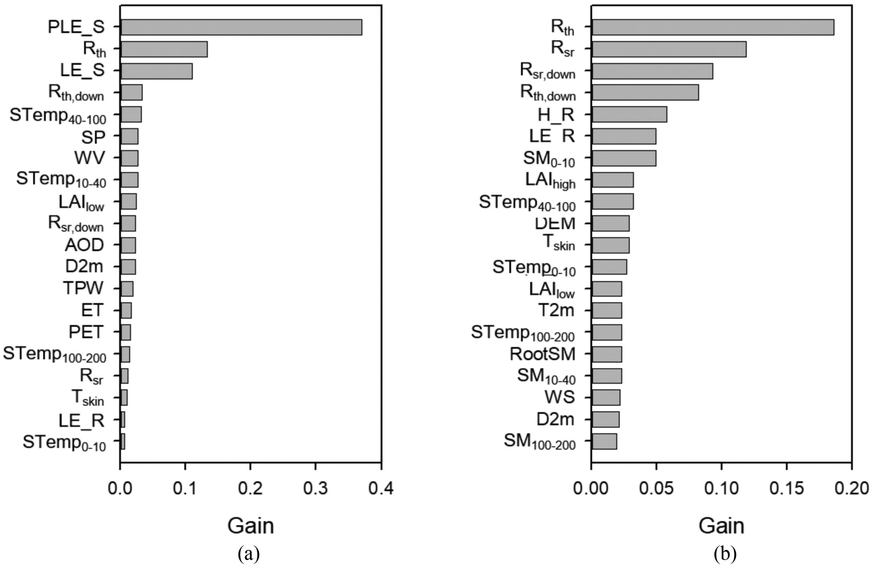


Fig. 4. Top 20 important variables of XGBoost for (a) scheme A with the oversampled dataset and (b) scheme B with the original dataset. Gain indicates the average of the relative contribution of each corresponding feature to the model for each tree. A higher value of gain indicates more relatively important variable for prediction.

temperature, which has the negative relationship with $\mathrm{RH}_{\mathrm{ns}}$. The negative relationship between R_{th} and $\mathrm{RH}_{\mathrm{ns}}$ comes from the definition of R_{th}, which is the difference between downward and upward longwave radiation at the surface of the Earth.

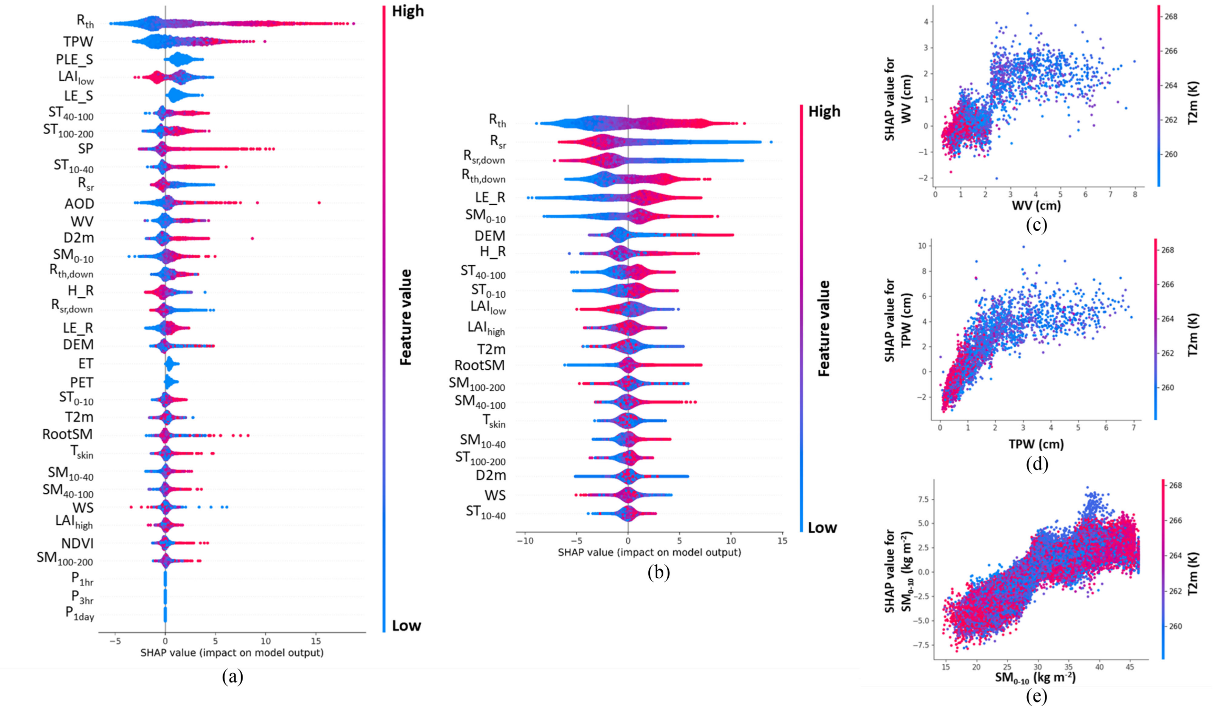
The satellite-derived latent heat variables, PLE_S and LE_S, were more contributing to the model than other radiation-related variables, except for R_{th}, for scheme A. The different relative contribution of the latent heat flux variables between schemes A and B may come from the varied performance of satellite-derived and reanalysis data. The satellite-derived latent heat flux product had the correlation coefficient (R) of 0.7 with the ground reference data [81], while the reanalysis-derived latent heat flux product had 0.4 [81]. The vegetation indices (e.g., NDVI and enhanced vegetation index), DEM, land surface temperature, and precipitable water were reported as the most relevant parameters to the $\mathrm{RH}_{\mathrm{ns}}$ in several previous studies [20], [32], [83], [84]. However, the surface energy balance-related variables were relatively more contributing to the models in this article.

In addition, the interaction between the humidity-related variables (i.e., WV, TPW, and $\mathrm{SM}_{0\text{-}10}$) and $\mathrm{T}_{2\mathrm{m}}$ to predict $\mathrm{RH}_{\mathrm{ns}}$ was further analyzed for scheme A [see Figs. 5(c) and (d)] and scheme B [see Fig. 5(e)]. WV and TPW have negative relationships with $\mathrm{T}_{2\mathrm{m}}$: the larger WV and TPW, the lower $\mathrm{T}_{2\mathrm{m}}$. The smaller WV and TPW (i.e., higher $\mathrm{T}_{2\mathrm{m}}$) tend to lower $\mathrm{RH}_{\mathrm{ns}}$. $\mathrm{SM}_{0\text{-}10}$ showed a linear relationship with SHAP values, but there was no clear trend for $\mathrm{T}_{2\mathrm{m}}$.

