Earth and Space Science

RESEARCH ARTICLE
10.1029/2019EA000740

Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas

Dongjin Cho[1], Cheolhee Yoo[1], Jungho Im[1], and Dong-Hyun Cha[1]

[1]School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea

Abstract Forecasts of maximum and minimum air temperatures are essential to mitigate the damage of extreme weather events such as heat waves and tropical nights. The Numerical Weather Prediction (NWP) model has been widely used for forecasting air temperature, but generally it has a systematic bias due to its coarse grid resolution and lack of parametrizations. This study used random forest (RF), support vector regression (SVR), artificial neural network (ANN) and a multi-model ensemble (MME) to correct the Local Data Assimilation and Prediction System (LDAPS; a local NWP model over Korea) model outputs of next-day maximum and minimum air temperatures ($T_{max_{t+1}}$ and $T_{min_{t+1}}$) in Seoul, South Korea. A total of 14 LDAPS model forecast data, the daily maximum and minimum air temperatures of *in-situ* observations, and five auxiliary data were used as input variables. The results showed that the LDAPS model had an R^2 of 0.69, a bias of −0.85 °C and an RMSE of 2.08 °C for $T_{max_{t+1}}$ forecast, whereas the proposed models resulted in the improvement with R^2 from 0.75 to 0.78, bias from −0.16 to −0.07 °C and RMSE from 1.55 to 1.66 °C by hindcast validation. For forecasting $T_{min_{t+1}}$, the LDAPS model had an R^2 of 0.77, a bias of 0.51 °C and an RMSE of 1.43 °C by hindcast, while the bias correction models showed R^2 values ranging from 0.86 to 0.87, biases from −0.03 to 0.03 °C, and RMSEs from 0.98 to 1.02 °C. The MME model had better generalization performance than the three single machine learning models by hindcast validation and leave-one-station-out cross-validation.

Key Points:
- Machine learning based bias correction of air temperature forecasts of a numerical model
- All machine learning models improved prediction skills of air temperature
- An ensemble of three machine learning models resulted in more robust bias correction than individual machine learning models

Correspondence to:
J. Im,
ersgis@unist.ac.kr

Citation:
Cho, D., Yoo, C., Im, J., & Cha, D.-H. (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and Space Science*, *7*, e2019EA000740. https://doi.org/10.1029/2019EA000740

Received 20 JUN 2019
Accepted 8 FEB 2020
Accepted article online 14 MAR 2020

©2020. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Introduction

Reliable forecasting of air temperature at 2 m above the land surface plays a significant role when preparing for potential weather-related disasters, such as heat waves (i.e., maximum daytime air temperature) and cold spells (i.e., minimum nighttime air temperature). Extreme air temperatures can also cause various social and economic problems such as heat-related disease and high energy consumption (Klinenberg, 2015; Russo et al., 2019). In particular, the increasing intensity, frequency and duration of extreme air temperatures during the summer season (Perkins et al., 2012), and the fact that more than half of the Earth's population now lives in cities (Schulze & Langenberg, 2014) suggest that accurate air temperature forecasting is essential for urban areas.

Numerical Weather Prediction (NWP) models based on the physical relationships of parameters and the mechanisms of atmospheric dynamics have become a valuable tool for forecasting various weather components, including air temperature. However, due to the coarse grid resolution and imperfectness of physical parameterizations, NWP models have typically simplified the detailed characteristics of land, atmosphere, and ocean systems. Despite continuous improvements to model performance, uncertainties of the NWP models, caused by uncertain physical parameterization, inaccurate initial/boundary conditions, and dependency to domain and resolution, result in model bias in air temperature forecasting. Thus, post-processing of the model output to reduce the bias may be required for the operational use of the models. Several statistical methods have been used for the bias correction of air temperature data produced from NWP models (Anadranistakis et al., 2004; de Carvalho et al., 2011; Stensrud & Yussouf, 2003). In many countries, these

techniques have been applied to weather elements generated in the NWP models to increase forecasting performance.

The most commonly used bias correction methods in the air temperature forecasting fields are the Model Output Statistics (MOS) and Kalman Filter (KF) techniques. The MOS improves forecasting accuracy by applying a statistical linear model developed between the past model results and observation data to the NWP model output (Stensrud & Yussouf, 2003). With recent advances in computer resources, KF has been widely used to solve nonlinear problems. When forecasting air temperature, KF first bias-corrects NWP model output. Observed air temperatures are then used to recursively update the parameters applied to the next forecast step (Anadranistakis et al., 2004; de Carvalho et al., 2011; Libonati et al., 2008). Recently, researchers have applied machine learning techniques to improve the prediction accuracy of various types of high-impact weather phenomena (Kim et al., 2017; Lee et al., 2017; Sim et al., 2018; Yoo et al., 2018). Machine learning techniques are not sensitive to the multi-collinearity of input variables, and thus can deal with many input variables. Unlike MOS and KF—in which bias-correction is required to construct a model for each station—machine learning can be used to develop a model that works for a multitude of stations. Because of these advantages, the spatial distributions of the predictand (e.g., air temperature) can be monitored when spatially continuous input variables are fed into machine learning models.

Among various machine learning classifiers, Artificial Neural Network (ANN) has been the most popular technique for air temperature forecasting in the literature (Isaksson, 2018; Marzban, 2003; Vashani et al., 2010; Zjavka, 2016). Marzban (2003) used ANN for post-processing of the Advanced Regional Prediction System (ARPS) model's hourly temperature outputs, getting an average 40% reduction in the mean squared error for all validated weather stations. Vashani et al. (2010) found that the ANN and KF methods demonstrate better bias correction performance than the other methods for the summing accuracy of 30 weather stations in Iran, and ANN produced slightly higher accuracy than KF for longer forecast ranges (i.e., 2 to 5 days ahead forecasts). Zjavka (2016) reported that a polynomial neural network could successfully bias-correct the National Oceanic and Atmospheric Administration (NOAA) meso-scale model to forecast hourly air temperature. Isaksson (2018) compared a deep neural network with KF for the bias correction of air temperature forecasted by the European Centre for Medium-Range Weather Forecasts (ECMWF) model. That paper found that the neural network model shows superior accuracy to KF in error reduction for most validated stations. In addition to ANN, some other machine learning approaches (i.e., Support Vector Regression (SVR) and Random Forest (RF)) have been used to correct the bias of the NWP model's air temperature outputs (Eccel et al., 2007; Yi et al., 2018). Eccel et al. (2007) tested various approaches, from simple correction (i.e., mean bias) to machine learning approaches—ANN and RF, to improve the minimum temperature forecasting skills of two NWP models, ECNWF and Local Area Model Italy (LAMI) in a region of the Italian Alps. They found that, compared to other approaches, RF yielded the best results with the advantage of an easily automated process. Yi et al. (2018) improved the accuracy of air temperature from the Local Data Assimilation and Prediction System (LDAPS) model in Seoul, South Korea, by using SVR and a linear regression model, finding that SVR showed higher correction accuracy than the linear regression model.

In fact, the choice of a regressor among various machine learning options significantly affects the prediction results (Lee et al., 2018; Liu et al., 2018; Park et al., 2018; Wylie et al., 2019). Although several machine learning algorithms have already been used in temperature bias-correction, improving the modeling accuracy remains challenging. Recently, some researchers have tried to increase predicting performance by combining (i.e., ensemble) the results of various machine learning approaches in several different fields (Chou & Pham, 2013; Healey et al., 2018; Ren et al., 2016). All of these studies showed that using various machine learning models together improves performance by overcoming the limitations of each individual classifier. To the best of our knowledge, however, there has been no research on bias correction for the NWP model-derived air temperature through an ensemble of multiple machine learning approaches. Because of the complex atmosphere–surface interactions, a single machine learning algorithm cannot reduce the NWP model bias consistently and effectively. Therefore, combining different machine learning models might reduce daily variations in the errors of the NWP models.

This study aims to correct the bias of the LDAPS air temperatures, one of the NWP model outputs produced by the Korea Meteorological Administration (KMA) using various machine learning–based post-processing

methods. We designed our bias correction models through the integration of NWP model's forecast, *in-situ* maximum and minimum air temperatures of present-day, and auxiliary data including location (i.e., coordinate) and topographic variables, to correct the NWP model's next-day maximum and minimum air temperatures forecast for urban areas. The key objectives of this study were to (1) develop machine learning-based bias correction models—RF, SVR, ANN, and a multi-model ensemble (MME)—to improve NWP model-derived daily maximum and minimum air temperature, and (2) examine the spatiotemporal characteristics of the corrected temperature in comparison with the NWP model output.

2. Study Area and Data

2.1. Study Area

The study area is a metropolitan city, Seoul (~605 km^2) with over 10 million people, which is situated in the northwestern part of South Korea (Figure 1). Seoul is geographically surrounded by four distinct mountains and is divided into the northern and southern parts by the Han River. It is hot and humid due to the East Asian monsoon in summer, so there are a lot of days with hot weather above 30 °C, and precipitation is concentrated in summer. (i.e., seasonal rainfall is 892.1 mm in summer and 67.3 mm in winter during 1981–2010.)

2.2. Local Data Assimilation and Prediction System (LDAPS) Model Data

The KMA has operated several NWP models which were constructed by adopting the United Model (UM) of the UK Met Office in May 2008. The Global Data Assimilation and Prediction System (GDAPS) and Regional Data Assimilation and Prediction System (RDAPS) models have been operating since May 2010. However, these models have a limitation to capture extreme weather events (e.g., heavy precipitation, severe snowfall and heat wave) in small areas due to their coarse spatial resolution. To overcome the spatial resolution and time scale limitations of the GDAPS and RDAPS models, the KMA has also been operating the LDAPS model based on the UM.