C. Temporal Pattern Analysis of Near-Surface Relative Humidity

Fig. 6 depicts the leave-one-year-out cross validation results for schemes A and B. The variation of the year-by-year RMSEs for 2012–2017 was not large, ~8.16–9.35% and 10.92–13.57% for schemes A and B, respectively, implying that the proposed model is temporally stable and transferable. Fig. 7 shows the national average of $\mathrm{RH}_{\mathrm{ns}}$ on the tenth day of each month during 2012–2017 (i.e., test set). The $\mathrm{RH}_{\mathrm{ns}}$ predicted from the final models of schemes A and AB were compared with the ground reference data. The surface relative humidity data of NCEP-NCAR reanalysis 1 with 6 h-interval [17][3] was additionally

[3] [Online]. Available: https://psl.noaa.gov/



(a)

(b)

(c)

(d)

(e)

Fig. 5. Summary plots for scheme A (a) and scheme B (b) of SHAP values. The horizontal axis indicates the impact of a feature value for prediction. The color and thickness of summary plots show the direction of each feature value from low to high and sample size, respectively. Features with larger contribution are placed in order from the top. The dependence plots of SHAP values between humidity-related features and air temperature are plotted on (c)–(e). The humidity-related values (*X*-axis) were selected differently on scheme A ((c) WV and (d) TPW) and scheme B ((e) SM_{0-10}) considering the contribution of input features for each scheme. Each *y*-axis of (c)–(e) shows the SHAP values for each feature. The color means the feature value of T2m.

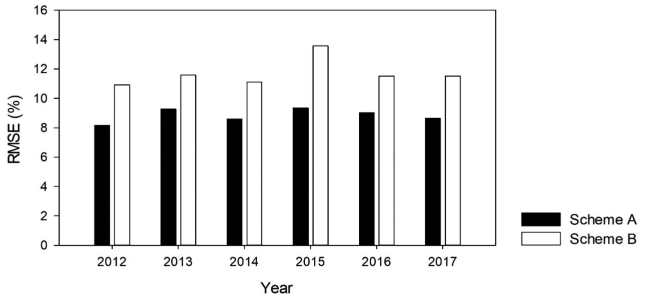


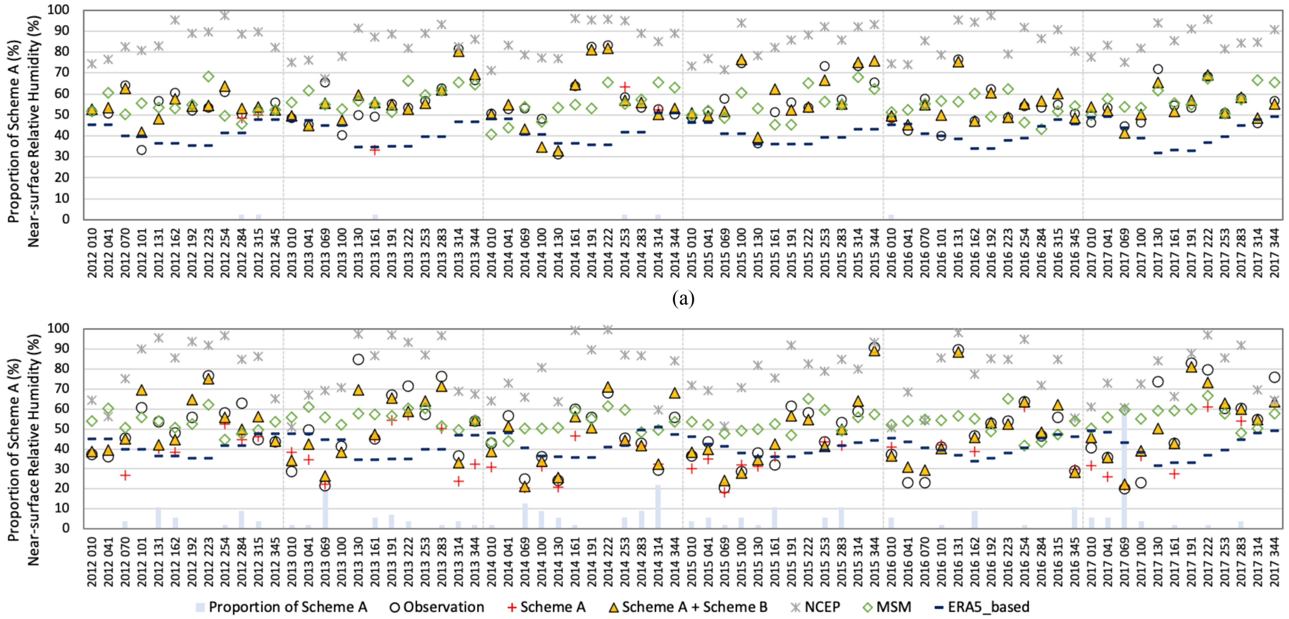
Fig. 6. Leave-one-year-out cross-validation results for schemes A and B.

compared after the linear temporal interpolation was applied. The scheme AB model well reflected the temporal pattern of the ground observations compared to the others. Since scheme A provided results over limited areas (generally less than 20% in most days), it is hard to represent the ground reference of RH_{ns}. While the NCEP-NCAR reanalysis data seemed to follow the temporal patterns of the observed $RH_{ns,}$ they were overall higher than ground reference data. This article also calculated ERA5-based RH_{ns} using T_a and T_{dew} (1), which is shown in Fig. 7. The RMSE of the ERA5-based RH_{ns} was 24.15%, greater than the error of the model proposed in this article. The coarse spatial resolution (~9 km) of ERA5-based data could be a reason to degrade the accuracy. Moreover, the interannual fluctuations of the ERA5-based RH_{ns} are not corresponding to *in situ* observations. The MSM-based RH_{ns} showed higher accuracy (i.e., RMSE of 15.47%) than the other two numerical model-based results (i.e., NCEP-NCAR and ERA5). It should be noted that the RH_{ns} estimated using the proposed approach has the highest accuracy (10.61%) when compared to the three-numerical model-based RH_{ns} (see Fig. 7). The more specific spatial modeling with a spatial resolution at ~5 km of MSM would derive relatively better estimations than ERA5 and NCEP. However, it is confirmed that the proposed model in this article showed higher accuracy than the other data sources.