The LDAPS model is designed to predict mesoscale weather phenomena in the region including the Korean peninsula and the surrounding seas. It is a grid-point model and uses non-hydrostatic dynamics with semi-Lagrangian advection and semi-implicit time stepping (Orr et al., 2014). The LDAPS has a rotated latitude-longitude grid with 1.5 km horizontal resolution and a hybrid-height vertical coordinate with 70 layers up to 40 km. It is operated every three hourly, that is, eight times a day (i.e., 0, 3, 6, 9, 12, 15, 18 and 21 UTC). However, the forecasts only for 0, 6, 12, and 18 UTC are official operations with 36-hour forecast leading time, while the others with 3-hour forecast time are conducted for model initialization.

The present study is focused on a 5-year period of data of July and August, from 2013 to 2017 when the LDAPS model data are available. The lowest layer (surface layer) data of LDAPS model produced at 21 KST (12 UTC) for the next-day forecast was used: next-day maximum and minimum air temperatures ($LDAPS_{T_{max}}$ and $LDAPS_{T_{min}}$, respectively) and relative humidity, next-day average wind speed and latent heat flux, and next-day 6-hour split average (e.g., each average of 0–5, 6–11, 12–17, and 18–23 hour) cloud cover and precipitation forecast data. The cloud cover value was calculated by averaging high, medium and low cloud cover values.

2.3. *In-situ* Data

The next-day maximum and minimum air temperatures ($T_{max_{t+1}}$ and $T_{min_{t+1}}$, respectively) during July and August from 2013 to 2017 were obtained from 25 Automatic Weather Stations (AWSs) operated by KMA in Seoul. Hourly air temperature data from the 25 AWSs were also collected to determine present-day maximum and minimum air temperatures (T_{max_t} and T_{min_t}, respectively). T_{max_t} and T_{min_t} were determined as the maximum and minimum air temperature values between 0 and 21 KST because the LDAPS model data produced at 21 KST was used, which was compared with post-processing bias correction models.

2.4. Auxiliary Variables

Latitude, longitude, elevation, slope, and solar radiation were used as additional auxiliary input data in this study. Latitude and longitude values were extracted as the median of the grid in the LDAPS model. The elevation and slope values were derived from Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) with 30 m spatial resolution. Solar radiation was calculated using the 'Area Solar radiation'

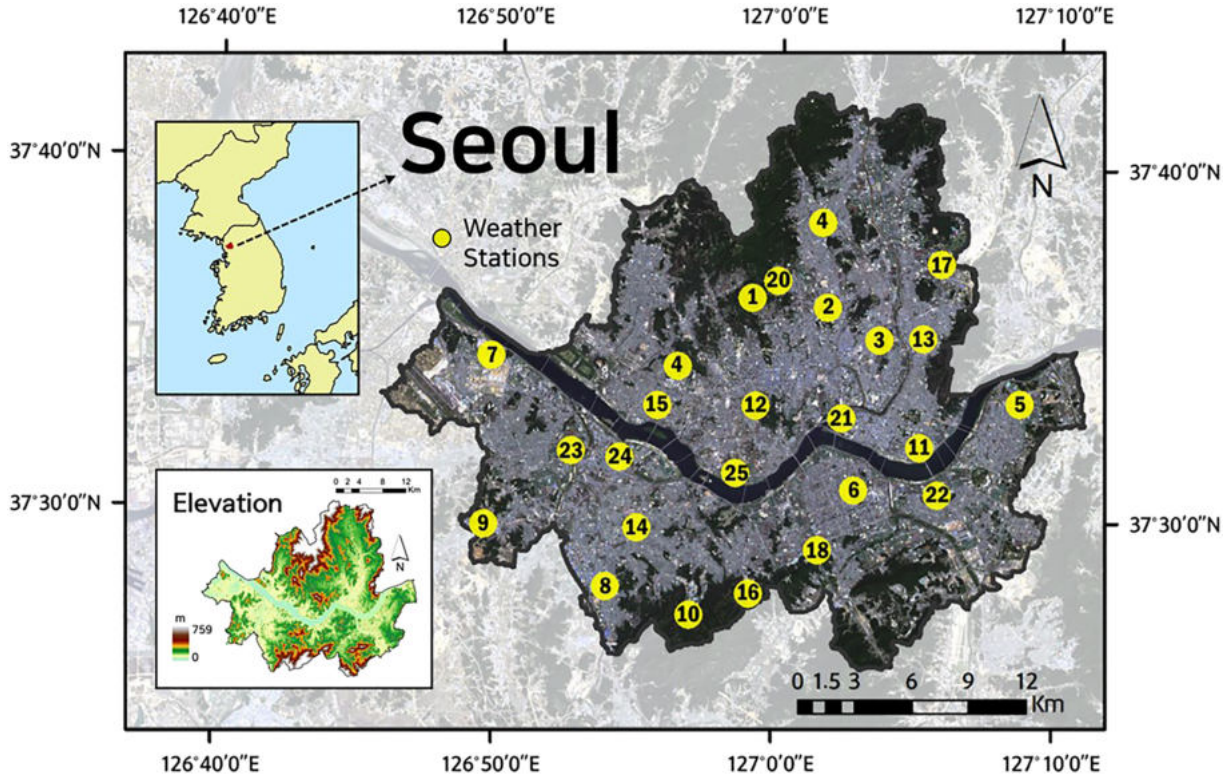

Figure 1. Study area and the location of automatic weather stations (AWS) operated by KMA. Landsat 8 RGB band composite acquired on may 19, 2016 is used as a background image. Elevation is from Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) with 30 m spatial resolution.

tool and elevation and day of year (DOY) values in ArcMap to obtain the daily average incoming solar radiation for each pixel. It was assumed that incoming solar radiation follows the same temporal pattern by year. Latitude and longitude were considered as location (i.e., coordinate) variables, while elevation, slope, and solar radiation were grouped as topographic variables in this study.

3. Methods

3.1. Data Processing

Figure 2 summarizes the process flow of our proposed methodology. In this study, a total of 14 LDAPS model forecast data, $\mathbf{T_{max_t}}$, $\mathbf{T_{min_t}}$, and five auxiliary variables were used as input variables, while $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ were used as target variables (Table 1). LDAPS-derived temperature forecast data were processed using lapse rates with elevation data developed by Yun et al. (2001) to calculate more reliable temperature values because the elevation information used in the LDAPS model was not accurate due to its smoothed surface which is applied to eliminate grid-scale numerical noise in NWP models (Bosart et al., 1998; Wallace et al., 1983; Webster et al., 2003). For mapping $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ using the developed models, point-based *in-situ* $\mathbf{T_{max_t}}$ and $\mathbf{T_{min_t}}$ data were rasterized through cokriging interpolation. The cokriging technique is known to have a higher estimation accuracy than the kriging or the Inverse Distance Weighted Method (IDW) in interpolating hourly air temperature (Ishida & Kawashima, 1993). With cokriging, the elevation data was used as a co-variable, which is considered to be one of the most important additional variables for estimating air temperature (Aznar et al., 2013). We selected the ordinary cokriging method among several cokriging techniques (e.g., simple cokriging and universal cokriging) through a comparison of their performance.

3.2. Machine Learning Approaches

The first bias correction model, RF, is an ensemble machine learning algorithm that predicts a target variable from a set of predictors by growing multiple trees (CART; Breiman, 2001) and aggregating their

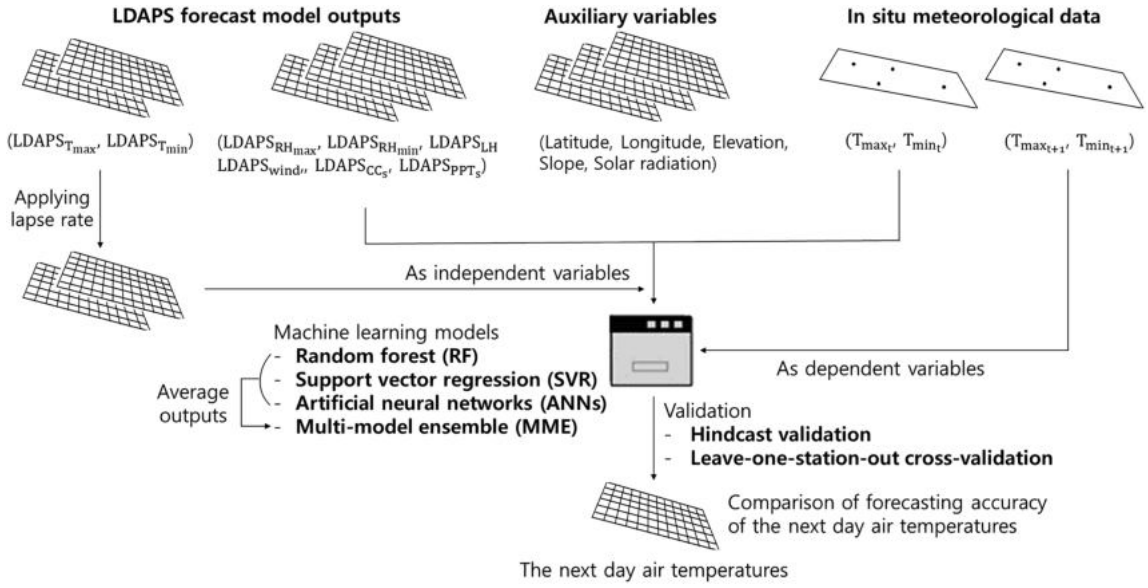

Figure 2. The process flow diagram of the proposed approach. The subscript of 's' in the $LDAPS_{CC}$ and $LDAPS_{PPT}$ implies the series of the 6-hour split.

results. RF has been widely used to solve a multitude of classification and regression problems (Kim et al., 2015; Lee, Im, et al., 2016; Liu et al., 2015; Yoo et al., 2012). RF generates a multitude of CARTs (typically 500–1000 trees) through bootstrapping-based randomization approaches for the selection of both training samples for a tree and predictors at each node of the tree (Im et al., 2016; Richardson et al., 2017). This approach alleviates any existing problems in the CART such as overfitting and sensitivity to training samples (Forkuor et al., 2018; Yoo et al., 2018). R software with the 'randomForest' package was used to develop and apply the statistical models using default model parameter settings (Ho et al., 2014; Yoo et al., 2019).