D. Spatial Pattern Analysis of Near-Surface Relative Humidity

The spatial distribution of seasonal RHns is depicted in Fig. 3. Both Japan and South Korea showed the highest RH_{ns} in summer followed by autumn, winter, and spring. The spatial patterns of RH_{ns} were also found in accordance with topographic characteristics such as the gradual difference from plains to mountainous areas (see Fig. 3). In common, areas with higher elevation tend to have smaller RH_{ns} values than those with lower elevation [85]. Duane *et al.* [86] also reported higher seasonal and diurnal variability of RH_{ns} as elevation increases. The land use [87] or clouds [88] could influence on the regional spatial patterns of RH_{ns}.

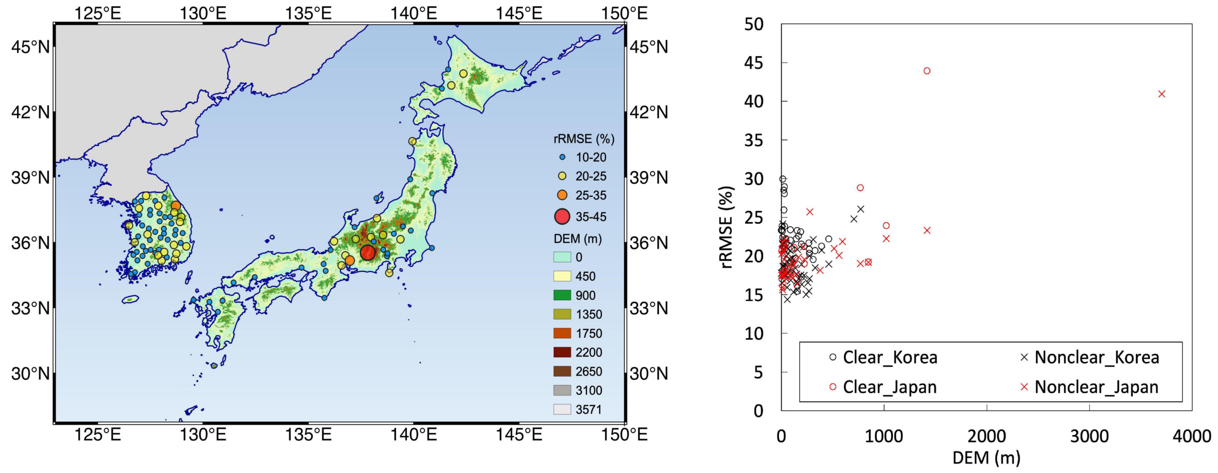
The model performance of the leave-one-station out validation was relatively higher than that of the leave-one-year out cross validation (see Fig. S2). Some stations in the leave-one-station out validation were abnormally less accurate, showing the higher maximum value of RMSE, MAE, and MBE, than the results of leave-one-year out validation [see Fig. S2 (a)]. It means



(a)

(b)

Fig. 7. National average of RH_{ns} in (a) Japan and (b) South Korea. The data of the tenth day of each month during 2012–2017 were compared. The black circles indicate *in situ* observations. The red cruciform and yellow triangle markers indicate the RH_{ns} estimated from schemes A and AB, respectively. The gray stellate markers show the NCEP-NCAR reanalysis data. The green diamond symbol describes MSM. The dark blue dash is the result of ERA5-based calculation using dew point temperature and air temperature. The blue-sky bar graphs show the proportion of the stations that had output for scheme A.



(a)

(b)

Fig. 8. Relationship between RMSE and DEM of all *in situ* observation stations (94 sites). Relative RMSE (rRMSE) of each station (a) on the SRTM DEM, and (b) with scatter plot with DEM for both scheme A (clear) and B (nonclear sky), and both countries.

the model proposed in this article might be better at training the characteristics of *in situ* RH_{ns} observations according to the geography (with 94 stations) rather than their annual variations (with six years). It also implies if more temporal observation data were accumulated, the model performance might improve.

In particular, the station sparsity such as few stations in a mountainous region often greatly increase the uncertainty of the RH_{ns} distribution [89]. The mountainous areas can cause two kinds of uncertainties: 1) station representativeness regarding topographic relief [90] and 2) low density of ground-level stations in high elevation [91].

The impact of data sparsity on the proposed RH_{ns} model was analyzed with the relationship between elevation and rRMSE (see Fig. 8). Although some stations at low elevation showed large variation between near to and far from coastal areas [see Fig. 8(a)], most stations resulted in the tendency of higher rRMSE with increasing elevation [see Fig. 8(b)]. The elevation of ground stations measuring RH_{ns} varied from 1 to 3702 m in our study area (the average was 211.18 m among 94 stations). In terms of the elevation difference by country, the average elevation of the 39 sites in Japan (308.41 m) showed higher than that of 55 sites in South Korea (142.24 m). In Fig. 8(b), the

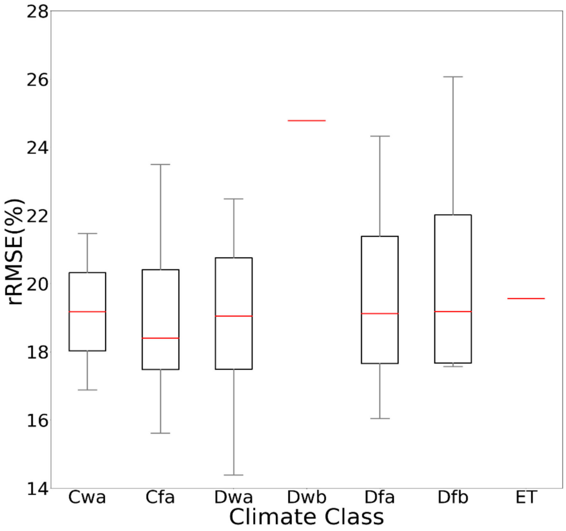


Fig. 9. Boxplots of RMSEs between estimated and *in situ* RH_{ns} with each climate zone: Cwa (temperate, dry winter, and hot summer), Cfa (temperate, no dry season, and hot summer), Dwa (cold, dry winter, and hot summer), Dwb (cold, dry winter, and warm summer), Dfa (cold, no dry season, and hot summer), Dfb (cold, no dry season, and warm summer), and ET (polar, and tundra). The line in the middle of each box shows the median RMSE of the data in the zone. Error ticks around the box represent the maximum and minimum. Boxes describe the first and third quartiles.