The second model we used, the SVR algorithm, aims to get the optimal hyperplane that fits the data and predicts with minimal empirical errors. SVR generally converts training data from the original dimension to a higher dimension to effectively find the optimum hyperplane. Selecting a kernel function that is used to convert training data to a higher dimension is crucial for successful implementation of SVR. Linear, polynomial and Gaussian functions are typically used as kernels. In this study, SVR was implemented in MATLAB 2018a using the 'fitrsvm' function (https://mathworks.com/help/stats/fitrsvm.html) with the linear kernel function. To determine the parameters of SVR, many previous studies used the grid search method (Das & Padhy, 2018; Han et al., 2017; Mansaray et al., 2019; Shao & Lunetta, 2012). However, we used an automatic kernel scaling approach, which selects the appropriate parameters using a heuristic procedure because of the time-consuming drawback of the grid search method in the hindcast validation (Boardman & Trappenberg, 2006; Wang et al., 2005). A Sequential Minimal Optimization (SMO) was used as a solver in the training process (Flake & Lawrence, 2002; Platt, 1998).

The third model, ANN, has an interconnected structure which emulates the operations and connectivity of biological neurons in human brain (Özçelik et al., 2010; Tiryaki & Aydın, 2014). Among numerous proposed ANN algorithms including Radial Basis Function (Kecman, 2001), Elman recurrent (Rakkiyappan & Balasubramaniam, 2008), and Hopfield neural networks (Nguyen et al., 2006), this study used a multi-layer perceptron (MLP) neural network that consists of input, output and hidden layers with a back-propagation algorithm, which is the most popular due to its ease of training, meaning it has been widely used in many applications including forecasting (Omrani et al., 2017; Pham et al., 2017; Sonobe et al., 2017; Yang et al., 2019). The architecture of ANN is crucial to its performance. A simple network with few hidden layers and neurons may restrict the learning ability, while a complex network with too many hidden layers and neurons may lead to overfitting and poor generalization (Moghim & Bras, 2017). A trial and error procedure determined that two hidden layers consisting of 22 and 19 neurons, respectively, with the

rectified linear unit (ReLU) activation function and Adam optimization algorithm is an appropriate choice for the bias correction model of temperature prediction. In addition, since ANN might not accurately predict the required output with the randomly selected initial weights and biases (Lee, Geem, & Suh, 2016), the near-global optimal weights and biases were determined through a trial and error procedure.

Previous studies have proposed MME to combine multiple machine learning models. Generally, this approach enables users to achieve higher accuracy and robustness when compared to using a single model (Adeodato et al., 2011; Van Wezel & Potharst, 2007). With the bias-variance trade-off in machine learning, the generalization of predictive models is more likely to be improved by an ensemble approach from the models. We constructed an MME model using simple average ensemble, which does not require extra training to finding the weights of each ensemble member and has a low level of complexity (Gorissen et al., 2009), although the other methods can be further explored to improve results. In addition, three cases—MME without location (i.e., coordinate) variables (case 1), MME without topographic ones (case 2), and MME without both location and topographic ones (case 3)—were further compared to the MME model developed with all variables to see the effect of using geographic information as input data for bias correction.

3.3. Accuracy Assessment

Hindcast validation was conducted for the period from 2015 to 2017. The validation results of the developed machine learning-based bias correction models were compared to those of the LDAPS model. For example, samples from the first day of the study period to July 31, 2016 were trained to predict air temperature on August 1, 2016. In addition, leave-one-station-out cross-validation (LOSOCV) was conducted to spatially compare the generalization ability of the developed bias correction models. Three accuracy metrics—coefficient of determination (R^2, Equation 1), bias (Equation 2), and root mean square error (RMSE, Equation 3—were used for accuracy assessment.

$$\mathbf{R^2} = \mathbf{1} - \frac{\sum_1^n \boldsymbol{y}_i - \boldsymbol{y}^{-2}}{\sum_1^n \boldsymbol{y}_i - \boldsymbol{y}^{-2}}, \boldsymbol{y}^{-} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_1^n \boldsymbol{y}_i \tag{1}$$

$$\mathbf{Bias} = \sum_1^n \frac{(\widehat{\boldsymbol{y}}_i - \boldsymbol{y}_i)}{\boldsymbol{n}}, \tag{2}$$

$$\mathbf{RMSE} = \sqrt{\sum_1^n \frac{(\widehat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2}{\boldsymbol{n}}}, \tag{3}$$

where $\boldsymbol{y_i}$ is the measured value, $\widehat{\boldsymbol{y}}_i$ is the predicted value of each model, and n is the number of samples. The skill score (SS) was also used to summarize the improvement of corrected forecasts with respect to the LDAPS model forecasts at each weather station (Libonati et al., 2008; Wilks, 2011). The SS as measured in terms of the RMSE is given by,

$$\mathbf{SS} = \frac{(\mathbf{RMSE_{LDAPS}} - \mathbf{RMSE_{BC}})}{\mathbf{RMSE_{LDAPS}}} \times \mathbf{100}\%, \tag{4}$$

where $\mathbf{RMSE_{LDAPS}}$ and $\mathbf{RMSE_{BC}}$ are the RMSEs of LDAPS model and bias corrected forecasts. Positive (negative) SS values indicate that bias corrected forecasts improved (worsened) the predictions.

4. Result and Discussion

4.1. LDAPS Model Performance Analysis

The LDAPS model forecast performances for each station are Presented in Table 2. The average R^2 and RMSE values of the LDAPS model for the 25 stations in Seoul were 0.74 and 1.91 °C for forecasting $\mathbf{T_{max_{t+1}}}$ and 0.83 and 1.38 °C for forecasting $\mathbf{T_{min_{t+1}}}$, respectively. Stations that had a higher RMSE than the average had more than one degree absolute bias value with a high R^2, implying that the LDAPS model might have systematic errors at these stations. These temperature errors of the numerical model can be associated with uncertain physical parameterizations within boundary layer/land surface (Zheng et al., 2017) and inaccurate lower boundary conditions (i.e., land-use and topography) (Li et al., 2018; Zhang et al., 2009).

Table 1
Description of Variables Used in This Study

Variable type		Acronym (unit)	Description
LDAPS model data		$LDAPS_{T_{max}}$ (°C)	LDAPS model forecast of next-day maximum air temperature
		$LDAPS_{T_{min}}$ (°C)	LDAPS model forecast of next-day minimum air temperature
		$LDAPS_{RH_{max}}$ (%)	LDAPS model forecast of next-day maximum relative humidity
		$LDAPS_{RH_{min}}$ (%)	LDAPS model forecast of next-day minimum relative humidity
		$LDAPS_{wind}$ (m/s)	LDAPS model forecast of next-day average wind speed
		$LDAPS_{LH}$ (W/m^2)	LDAPS model forecast of next-day average latent heat flux
		$LDAPS_{CC_1}$ (%)	LDAPS model forecast of next-day 1st 6-hour split average cloud cover (0–5 h)
		$LDAPS_{CC_2}$ (%)	LDAPS model forecast of next-day 2nd 6-hour split average cloud cover (6–11 h)
		$LDAPS_{CC_3}$ (%)	LDAPS model forecast of next-day 3rd 6-hour split average cloud cover (12–17 h)
		$LDAPS_{CC_4}$ (%)	LDAPS model forecast of next-day 4th 6-hour split average cloud cover (18–23 h)
		$LDAPS_{PPT_1}$ (%)	LDAPS model forecast of next-day 1st 6-hour split average precipitation (0–5 h)
		$LDAPS_{PPT_2}$ (%)	LDAPS model forecast of next-day 2nd 6-hour split average precipitation (6–11 h)
		$LDAPS_{PPT_3}$ (%)	LDAPS model forecast of next-day 3rd 6-hour split average precipitation (12–17 h)
		$LDAPS_{PPT_4}$ (%)	LDAPS model forecast of next-day 4th 6-hour split average precipitation (18–23 h)
In-situ data		T_{max_t} (°C)	Maximum air temperature between 0 and 21 h on the present-day
		T_{min_t} (°C)	Minimum air temperature between 0 and 21 h on the present day
		$T_{max_{t+1}}$ (°C)	The next-day maximum air temperature
		T_{min_t+1} (°C)	The next-day minimum air temperature
Auxiliary data	Location variables	Lat (°)	Latitude
		Lon (°)	Longitude
	Topographic variables	Elev (m)	Elevation
		Slop (°)	Slope
		Sol (wh/m^2)	Daily incoming solar radiation

Note. All data were aggregated to the spatial resolution of the LDAPS model (1.5 km).

For stations 4 and 20, the differences between the SRTM elevation aggregated with the LDAPS model spatial resolution and the topographic height in the LDAPS model were greater than 200 m. Such elevation differences reaching several hundreds of meters are one of the main reasons that large systematic errors in the temperature forecasts are generated (Hart et al., 2004; Libonati et al., 2008). As a result, it is concluded that the LDAPS model had a distinct high cold bias at stations 4 and 20 when forecasting both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$. For the other stations in terms of systematic errors, such as stations 7 and 18 for $\mathbf{T_{max_{t+1}}}$ and stations 12 and 16 for $\mathbf{T_{min_{t+1}}}$, the LDAPS model forecast mostly showed a cold bias for $\mathbf{T_{max_{t+1}}}$, whereas it showed a warm bias for $\mathbf{T_{min_{t+1}}}$, which implies that the LDAPS model generally underestimates the next-day daily temperature range.

4.2. Comparative Evaluation of Bias Correction Models

The hindcast validation results of LDAPS and all bias corrected models during July and August from 2015 to 2017 for $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ are depicted in Figures 3 and 4, respectively. The LDAPS model had an R^2 of 0.69, a bias of −0.85 °C, and an RMSE of 2.08 °C for forecasting $\mathbf{T_{max_{t+1}}}$ (Figure 3(a)), while all three bias correction models and the MME model resulted in various levels of improvement, with R^2 ranging from 0.75 to 0.78,

Table 2
The LDAPS Model Forecast Performances for Each Station During July and August From 2013 to 2017

	$T_{max_{t+1}}$			$T_{min_{t+1}}$			
Station number	R^2	Bias (°C)	RMSE (°C)	R^2	Bias (°C)	RMSE (°C)	ΔElev (m)
1	0.76	0.10	1.44	0.87	1.01	1.29	41
2	0.77	−0.37	1.51	0.85	1.35	1.64	52
3	0.74	0.54	1.68	0.86	0.87	1.24	3
4	0.75	−3.46	3.79	0.83	−2.14	2.35	247
5	0.71	−0.85	1.83	0.85	0.37	0.96	21
6	0.77	−0.23	1.47	0.84	0.33	0.98	−32
7	0.73	−1.58	2.18	0.81	−0.07	1.02	6
8	0.76	−0.30	1.51	0.84	0.42	1.04	−35
9	0.71	−1.51	2.23	0.82	0.40	1.06	−18
10	0.75	−0.29	1.50	0.77	1.48	1.88	−97
11	0.74	−0.04	1.58	0.83	0.69	1.15	−10
12	0.69	0.23	1.86	0.86	1.91	2.11	−64
13	0.74	0.15	1.62	0.87	0.30	0.91	−40
14	0.73	0.08	1.55	0.81	1.09	1.48	−20
15	0.69	−0.06	1.71	0.86	1.22	1.51	−5
16	0.73	−0.06	1.50	0.82	1.59	1.86	−14
17	0.74	−0.33	1.57	0.84	1.51	1.84	−14
18	0.71	−1.94	2.63	0.81	−0.13	1.01	−24
19	0.73	−0.24	1.63	0.85	1.48	1.76	−26
20	0.76	−3.57	3.86	0.84	−1.75	1.97	230
21	0.75	0.58	1.63	0.83	0.45	1.07	20
22	0.73	−0.80	1.78	0.82	0.39	1.09	−6
23	0.73	−1.71	2.32	0.82	0.11	1.00	7
24	0.75	−0.88	1.73	0.82	0.20	1.02	−1
25	0.75	−0.14	1.57	0.84	0.60	1.14	−7
Average	0.74	−0.67	1.91	0.83	0.55	1.38	9

Note. ΔElev represents the difference between the SRTM elevation aggregated to 1.5 km and topography height in the LDAPS model.

bias from −0.16 to −0.07 °C and RMSEs from 1.55 to 1.66 °C (Figure 3(b)-(e)). Among the bias correction models, the RF model had the largest RMSE (1.66 °C) when compared to the other models for forecasting $\mathbf{T_{max_{t+1}}}$. Especially, the RF model overestimated low temperatures which were outside the range of the observed values in the training data. RF is known to be prone to overestimating low values and underestimating high values because it averages over all independent trees, which means that RF cannot extrapolate outside of the training data (Horning, 2013; Kühnlein et al., 2014; Shah et al., 2014). While the SVR and ANN models produced similar results, the SVR model yielded better estimates of low temperatures than the ANN model. This is consistent with the findings of a previous study, which noted that ANN is very accurate but not able to solve extrapolation problems, whereas SVR is capable of solving extrapolation tasks (Balabin & Smirnov, 2012). The MME model showed the lowest RMSE (1.55 °C) compared to the other models. When forecasting $\mathbf{T_{min_{t+1}}}$, the LDAPS model had an R^2 of 0.77, a bias of 0.51 °C and an RMSE of 1.43 °C (Figure 4(a)), whereas the bias correction models showed R^2 ranging from 0.86 to 0.87, biases of −0.03 to 0.03 °C and RMSEs of 0.98 to 1.02 °C (Figure 4(b)-(e)). Among the bias correction models, all single machine learning-based bias correction models yielded similar results in terms of R^2, bias and RMSE. The MME model showed slightly better performance when compared to the other methods in terms of RMSE when forecasting $\mathbf{T_{min_{t+1}}}$, which is a similar result to that noted when it was used to correct $\mathbf{T_{max_{t+1}}}$.This result is consistent with the findings of previous studies which found that a machine learning-based ensemble model outperformed individual machine learning models (Adeodato et al., 2011; Chou & Pham, 2013).

All bias correction models forecasting both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ improved upon the LDAPS model by year (Table 3). In particular, they had the most improved performance in 2016 for forecasting $\mathbf{T_{max_{t+1}}}$ and in 2017 for forecasting $\mathbf{T_{min_{t+1}}}$, because the LDAPS model forecast had the highest absolute bias in 2016 for $\mathbf{T_{max_{t+1}}}$, and in 2017 for $\mathbf{T_{min_{t+1}}}$. When comparing the individual machine learning-based bias correction models, we found that the best model among the three single machine learning-based bias correction models varied by year. For example, the SVR model had the lowest RMSEs of 1.06 °C in 2015 and 2016 for forecasting $\mathbf{T_{min_{t+1}}}$, whereas the RF model had the lowest RMSE of 0.85 °C in 2017. However, the MME model showed an RMSE lower than the other models in most years for forecasting both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$. Since this work used a hindcast validation approach, this result ensures that the generalization performance of forecasting air temperature is improved by the MME model.

Figure 5 shows the SS values of each bias correction model from the hindcast validation results at the 25 stations. The SS values were strongly station-dependent. All bias correction models yielded positive SS values between 20% and 70% at the stations where the LDAPS model had a remarkable bias (e.g., stations 4, 7, 20 and 23 for $\mathbf{T_{max_{t+1}}}$ and stations 4, 10, 12 and 20 for $\mathbf{T_{min_{t+1}}}$). Especially, for forecasting both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$, great improvements reaching around 60% were observed at stations 4 and 20, which had a negative bias due to the high elevation errors in the LDAPS model. For the RF, SVR, and ANN models, the SS value at each station differed from model to model. On the contrary, the MME model had values similar to the highest SS among the three single machine learning models at all stations, and achieved higher improvements when compared to the other models for most stations. This shows that the MME model produced better accuracy and robustness than a single machine learning model. However, there were some stations with negative SS values in all bias correction models, including MME. One possible reason is that the models were trained for multiple stations (i.e., 25 stations in Seoul). Isaksson (2018) found that the results of training using multiple stations were not as good as using a single station when correcting bias in temperature forecasts produced by NWP models. Another possible reason is that the LDAPS model has been constantly

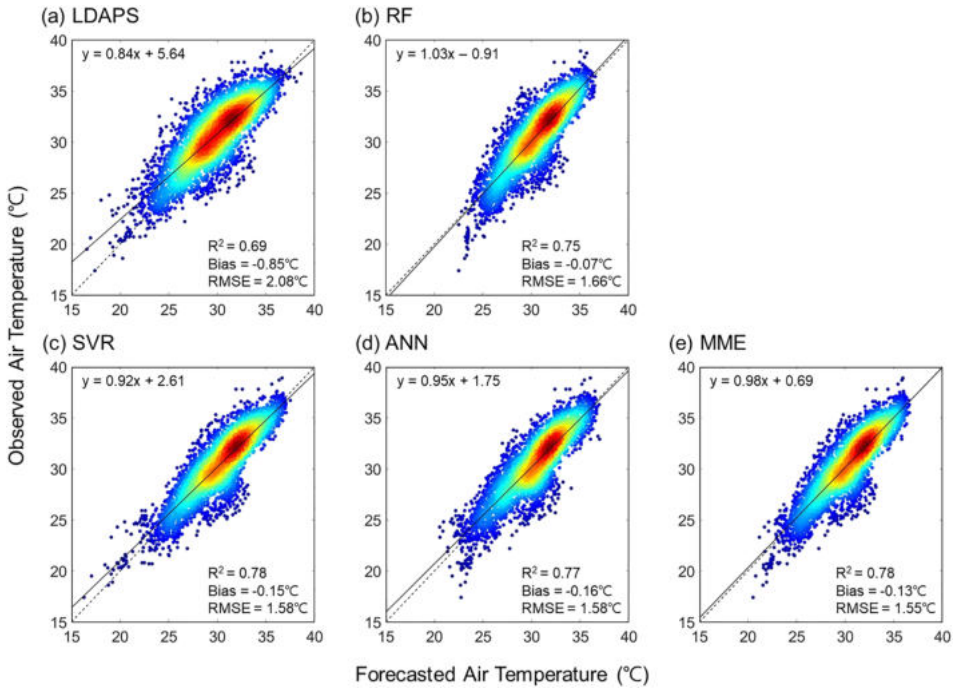

Figure 3. Scatter plots between forecasted and observed air temperatures from hindcast validation results based on (a) LDAPS and bias corrected models using (b) RF, (c) SVR, (d) ANN, and (e) MME during July and august from 2015 to 2017 for $\mathbf{T_{max_{t+1}}}$ forecast. The color ramp from blue to red corresponds to increasing point density.

updated throughout the study period, which may affect the performance of the machine learning-based bias correction models.