stations in Japan, which are located in relatively high altitudes (>1000 m), showed large rRMSE. It might mean that RH_{ns} data at high elevation in Japan were not well learned by the model, whereas the most observations of both countries in low elevation were trained well resulting in relatively high accuracy. Gervais *et al.* [91] and Herrera *et al.* [92] showed consistent results that the density of ground stations has a significant effect on the model performance. This article also implies that RH_{ns} in high elevation might not be trained well due to the lack of *in situ* stations in high altitudes. Additionally, we tested the dependency of DEM for both schemes *with/without* DEM as an input variable in the XGBoost model. The results showed that scheme A produced RMSE of 7.45% *with* DEM, and 7.93% *without* DEM. Scheme B also showed RMSE of 10.68% *with* DEM, and 10.79% *without* DEM. In all cases, the slight decrease in the model accuracy was found when DEM was excluded. The relationship between elevation and RH_{ns} should be further examined in the future by adding more observation data in high altitude regions.

Climate zones or climate factors are often used as a proxy for evaluating the reliability of a meteorological humidity model based on *in situ* observations [93]. To discuss the reliability of the proposed RH_{ns} model over varied climate classes, the boxplots of rRMSE were examined by climate zone (see Fig. 9). The climate zones of the study area, according to the scheme of Köppen-Geiger [14], [42], consist of Cwa (temperate, dry winter, hot summer, 2 sties), Cfa (temperate, no dry season, hot summer, 28 sites), Dwa (cold, dry winter, hot summer, 39 sites), Dwb (cold, dry winter, warm summer, 1 site), Dfa (cold, no dry season, hot summer, 19 sites), Dfb (cold, no dry season, warm summer, 4 sties), and ET (polar, tundra, 1 site). Overall, the temperate classes (Cwa and Cfa) showed high accuracy in terms of rRMSE within 18 to 20% and the standard deviation was also relatively small. In contrast, cold climates (Dwb, Dfa, Dfb, and ET, excluding Dwa), showed more variations in rRMSE when using daily RH_{ns}. The relatively low accuracy in cold climate zones, in this article, area might be due to snow. Snow plays roles to make soil moisture decrease by delaying water discharge due to the time lag of melting [94], [95] and to change an evaporation rate [96], [97]. Furthermore, snow on canopy and subcanopy contributes to stimulating more evaporation since the sublimation of snow is more active with the difference between snow and air temperature [96], [97]. In Suzuki and Nakai [96], the sublimation of intercepted snow was accounted to be 26% of total precipitation at a coniferous forest site in northern Japan. Accordingly, relatively low model performances with rRMSE in cold climate regions in this article might be related to the snow effect.

The spatial distribution of RH_{ns} stations is not uniform by land cover type: about half of the stations (i.e., 57 stations) were located on nonvegetated areas (e.g., urban, cropland/natural vegetation mosaics, and barren) according to the International Geosphere-Biosphere Programme land cover classes of MCD12Q1 in 2017). The rRMSE of each station was plotted according to the land cover types (see Fig. S4). The rRMSE values on both vegetated and nonvegetated stations were similar to each other. This indicates that our model has consistent performance irrespective of the land cover types, although this article focused on vegetation presence using NDVI, ET, and LAI as predictors. However, RH_{ns} is also known to be related with aerosols and heat island phenomena [98], [99]. In order to further improve the proposed model, the representativeness of *in situ* observations by environmental parameters such as land cover, aerosols, and urban/rural landscapes should be carefully considered.

V. Conclusion

In this article, the spatially continuous RH_{ns} was modeled using machine learning approaches over Japan and South Korea, a part of northeast Asia. Under cloud free conditions, satellite-based data and reanalysis data were synergistically used for modeling RH_{ns}, named scheme A. To estimate the RH_{ns} under the cloudy sky conditions, spatially continuous reanalysis data was solely used (scheme B). Combining scheme A and B (i.e., scheme AB) enabled to generate the map of all-sky RH_{ns}. RF and XGBoost machine learning approaches were used for both schemes. The results showed that XGBoost yielded slightly higher accuracy than RF with the more appropriate spatial distribution of RH_{ns}. Consequently, the combined scheme AB based on XGBoost method for all sky conditions produced comparable or even better performance with the literature, resulting in an R^2 of 0.72, MAE of 7.88%, and RMSE of 10.61% using the test data. The spatiotemporal patterns of the RH_{ns} predicted using scheme AB agreed with the *in situ* observations in both Japan and South Korea. However, some stations that were mostly located in data scarcity areas still showed relatively low accuracy in the leave-one-station out cross validation. A few *in situ* stations located at high altitudes were revealed as a possible uncertainty of $\mathrm{RH}_{\mathrm{ns}}$ distribution, showing rRMSE over 30%. The performance of the proposed approach could be further improved by modifying the input meteorological and climatic factors under different conditions or reducing uncertainties associated with input data. Besides, the spatial continuity between the results of two schemes should be tested carefully in the future. Future research should also focus on the improvement of all-sky $\mathrm{RH}_{\mathrm{ns}}$ estimation over the globe including ocean (which was not consider in this article) through multisensor (i.e., optical and microwave) data fusion. The spatially continuous $\mathrm{RH}_{\mathrm{ns}}$ on land surface based on the proposed approach can be used for various environmental studies, which use $\mathrm{RH}_{\mathrm{ns}}$ as input data

REFERENCES

[1] M. Ek and L. Mahrt, “Daytime evolution of relative humidity at the boundary layer top,” *Monthly Weather Rev.*, vol. 122, no. 12, pp. 2709–2721, 1994.

[2] D. Jacob, “The role of water vapour in the atmosphere. A short overview from a climate modeller’s point of view,” *Phys. Chem. Earth, a Solid Earth Geod.*, vol. 26, no. 6–8, pp. 523–527, 2001.