The time-series of the daily RMSE and R^2 values of the LDAPS and all bias correction models were compared for 2017, the last year of the study period (Figure 6). Both bias corrected $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecasts

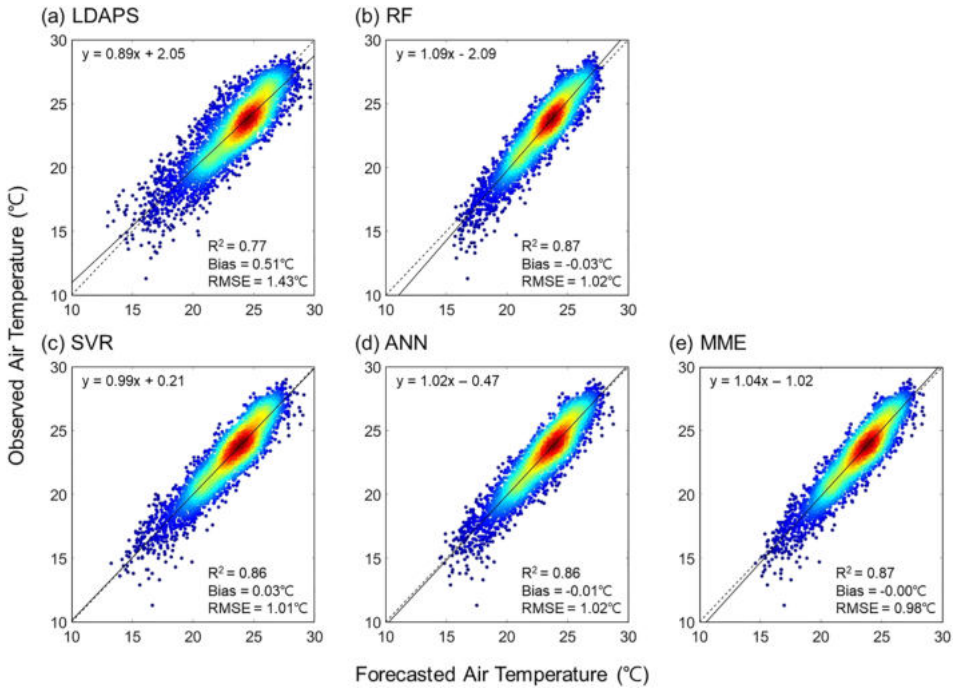

Figure 4. Same as Figure 3 but for $\mathbf{T_{min_{t+1}}}$ forecast.

Table 3
Yearly Hindcast Validation Results of LDAPS, RF, SVR, ANN, and MME Models for $\mathbf{T}_{\mathbf{max}_{t+1}}$ and $\mathbf{T}_{\mathbf{min}_{t+1}}$ Forecasts

$\mathbf{T}_{\mathbf{max}_{t+1}}$

Year		LDAPS	RF	SVR	ANN	MME
2015	R^2	0.58	0.62	0.67	0.68	0.68
	Bias (°C)	−0.93	−0.23	−0.34	−0.36	−0.31
	RMSE (°C)	2.07	1.65	1.56	1.56	1.53
2016	R^2	0.80	0.85	0.87	0.85	0.87
	Bias (°C)	−1.22	−0.33	−0.45	−0.41	−0.40
	RMSE (°C)	2.15	1.56	1.49	1.55	1.45
2017	R^2	0.65	0.70	0.74	0.74	0.74
	Bias (°C)	−0.41	0.35	0.33	0.28	0.32
	RMSE (°C)	2.04	1.76	1.69	1.64	1.65

$\mathbf{T}_{\mathbf{min}_{t+1}}$

Year		LDAPS	RF	SVR	ANN	MME
2015	R^2	0.69	0.79	0.80	0.78	0.80
	Bias (°C)	0.57	−0.01	0.04	0.02	0.02
	RMSE (°C)	1.47	1.09	1.06	1.12	1.05
2016	R^2	0.81	0.89	0.89	0.89	0.90
	Bias (°C)	0.38	−0.11	−0.09	−0.08	−0.09
	RMSE (°C)	1.43	1.09	1.06	1.07	1.03
2017	R^2	0.77	0.89	0.87	0.88	0.89
	Bias (°C)	0.58	0.03	0.14	0.04	0.07
	RMSE (°C)	1.39	0.85	0.92	0.87	0.84

generally showed a lower daily RMSE than the LDAPS model, resulting in an increasing pattern as the RMSE of the LDAPS model increased (Figure 6(a), (c)). This is because the time-series of the bias corrected temperatures were closer to the observations but behaved similarly to the LDAPS model (not shown), which is consistent with the results of Isaksson (2018), in that the machine learning-based bias correction model depends on the initial forecast without radically producing different results. For the $\mathbf{T}_{\mathbf{max}_{t+1}}$ forecast, there were days when all bias correction models had higher RMSEs than that of the LDAPS model such as DOY 196, 204 and 208. $\mathbf{T}_{\mathbf{max}_t}$ was used as an input variable of the bias correction models, and for these days, the observed $\mathbf{T}_{\mathbf{max}_{t+1}}$ tended to be significantly lower than $\mathbf{T}_{\mathbf{max}_t}$ (not shown), because it was clear on the present-day, but mostly cloudy and rainy the next-day. Thus, bias corrected models overestimated $\mathbf{T}_{\mathbf{max}_{t+1}}$, resulting in higher RMSEs than the LDAPS model.

The LDAPS model generally showed very low R^2 for both $\mathbf{T}_{\mathbf{max}_{t+1}}$ and $\mathbf{T}_{\mathbf{min}_{t+1}}$ forecasts (Figure 6(b), (d)), which means that the LDAPS model does not properly simulate the spatial distribution of the temperature within a city (i.e., local scale). The main reason is that the LDAPS model has a high cold bias at stations 4 and 20. All bias correction models had high R^2 variations, but generally higher than those obtained with the LDAPS model. Among them, the RF model generally showed the highest R^2 for the $\mathbf{T}_{\mathbf{max}_{t+1}}$ forecast, while the RMSE of the RF model was higher than the other bias correction models as well as for the $\mathbf{T}_{\mathbf{min}_{t+1}}$ forecast. The MME model had a higher R^2 than SVR and ANN models, which is considered to be due to the high R^2 of the RF model.

4.3. Spatial Performance of the Bias Correction Models

All bias correction models were validated through LOSOCV for forecasting both $\mathbf{T}_{\mathbf{max}_{t+1}}$ and $\mathbf{T}_{\mathbf{min}_{t+1}}$ (Figure 7). The RF model had the highest RMSE as well as the lowest R^2 for forecasting both $\mathbf{T}_{\mathbf{max}_{t+1}}$ and

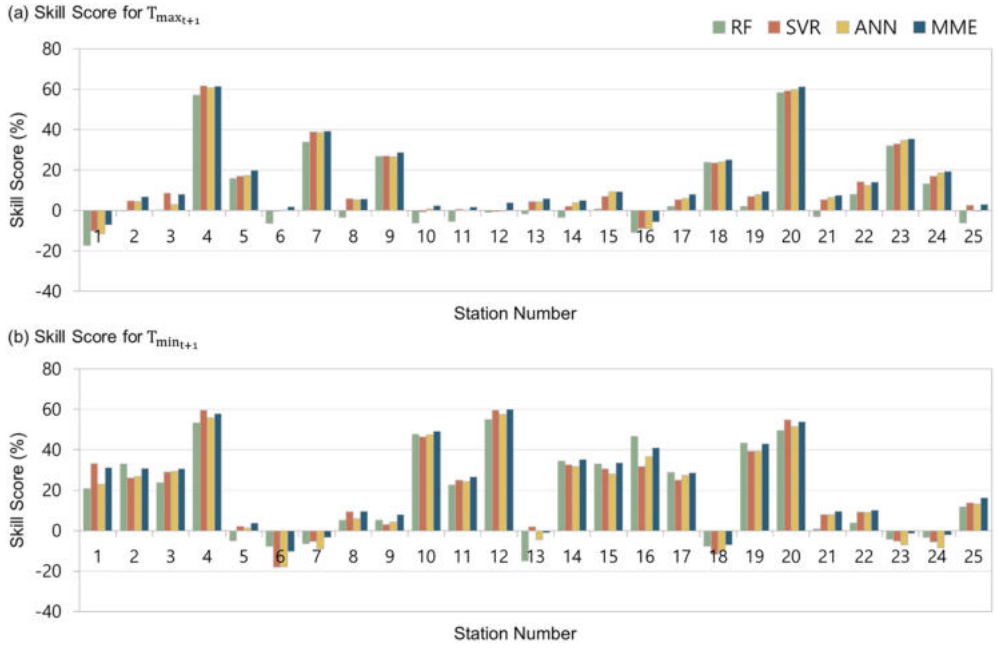

Figure 5. Skill scores from hindcast validation results based on RF, SVR, ANN, and MME models at the 25 stations for (a) $\mathbf{T}_{\mathbf{max}_{t+1}}$ and (b) $\mathbf{T}_{\mathbf{min}_{t+1}}$ forecasts.

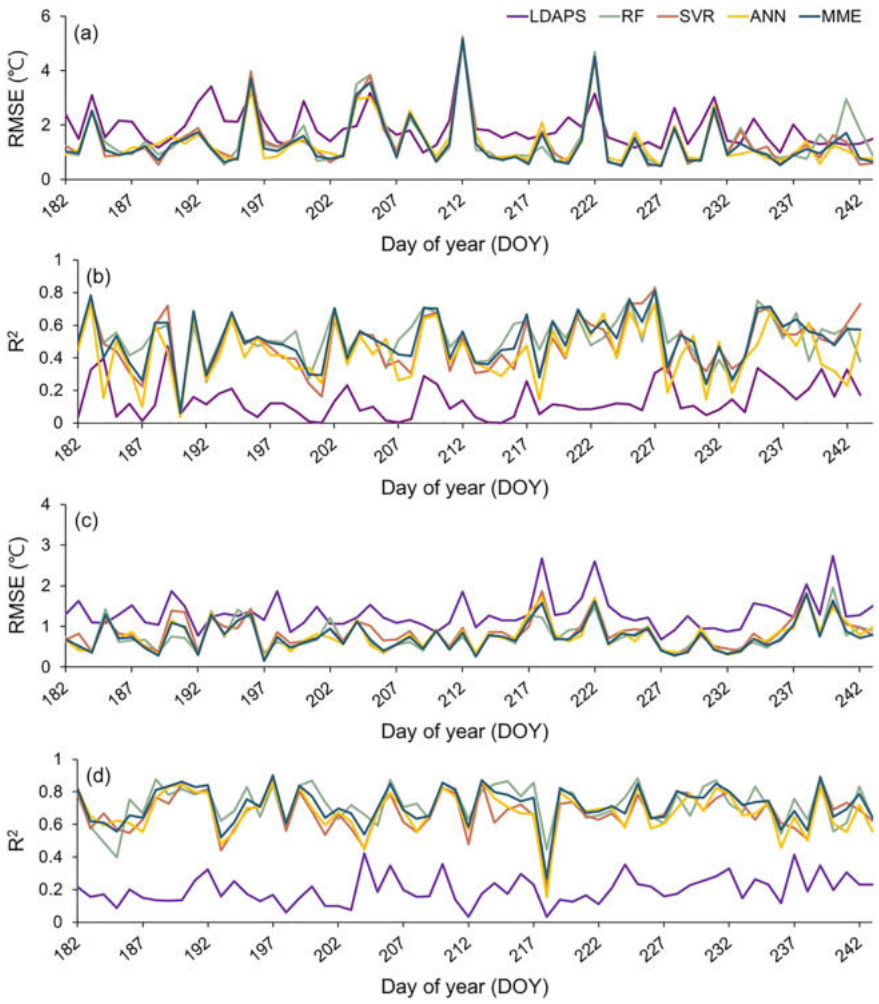

Figure 6. Time-series of daily RMSE and R^2 of LDAPS, RF, SVR, ANN, and MME models for (a) and (c) $\mathbf{T_{max_{t+1}}}$ and (b) and (d) $\mathbf{T_{min_{t+1}}}$ forecast in 2017.

$\mathbf{T_{min_{t+1}}}$. For SVR and ANN models, they showed similar results from the hindcast validation, but the SVR model had a lower RMSE than the ANN model from LOSOCV, implying that the SVR model had a better spatial generalization ability. The MME model showed a slightly lower RMSE than the other models both forecasting $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$. This result indicates that the spatially generalized ability is improved by the MME model.

The SS values of each bias correction model from the LOSOCV results at the 25 stations are shown in Figure 8. The SS values of all bias correction models were lower than the hindcast validation results at all stations because the models were not trained for each station, but they still showed positive values at most stations, especially for $\mathbf{T_{min_{t+1}}}$ forecast. These bias correction models can be judged to have been spatially generalized using multiple stations, which implies that they were able to bias correct the LDAPS model output where stations were not provided in Seoul. However, at station 1 for $\mathbf{T_{max_{t+1}}}$ forecast, all bias correction models had SS values 20% lower than the hindcast validation results (Figure 8(a)), which may be because the observed $\mathbf{T_{max_{t+1}}}$ was substantially lower than the readings at the other stations. There were no stations at which all bias correction models had significantly lower SS values than the hindcast validation results for $\mathbf{T_{min_{t+1}}}$ forecast (Figure 8(b)), but only the RF model had such stations (i.e., stations 1, 4, 17 and 20). One possible reason is that the LDAPS model had a distinct cold bias at stations 4 and 20, but a warm bias at the others, and stations 1 and 17 had lower $\mathbf{T_{min_{t+1}}}$ than the others, which may have affected the RF model, so that it was unable to extrapolate beyond the range of the training data (Shah et al., 2014).

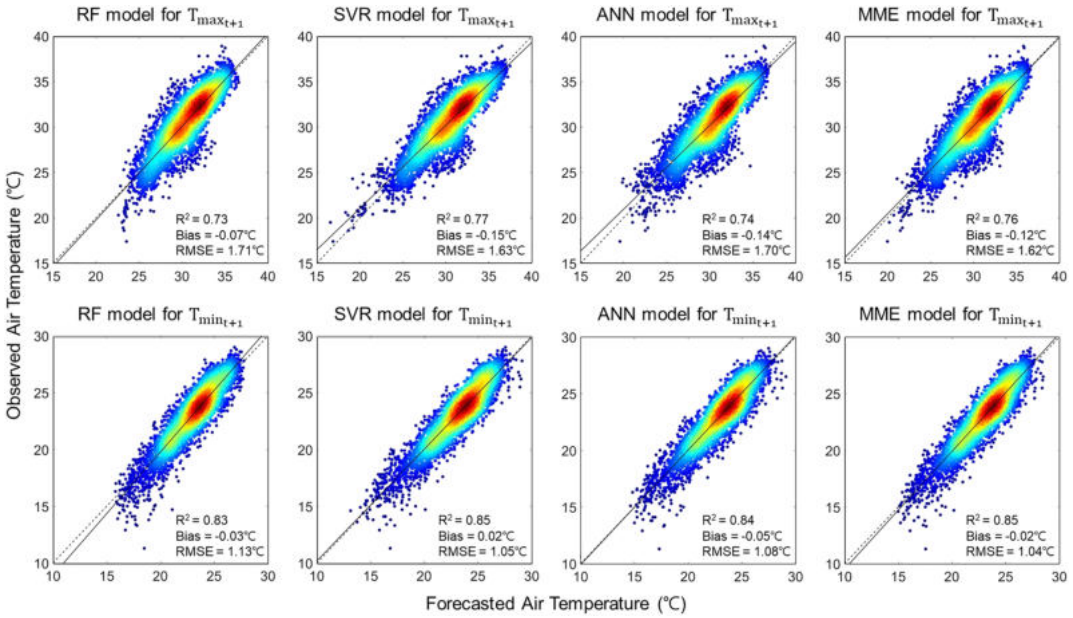

Figure 7. Scatter plots between forecasted and observed air temperatures from leave-one-station-out cross-validation results based on RF, SVR, ANN, and MME models for $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecasts.

To produce the spatially distributed maps of $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ from machine learning-based bias correction models, $\mathbf{T_{max_t}}$ and $\mathbf{T_{min_t}}$ of *in-situ* observations were interpolated using a cokriging technique with elevation, which is useful for predictions (Baskan et al., 2009). Figures 9 and 10 show the $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecast maps of the LDAPS model and all bias correction models which averaged across days with non-missing *in-situ* observations at all stations (i.e., 135 days) during July and August from 2015 to 2017.

$\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecast maps of the LDAPS model showed low temperatures in the northern part of Seoul around stations 4 and 20 wider than actual high elevation area (Figures 9(a), 10(a)), which may have

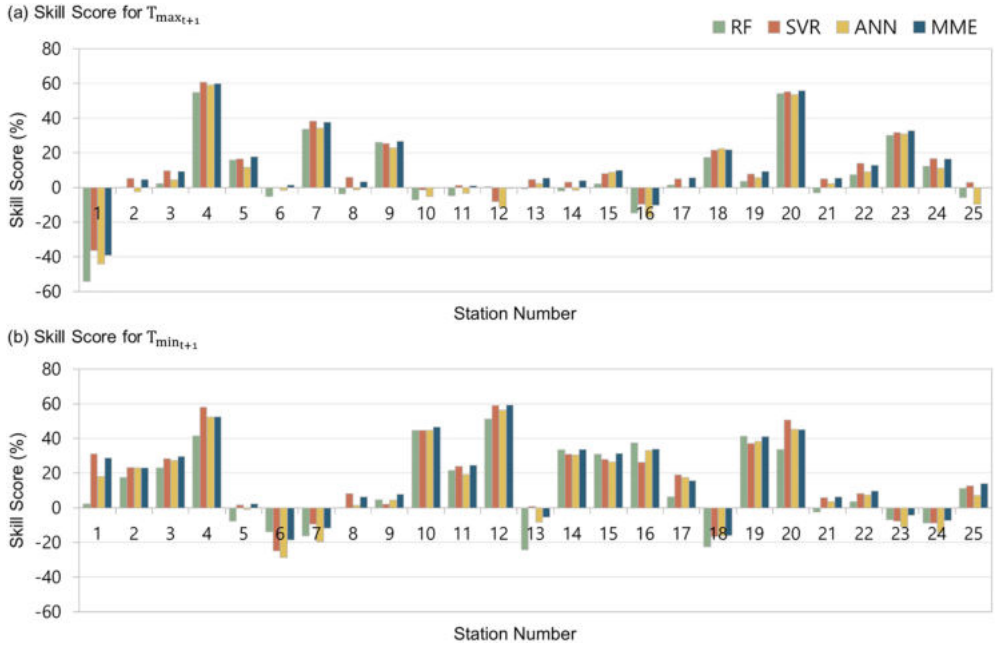

Figure 8. Skill scores from leave-one-station-out cross-validation results based on RF, SVR, ANN, and MME models at the 25 stations for (a) $\mathbf{T_{max_{t+1}}}$ and (b) $\mathbf{T_{min_{t+1}}}$ forecasts.