[3] K. Ishida, M. L. Kavvas, S. Jang, Z. Q. Chen, N. Ohara, and M. L. Anderson, “Physically based estimation of maximum precipitation over three watersheds in Northern California: Relative humidity maximization method,” *J. Hydrol. Eng.*, vol. 20, no. 4, 04015014, 2015.

[4] S. Park *et al.*, “Estimation of spatially continuous daytime particulate matter concentrations under all sky conditions through the synergistic use of satellite-based AOD and numerical models,” *Sci. Total Environ.*, vol. 713, 136516, 2020.

[5] M. Valipour, M. A. G. Sefidkouhi, and M. Raeini, “Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events,” *Agricultural Water Manag.*, vol. 180, pp. 50–60, 2017.

[6] H. Yan and H. H. Shugart, “An air relative-humidity-based evapotranspiration model from eddy covariance data,” *J. Geophys. Res. Atmos.*, vol. 115, no. D16, D16106, 2010.

[7] M. Liu, R. Tang, Z. L. Li, Y. Yao, and G. Yan, “Global land surface evapotranspiration estimation from meteorological and satellite data using the support vector machine and semiempirical algorithm,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 513–521, Feb. 2018.

[8] S. L. Goodrick, “Modification of the fosberg fire weather index to include drought,” *Int. J. Wildland Fire*, vol. 11, no. 4, pp. 205–211, 2002.

[9] A. Farahmand and A. AghaKouchak, “A generalized framework for deriving nonparametric standardized drought indicators,” *Adv. Water Resour.*, vol. 76, pp. 140–145, 2015.

[10] M. A. White, P. E. Thornton, S. W. Running, and R. R. Nemani, “Parameterization and sensitivity analysis of the BIOME–BGC terrestrial ecosystem model: Net primary production controls,” *Earth Interact,* vol. 4, no. 3, pp. 1–85, 2000.

[11] K. S. Kim, S. E. Taylor, M. L. Gleason, R. Villalobos, and L. F. Arauz, “Estimation of leaf wetness duration using empirical models in Northwestern Costa Rica,” *Agricultural Forest Meteorol.*, vol. 129, no. 1/2, pp. 53–67, 2005.

[12] P. C. Sentelhas, A. Dalla Marta, S. Orlandini, E. A. Santos, T. J. Gillespie, and M. L. Gleason, “Suitability of relative humidity as an estimator of leaf wetness duration,” *Agricultural Forest Meteorol.*, vol. 148, no. 3, pp. 392–400, 2008.

[13] S. Park *et al.*, “Delineation of high resolution climate regions over the Korean peninsula using machine learning approaches,” *PLoS One*, vol. 14, no. 10, e0223362, 2019.

[14] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, “Present and future Köppen-Geiger climate classification maps at 1-km resolution,” *Sci. Data*, vol. 5, 180214, 2018.

[15] P. J. Robinson, “On the definition of a heat wave,” *J. Appl. Meteorol.*, vol. 40, no. 4, pp. 762–775, 2001.

[16] W. P. Elliott and D. J. Gaffen, “On the utility of radiosonde humidity archives for climate studies,” *Bull. Amer. Meteorol. Soc.*, vol. 72, no. 10, pp. 1507–1520k, 1991.

[17] E. Kalnay *et al.*, “The NCEP/NCAR 40-year reanalysis project,” *Bull. Amer. Meteorol. Soc.*, vol. 77, no. 3, pp. 437–472, 1996.

[18] I. Harris, T. J. Osborn, P. Jones, and D. Lister, “Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset,” *Sci. Data*, vol. 7, no. 1, pp. 1–18, 2020.

[19] G. Peng, J. Li, Y. Chen, A. P. Norizan, and L. Tay, “High-resolution surface relative humidity computation using MODIS image in peninsular Malaysia,” *Chin. Geogr. Sci.*, vol. 16, no. 3, pp. 260–264, 2006.

[20] N. D. Ramírez-Beltrán, C. M. Salazar, J. M. Castro Sánchez, and J. E. González, “A satellite algorithm for estimating relative humidity, based on GOES and MODIS satellite data,” *Int. J. Remote Sens.*, vol. 40, no. 24, pp. 9237–9259, 2019.

[21] B. D. Santer *et al.*, “Uncertainties in observationally based estimates of temperature change in the free atmosphere,” *J. Geophys. Res. Atmos.*, vol. 104, no. D6, pp. 6305–6333, 1999.

[22] M. Zhao, S. W. Running, and R. R. Nemani, “Sensitivity of moderate resolution imaging spectroradiometer (MODIS) terrestrial primary production to the accuracy of meteorological reanalyses,” *J. Geophys. Res. Biogeosciences*, vol. 111, no. G1, G01002, 2006.

[23] A. A. Berg, J. S. Famiglietti, J. P. Walker, and P. R. Houser, “Impact of bias correction to reanalysis products on simulations of North American soil moisture and hydrological fluxes,” *J. Geophys. Res. Atmos.*, vol. 108, no. D16, 4490, 2003.

[24] A. Yatagai, K. Kamiguchi, O. Arakawa, A. Hamada, N. Yasutomi, and A. Kitoh, “Aphrodite constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges,” *Bull. Amer. Meteorol. Soc.*, vol. 93, no. 9, pp. 1401–1415, 2012.

[25] E. Palazzi, J. Von Hardenberg, and A. Provenzale, “Precipitation in the Hindu-Kush Karakoram Himalaya: Observations and future scenarios,” *J. Geophys. Res. Atmos.*, vol. 118, no. 1, pp. 85–100, 2013.

[26] N. Kanda, H. S. Negi, M. Rishi, and A. Kumar, “Performance of various gridded temperature and precipitation datasets over northwest himalayan region,” *Environ. Res. Commun.*, vol. 2, 085002, 2020.

[27] A. J. Newman, M. P. Clark, R. J. Longman, E. Gilleland, T. W. Giambelluca, and J. R. Arnold, “Use of daily station observations to produce high-resolution gridded probabilistic precipitation and temperature time series for the Hawaiian islands,” *J. Hydrometeorol.*, vol. 20, no. 3, pp. 509–529, 2019.

[28] H. Hashimoto *et al.*, “Satellite-based estimation of surface vapor pressure deficits using MODIS land surface temperature data,” *Remote Sens. Environ.*, vol. 112, no. 1, pp. 142–155, 2008.