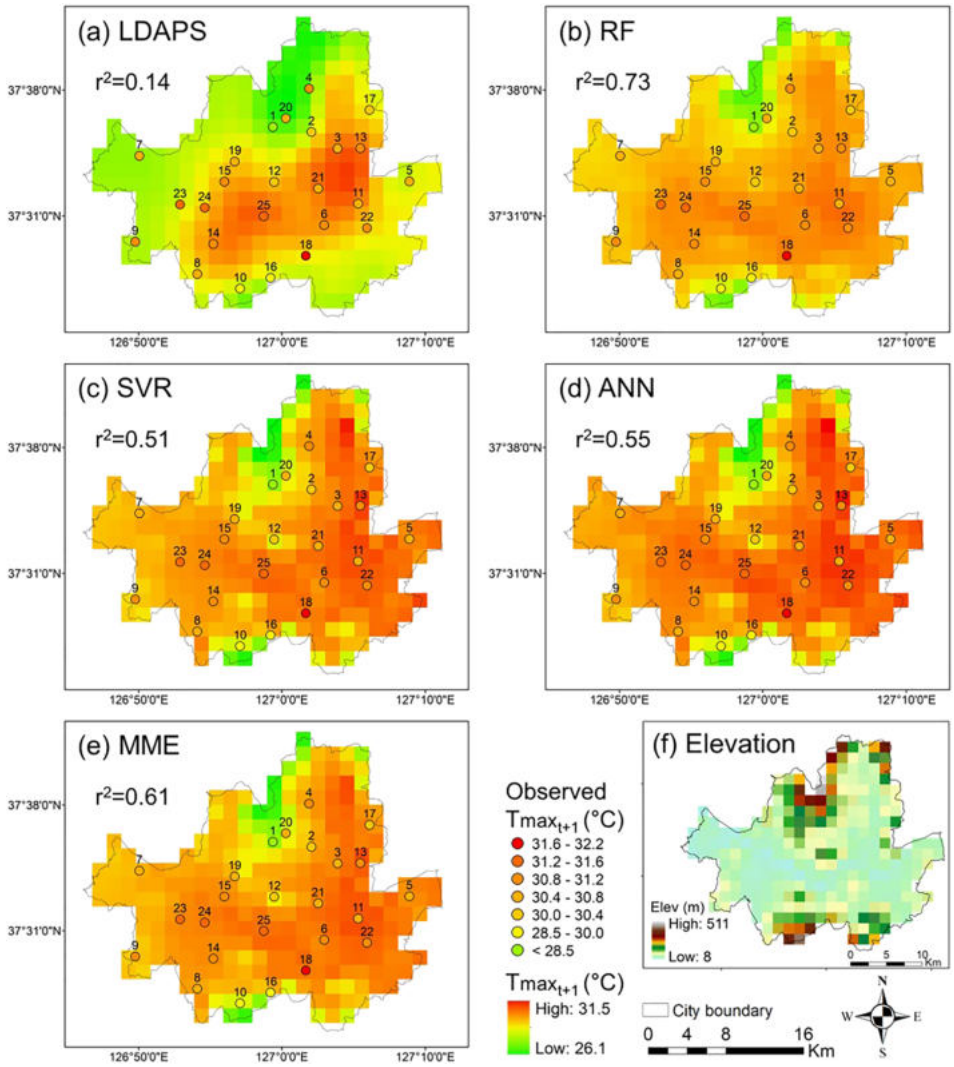

Figure 9. Map of spatial distribution of average forecasted $\mathbf{T_{max_{t+1}}}$ based on (a) LDAPS, (b) RF, (c) SVR, (d) ANN and (e) MME models, and (f) SRTM elevation aggregated to 1.5 km resolution.

been affected by the smoothed elevation surface used as the initial condition of the LDAPS model (Hart et al., 2004; Libonati et al., 2008). All bias correction models were similar in terms of the spatial patterns of $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecast maps (Figures 9(b)-e, 10(b)-(e)). The corrected temperature distributions were generally higher than the temperature distribution of the LDAPS model for $\mathbf{T_{max_{t+1}}}$ forecast, while lower than that of the LDAPS model for $\mathbf{T_{min_{t+1}}}$ forecast. This is because the models corrected cold and warm biases of the LDAPS model for $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecasts, respectively. All bias correction models also showed relatively low temperatures in high elevation areas when compared to low elevation areas, having a similar pattern by elevation. This result implies that the bias correction models clearly reflected the elevation which is one of the main factors affecting temperature.

However, the maximum values of $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecast maps of all bias correction models were lower than the maximum values of the *in-situ* observations. This may be because the $\mathbf{T_{max_t}}$ and $\mathbf{T_{min_t}}$ of the *in-situ* observations were calculated using the cokriging technique, which produces more plausible interpolation fields than the other techniques but does not perform well with respect to typical validation indices such as the mean square error (Appelhans et al., 2015; Collins, 1995). Moreover, the RF model had narrower temperature ranges than the other bias correction models for $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecasts. This might be because

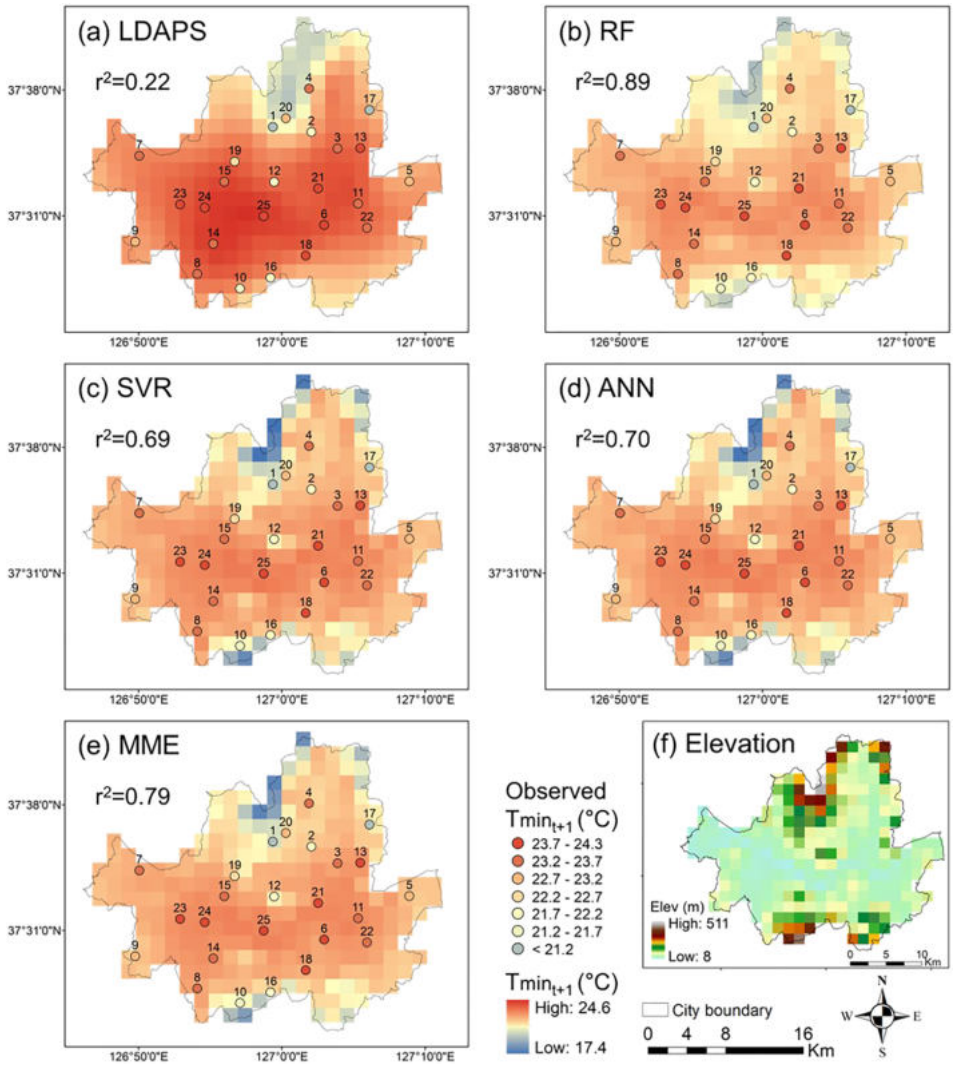

Figure 10. Same as Figure 9 but for $\mathbf{T_{min_{t+1}}}$ forecast.

RF cannot extrapolate temperature values beyond the range of the observed temperatures in the training data, while the SVR and ANN models can (Bramer, 2006; Crone et al., 2006; Shah et al., 2014).