[29] Q. - Y. Liao *et al.*, “A Method for Deriving Relative Humidity From MODIS Data Under All-Sky Conditions,” *IEEE Trans. Geosci. Remote Sens.*, pp. 1–15, 2020, doi: 10.1109/TGRS.2020.3036248.

[30] D. B. Shank, G. Hoogenboom, and R. W. McClendon, “Dewpoint temperature prediction using artificial neural networks,” *J. Appl. Meteorol. Climatol.*, vol. 47, no. 6, pp. 1757–1769, 2008.

[31] L. Li and Y. Zha, “Mapping relative humidity, average and extreme temperature in hot summer over china,” *Sci. Total Environ.*, vol. 615, pp. 875–881, 2018.

[32] C. R. Cabrera-Mercader and D. H. Staelin, “Passive microwave relative humidity retrievals using feedforward neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 33, no. 6, pp. 1324–1328, Nov. 1995.

[33] X. Zhang, J. Zhou, F. M. Göttsche, W. Zhan, S. Liu, and R. Cao, “A method based on temporal component decomposition for estimating 1-km all-weather land surface temperature by merging satellite thermal infrared and passive microwave observations,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4670–4691, Jul. 2019.

[34] S. Kim, J. Jeong, M. Zohaib, and M. Choi, “Spatial disaggregation of ASCAT soil moisture under all sky condition using support vector machine,” *Stoch. Environ. Res. Risk Assess.*, vol. 32, no. 12, pp. 3455–3473, 2018.

[35] H. R. Shwetha and D. N. Kumar, “Estimation of daily actual evapotranspiration using vegetation coefficient method for clear and cloudy sky conditions,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2385–2395, May 2020.

[36] J. B. Roberts, C. A. Clayson, and F. R. Robertson, “Improving near-surface retrievals of surface humidity over the global open oceans from passive microwave observations,” *Earth Space Sci.*, vol. 6, no. 7, pp. 1220–1233, 2019.

[37] T. Okada, “Integrated water resources management and drought risk management in japan,” *Water Policy*, vol. 18, no. S2, pp. 70–88, 2016.

[38] I. Hong, J.-H. Lee, and H.-S. Cho, “National drought management framework for drought preparedness in Korea (lessons from the 2014–2015 drought),” *Water Policy*, vol. 18, no. S2, pp. 89–106, 2016.

[39] J.-Y. Chung, Y. Honda, Y.-C. Hong, X.-C. Pan, Y.-L. Guo, and H. Kim, "Ambient temperature and mortality: An international study in four capital cities of east Asia," *Sci. Total Environ.*, vol. 408, no. 2, pp. 390–396, 2009.

[40] H. Fukushima, "Air pollution monitoring in east Asia-Japan's role as an environmentally advanced Asian country," *NISTEP Sci. Technol. Foresight Center*, 2006.

[41] J. Kim, S.-C. Yoon, A. Jefferson, and S.-W. Kim, "Aerosol hygroscopic properties during Asian dust, pollution, and biomass burning episodes at Gosan, Korea in April 2001," *Atmos. Environ.*, vol. 40, no. 8, pp. 1550–1560, 2006.

[42] M. C. Peel, B. L. Finlayson, and T. A. McMahon, "Updated world map of the Köppen-Geiger climate classification," *Hydrol. Earth Syst. Sci.*, vol. 11, no. 5, pp. 1633–1644, 2007.

[43] T. G. Farr *et al.*, "The shuttle radar topography mission," *Rev. Geophys.*, vol. 45, no. 2, RG2004, 2007.

[44] R. B. Stull, *An Introduction to Boundary Layer Meteorology*. Berlin, Germany: Springer, 2012.

[45] Y. Sun *et al.*, "The impact of relative humidity on aerosol composition and evolution processes during wintertime in Beijing, China," *Atmos. Environ.*, vol. 77, pp. 927–934, 2013.

[46] E. Zsótér *et al.*, "Trends in the glofas-ERA5 river discharge reanalysis," *ECMWF Tech. Memo.*, vol. 871, 2020.

[47] K. Ruosteenoja and P. Räisänen, "Seasonal changes in solar radiation and relative humidity in Europe in response to global warming," *J. Clim.*, vol. 26, no. 8, pp. 2467–2481, 2013.

[48] S. J. Schymanski and D. Or, "Wind increases leaf water use efficiency," *Plant Cell Environ.*, vol. 39, no. 7, pp. 1448–1459, 2016.

[49] S. Ravi and P. D'Odorico, "A field-scale analysis of the dependence of wind erosion threshold velocity on air humidity," *Geophys. Res. Lett.*, vol. 32, no. 21, L21404, 2005.

[50] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, and N. Japkowicz, "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance," in *Proc. IEEE Int. Conf. Data Mining*, 2018, pp. 447–456.

[51] S. Park *et al.*, "Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over south Korea," *Atmos. Chem. Phys.*, vol. 19, pp. 1097–1113, 2019.

[52] P. Filzmoser, R. G. Garrett, and C. Reimann, "Multivariate outlier detection in exploration geochemistry," *Comput. Geosci.*, vol. 31, no. 5, pp. 579–587, 2005.

[53] J. R. Berrendero, B. Bueno-Larraz, and A. Cuevas, "On mahalanobis distance in functional settings," *J. Mach. Learn. Res.*, vol. 21, no. 9, pp. 1–33, 2020.

[54] A. M. Abdi, "Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data," *GIScience Remote Sens,* vol. 57, no. 1, pp. 1–20, 2020.

[55] A. C. Just, M. M. De Carli, A. Shtein, M. Dorman, A. Lyapustin, and I. Kloog, "Correcting measurement error in satellite aerosol optical depth with machine learning for modeling PM2.5 in the northeastern USA," *Remote Sens,* vol. 10, no. 5, 803, 2018.

[56] S. Georganos *et al.*, "Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application," *GIScience Remote Sens.*, vol. 55, no. 2, pp. 221–242, 2018.

[57] Y. Li, C. Li, M. Li, and Z. Liu, "Influence of variable selection and forest type on forest aboveground biomass estimation using machine learning algorithms," *Forests*, vol. 10, no. 12, 1073, 2019.

[58] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "PM2. 5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data," *Atmos. (Basel)*, vol. 10, no. 7, 373, 2019.

[59] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sens. Environ.*, vol. 221, pp. 430–443, 2019.

[60] S. Araki, M. Shima, and K. Yamamoto, "Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan," *Sci. Total Environ.*, vol. 634, pp. 1269–1277, 2018.

[61] M. Gumma *et al.*, "Agricultural cropland extent and areas of south Asia derived using landsat satellite 30m time series big data using random forest machine learning algorithms on the Google Earth Engine Cloud," *GIScience Remote Sens,* vol. 57, no. 3, pp. 302–322, 2020.

[62] K. McLaren, K. McIntyre, and K. Prospere, "Using the random forest algorithm to integrate hydroacoustic data with satellite images to improve the mapping of shallow nearshore benthic features in a marine protected area in Jamaica," *GIScience Remote Sens,* vol. 56, no. 7, pp. 1065–1092, 2019.

[63] G. Mutowo, O. Mutanga, and M. Masocha, "Including shaded leaves in a sample affects the accuracy of remotely estimating foliar nitrogen," *GIScience Remote Sens.*, vol. 56, no. 7, pp. 1114–1127, 2019.

[64] C. Sothe *et al.*, "Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data," *GIScience Remote Sens.*, vol. 57, no. 3, pp. 369–394, 2020.

[65] C. Yoo, D. Han, J. Im, and B. Bechtel, "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using landsat images," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 155–170, 2019.

[66] J. Yin *et al.*, "Estimation of grassland height based on the random forest algorithm and remote sensing in the Tibetan Plateau," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 178–186, Dec. 2020.

[67] A. I. Weinberg and M. Last, "Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification," *J. Big Data*, vol. 6, no. 1, 23, 2019.

[68] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[69] G. Ridgeway, "Looking for lumps: Boosting and bagging for density estimation," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 379–392, 2002.

[70] R. A. Viscarra Rossel, "Robust modelling of soil diffuse reflectance spectra by 'bagging-partial least squares regression,"' *J. Near Infrared Spectrosc.*, vol. 15, no. 1, pp. 39–47, 2007.

[71] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[72] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv1705.07874*.

[73] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, "Failure mode and effects analysis of RC members based on machine-learning-based SHapley additive exPlanations (SHAP) approach," *Eng. Struct.*, vol. 219, 2020, Art. no. 110927.

[74] H.-C. Thorsen-Meyer *et al.*, "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records," *Lancet Digit. Health*, vol. 2, no. 4, pp. e179–e191, 2020.

[75] R. G. Allen, L. S. Pereira, D. Raes, and M. Smith, *FAO Irrigation and Drainage Paper No. 56*. Rome, Italy: Food Agriculture Org. United Nations, 1998.

[76] K. Saito *et al.*, "The operational JMA nonhydrostatic mesoscale model," *Monthly Weather Rev.*, vol. 134, no. 4, pp. 1266–1298, 2006.

[77] H. Meyer, C. Reudenbach, S. Wöllauer, and T. Nauss, "Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction," *Ecological Model.,* vol. 411, 2019, Art. no. 108815.

[78] W. Li, L. Ni, Z. Li, S. Duan, and H. Wu, "Evaluation of machine learning algorithms in spatial downscaling of MODIS land surface temperature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2299–2307, Jul. 2019.

[79] R. Ren, Y. Yang, and L. Sun, "Oversampling technique based on fuzzy representativeness difference for classifying imbalanced data," *Appl. Intell.*, pp. 1–23, 2020.

[80] D. J. Lorenz, E. T. DeWeaver, and D. J. Vimont, "Evaporation change and global warming: The role of net radiation and relative humidity," *J. Geophys. Res. Atmos.*, vol. 115, no. D20, D20118, 2010.

[81] Y. Yao *et al.*, "A satellite-based hybrid algorithm to determine the Priestley–Taylor parameter for global terrestrial latent heat flux estimation across multiple biomes," *Remote Sens. Environ.*, vol. 165, pp. 216–233, 2015.

[82] B. Martens, D. L. Schumacher, H. Wouters, J. Muñoz-Sabater, N. E. C. Verhoest, and D. G. Miralles, "Evaluating the surface energy partitioning in ERA5," *Geosci. Model Develop. Discuss.*, pp. 1–35, 2020.

[83] B. J. Choudhury, T. J. Dorman, and A. Y. Hsu, "Modeled and observed relations between the AVHRR split window temperature difference and atmospheric precipitable water over land surfaces," *Remote Sens. Environ.*, vol. 51, no. 2, pp. 281–290, 1995.

[84] K.-S. Han, A. A. Viau, Y.-S. Kim, and J.-L. Roujean, "Statistical estimate of the hourly near-surface air humidity in Eastern Canada in merging NOAA/AVHRR and GOES/IMAGER observations," *Int. J. Remote Sens.*, vol. 26, no. 21, pp. 4763–4784, 2005.

[85] K. E. Kunkel, "Simple procedures for extrapolation of humidity variables in the mountainous Western United States," *J. Clim.*, vol. 2, no. 7, pp. 656–669, 1989.

[86] W. J. Duane, N. C. Pepin, M. L. Losleben, and D. R. Hardy, “General characteristics of temperature and humidity variability on Kilimanjaro, Tanzania,” *Arctic Antarctic Alpine Res.*, vol. 40, no. 2, pp. 323–334, 2008.

[87] A. Fries, R. Rollenbeck, T. Nauß, T. Peters, and J. Bendix, “Near surface air humidity in a megadiverse Andean mountain ecosystem of Southern Ecuador and its regionalization,” *Agricultural Forest Meteorol.*, vol. 152, pp. 17–30, 2012.

[88] P. Y. Groisman, R. S. Bradley, and B. Sun, “The relationship of cloud cover to near-surface temperature and humidity: Comparison of GCM simulations with empirical data,” *J. Clim.*, vol. 13, no. 11, pp. 1858–1878, 2000.

[89] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, “Very high resolution interpolated climate surfaces for global land areas,” *Int. J. Climatol.*, vol. 25, no. 15, pp. 1965–1978, 2005.

[90] L. Ding, J. Zhou, X. Zhang, S. Liu, and R. Cao, “Downscaling of surface air temperature over the Tibetan plateau based on DEM,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 73, pp. 136–147, 2018.

[91] M. Gervais, L. B. Tremblay, J. R. Gyakum, and E. Atallah, “Representing extremes in a daily gridded precipitation analysis over the united states: Impacts of station density, resolution, and gridding methods,” *J. Clim.*, vol. 27, no. 14, pp. 5201–5218, 2014.

[92] S. Herrera *et al.*, “Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution,” *Int. J. Climatol.*, vol. 39, no. 9, pp. 3717–3729, 2019.

[93] C. A. Famiglietti, J. B. Fisher, G. Halverson, and E. E. Borbas, “Global validation of MODIS near-surface air and dew point temperatures,” *Geophys. Res. Lett.*, vol. 45, no. 15, pp. 7772–7780, 2018.

[94] C. W. Andrew and S. Hironobu, “Seasonal snowpack dynamics and runoff in a cool temperate forest: Lysimeter experiment in Niigata, Japan,” *Hydrol. Process.*, vol. 19, no. 20, pp. 4179–4200, 2005.

[95] K. Suzuki and Y. Nakai, “Canopy snow influence on water and energy balances in a coniferous forest plantation in northern Japan,” *J. Hydrol.*, vol. 352, no. 1/2, pp. 126–138, 2008.

[96] Y. Nakai, T. Sakamoto, T. Terajima, K. Kitamura, and T. Shirai, “Energy balance above a boreal coniferous forest: A difference in turbulent fluxes between snow-covered and snow-free canopies,” *Hydrol. Process.*, vol. 13, no. 4, pp. 515–529, 1999.

[97] N. P. Molotch *et al.*, “Estimating sublimation of intercepted and sub-canopy snow using eddy covariance systems,” *Hydrol. Process.*, vol. 21, no. 12, pp. 1567–1575, 2007.

[98] T. F. Eck *et al.*, “Influence of cloud, fog, and high relative humidity during pollution transport events in South Korea: Aerosol properties and $pM_{2.5}$ variability,” *Atmos. Environ.*, vol. 232, 2020, Art. no. 117530.

[99] Y. H. Kim and J. J. Baik, “Maximum urban heat island intensity in Seoul,” *J. Appl. Meteorol.*, vol. 41, no. 6, pp. 651–659, 2002.

Haemi Park received the B.S. and M.S. in Earth science from Chiba University, Chiba, Japan, in 2010 and 2012, respectively. She received the Ph.D. degree in environmental remote sensing from the department of civil engineering of the university of Tokyo (UT), Tokyo, Japan, in 2015.

She worked as a postdoc at IRIS lab. in Ulsan National Institute of Science and Technology (UNIST), South Korea from 2015 to 2018. From 2019 to 2020, she was a postdoc of Institute of Industrial Science of UT. Since 2020, she has been working with Japan Aerospace Exploration Agency - Earth Observation Research Center as a researcher. Research interests include carbon/water balance in wetlands, soil moisture, sustainability of forests, and effects of human activity.

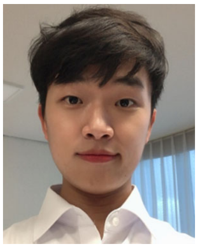
Her major is Environmental Remote Sensing, especially the carbon and water cycle modeling in land area. Between 2015 and 2018, she was a Postdoc of the IRIS Laboratory in Ulsan National Institute of Science and Technology, South Korea. She went back to Institute of Industrial Science, UT and was a Postdoc in 2019. Since 2020, she has been working with Japan Aerospace Exploration Agency – Earth Observation Research Center as an invited researcher. The research interests include carbon/water balance in wetlands, soil moisture, sustainability of forests, and effects of human activity.

Junghee Lee received the B.S. degree in Earth science and engineering from Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, in 2013. She finished the Ph.D. degree in environmental science and engineering from UNIST in 2021. She starts her postdoc in Ecological Sensing Lab at Seoul National University. Her research interests are focused on understanding the interactions between the terrestrial ecosystem and the atmosphere with remote sensing and GIS technologies.

In 2016, she was a Researcher with the National Institute of Forest Science for 11 months. Her research interests include the interactions between the terrestrial ecosystem and the atmosphere using remote sensing and GIS technologies to further understand carbon cycles.

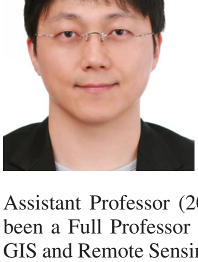
Cheolhee Yoo received the B.S. degree in environmental science and engineering from the Ulsan National Institute of Science and Technology (UNIST) in the Republic of Korea, in 2017. He is currently pursuing a Ph.D. degree in environmental science and engineering from the school of urban and environmental engineering at UNIST. His research interests include urban climatology using satellite remote sensing data and machine learning techniques.

In 2019, he was a Research Intern with Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Japan for three months. His research interests include urban climatology using satellite remote sensing data and machine learning techniques.



Seongmun Sim received the B.S. degree on environmental analysis & pollution control and the earth science & engineering from the Ulsan National Institute of Science and Technology (UNIST) in the Republic of Korea, in 2015. He is currently pursuing a doctoral degree majored on the ocean fog detection/prediction using remote sensing and numerical predicted data with artificial intelligence techniques at the school of urban and environmental engineering of UNIST. His research interests include monitoring of oceanic & atmospheric phenomena using remote sensing and artificial intelligence.

His research interests include monitoring of ocean and atmospheric phenomena using remote sensing and machine learning techniques.



Jungho Im (Member, IEEE) received the B.S. degree in oceanography and the M.C.P. degree in environmental studies from Seoul National University, Seoul, South Korea, in 1998 and 2000, respectively, and the Ph.D. degree in remote sensing and GIS from the University of South Carolina, SC, USA, in 2006.

He was with the SUNY-ESF, Syracuse, NY, USA between 2007 and 2012, serving as an Assistant Professor. He was with Ulsan National Institute of Science and Technology, Ulsan, South Korea as an Assistant Professor (2012–2014), Associate Professor (2014–2019), and has been a Full Professor since 2019. He has been serving as Editor-in-Chief of GIS and Remote Sensing since 2014. His research seeks to broaden and deepen our understanding of the Earth systems on which society depends using remote sensing and artificial intelligence, and leverage this knowledge to better manage and control critical functions related to urban ecology, terrestrial and coastal ecosystems, water resources, natural and man-made disasters, and carbon cycles.