4.4. Effect of Local Characteristics

The MME model used all station data with the same input variables. The importance of input variables for bias correction might vary by location, and thus the locality could be a crucial element of daily temperature fluctuations. In particular, inaccurate simulation of elevations of LDAPS may be a major cause of different biases for each station. With this in mind, the LDAPS-derived temperature forecast data were first corrected using lapse rates with elevation in this study. After this process, the accuracy of LDAPS model forecasting for $\mathbf{T_{max_{t+1}}}$ increased resulting in R^2 of 0.73 and RMSE of 1.91 °C when compared to the original data described in Figure 3(a) (i.e., R^2 of 0.69 and RMSE of 2.08 °C). Similarly, the accuracy of LDAPS model forecasting for $\mathbf{T_{min_{t+1}}}$ also increased resulting in R^2 of 0.82 and RMSE of 1.28 °C (from R^2 of 0.77 and RMSE of 1.43 °C in Figure 4(a)). It is widely known that air temperature of a day (i.e., $\mathbf{T_{max_t}}$ and $\mathbf{T_{min_t}}$) shows a very high correlation with air temperature of the next day (i.e., $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$) for each station (Ustaoglu et al., 2008). In addition, the LDAPS model also incorporates local characteristics such as elevation and land cover, which can be reflected in the forecast data. These suggest that locality characteristics could be included in the LDAPS model forecast and *in-situ* data, which were used as input variables in this study.

Table 4
Hindcast Validation Results of MME Model Runs Without Location and Topographic Variables for $\mathbf{T_{max_{t+1}}}$ *and* $\mathbf{T_{min_{t+1}}}$ *Forecasts*

MME model run without auxiliary variables, for $\mathbf{T_{max_{t+1}}}$			
	Case 1	Case 2	Case 3
R^2	0.78	0.78	0.78
RMSE (°C)	1.56	1.56	1.57
MME model run without auxiliary variables, for $\mathbf{T_{min_{t+1}}}$			
	Case 1	Case 2	Case 3
R^2	0.87	0.86	0.86
RMSE (°C)	0.99	1	1.03

Note. Case 1 excluded location variables (i.e., Lat and Lon); case 2 excluded topographic variables (i.e., Elev, Slop, and Sol); and case 3 excluded all five variables.

We further tested the impact of the five location and topographic variables (i.e., latitude and longitude, elevation, slope, and solar radiation) to the MME model by selectively excluding them from input variables (Table 4). When compared to the hindcast validation results of the MME model in Figures 3(e) and 4(e) for $\mathbf{T_{max_{t+1}}}$, all three cases (cases 1–3) produced similar performance with a slightly increased RMSE (0.01–0.02 °C). Similarly, a slightly decreased R^2 (~0.01) and increased RMSE (0.01–0.05 °C) were achieved by three case models for $\mathbf{T_{min_{t+1}}}$ when compared to the MME model developed using all variables. However, the decrease of accuracy for both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ was not large, which implies that using additional local characteristics as input variables to machine learning might not be always required for air temperature forecasting at a city level (e.g., Seoul in this study). As mentioned before, the LDAPS model forecast and *in-situ* data that were used as input variables in this study already contain local characteristics to some extent. However, the use of additional local variables is likely beneficial to bias correction of air temperature for a large area (i.e., country to continental scale) as the spatial variation of air temperature is quite large when compared to a city.

4.5. Novelty and Limitations

Most previous studies that have used machine learning approaches to correct the bias of the NWP model temperature forecasts only applied single machine learning methods (Marzban, 2003; Yi et al., 2018; Zjavka, 2016). In contrast, this study utilized three machine learning models (i.e., RF, SVR and ANN) and their ensemble (MME) for the bias correction of the LDAPS model's $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecasts. We found that the MME model with an ensemble of three single machine learning showed a better generalization performance than the single machine learning models for forecasting both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$.

In addition, many previous studies that implemented various statistical techniques (i.e., MOS and KF) to correct the NWP model temperature forecasts compared the models by performing bias corrections for each station, which cannot produce the spatial distribution maps of air temperature (Isaksson, 2018; Vashani et al., 2010). However, machine learning methods that use all stations can produce the spatial distribution of air temperature. Nevertheless, few researchers have conducted comparative bias corrected model studies with NWP models through spatial distribution as well as by station. In this study, we examined the performances of the bias correction models by conducting two validations (i.e., hindcast validation and LOSOCV), and compared the results with the LDAPS model in detail, not only by station but also based on the spatial distribution.

This study still has some limitations. The bias of the LDAPS model was corrected based on machine learning methods without considering the fact that the LDAPS model was being updated throughout the study period. In order to further improve the forecast accuracy of the NWP model, it is thus necessary to consider the update of the NWP model in future research. The cokriging technique was applied to the $\mathbf{T_{max_t}}$ and $\mathbf{T_{min_t}}$ of *in-situ* observations to produce the spatial maps of the bias corrected $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ forecasts. However, not only does cokriging require regularly spaced input data (Hulme et al., 1995; Price et al., 2000), but also does not perform well with respect to typical validation indices, such as the mean square error. Appelhans et al. (2015) evaluated cokriging and machine learning approaches for interpolating the monthly mean air temperature with geographical variables (i.e., DEM, slope, aspect and sky-view factor) and Normalized Difference Vegetation Index (NDVI) at Mt. Kilimanjaro, Tanzania. They found that machine learning-based interpolation, which does not require regularly spaced input data, could produce more reliable spatial estimates than cokriging. However, monthly data was used in their study, and the study area was not an urban area. Thus, daily air temperature interpolation based on machine learning in urban areas deserves future exploration, which might further improve the forecasting skills of the air temperature bias correction models. Another solution is to use air temperatures estimated from remote sensing-based data, such as land surface temperatures (Noi et al., 2017; Salcedo-Sanz et al., 2016; Shamshirband et al., 2015; Xu et al., 2014), although it requires careful gap-filling over cloud-contaminated areas.

5. Conclusion

In this study, we evaluated the bias correction performance of three machine learning methods and their ensemble for improving the LDAPS model outputs of $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$ in Seoul Metropolitan Area. The bias correction models were developed by fusing a total of 14 LDAPS model forecast data, $\mathbf{T_{max_t}}$ and $\mathbf{T_{min_t}}$ of *in-situ* observations, and five auxiliary data as input variables. Hindcast validation and LOSOCV were conducted to evaluate the four machine learning approaches and the LDAPS model.

When forecasting $\mathbf{T_{max_{t+1}}}$, the LDAPS model had an R^2 of 0.69, a bias of −0.85 °C and an RMSE of 2.08 °C, whereas according to hindcast validation all bias correction models improved performance with R^2 ranging from 0.75 to 0.78, biases from −0.16 to −0.07 °C and RMSEs from 1.55 to 1.66 °C. When forecasting $\mathbf{T_{min_{t+1}}}$, the LDAPS model had an R^2 of 0.77, a bias of 0.51 °C and an RMSE of 1.43 °C, whereas the bias correction models showed R^2 from 0.86 to 0.87, biases from −0.03 to 0.03 °C and RMSEs from 0.98 to 1.02 °C. LOSOCV for the four machine learning approaches also revealed corresponding improvements when compared to the NWP model for both $\mathbf{T_{max_{t+1}}}$ and $\mathbf{T_{min_{t+1}}}$. In particular, the MME model, the ensemble method of three different machine learning approaches, demonstrated its strength by producing more stable and accurate results compared to single machine learning models in terms of temporal and spatial aspects. We found that the weather stations where the LDAPS model had a large elevation bias showed a distinct improvement in performance after bias correction, which helps the spatial distribution of air temperature derived from the machine learning models be much more similar to that of *in-situ* observed air temperatures. Despite the necessity for more investigation with other NWP models, this approach is likely to be successful if applied to other NWP models for the study area that can predict temperatures deterministically over the next-day. This is because we confirmed that the LDAPS model has cold and warm biases for $T_{max_{t+1}}$ and $T_{min_{t+1}}$ forecasts, respectively, and the MME model generally reduced these biases. This study used a simple averaging technique for the MME model, but it is expected that a more sophisticated ensemble technique (i.e. weighted) can be utilized for operational purposes in the future.

Acknowledgments
This research was supported by the Space Technology Development Program and the Basic Science Research Program through the National Foundation of Korea (NRF) funded by the Ministry of Science, ICT, & Future Planning and the Ministry of Education of Korea, respectively (Grant: NRF-2017M1A3A3A02015981; NRF-2017R1D1A1B03028129), the Korea Meteorological Administration Research and Development Program under Grant KMIPA 2017-7010, and Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (NRF-2016M3C4A7952600). CY was also supported by Global PhD Fellowship Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2018H1A2A1062207). Data used in this study will be freely available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast).

References

Adeodato, P. J., Arnaud, A. L., Vasconcelos, G. C., Cunha, R. C., & Monteiro, D. S. (2011). MLP ensembles improve long term prediction accuracy over single networks. *International Journal of Forecasting*, *27*(3), 661–671.

Anadranistakis, M., Lagouvardos, K., Kotroni, V., & Elefteriadis, H. (2004). Correcting temperature and humidity forecasts using Kalman filtering: Potential for agricultural protection in northern Greece. *Atmospheric Research*, *71*(3), 115–125.

Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., & Nauss, T. (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spatial Statistics*, *14*, 91–113.

Aznar, J. C., Gloaguen, E., Tapsoba, D., Hachem, S., Caya, D., & Bégin, Y. (2013). Interpolation of monthly mean temperatures using cokriging in spherical coordinates. *International Journal of Climatology*, *33*(3), 758–769.

Balabin, R. M., & Smirnov, S. V. (2012). Interpolation and extrapolation problems of multivariate regression in analytical chemistry: Benchmarking the robustness on near-infrared (NIR) spectroscopy data. *Analyst*, *137*(7), 1604–1610. https://doi.org/10.1039/c2an15972d

Baskan, O., Erpul, G., & Dengiz, O. (2009). Comparing the efficiency of ordinary kriging and cokriging to estimate the Atterberg limits spatially using some soil physical properties. *Clay Minerals*, *44*(2), 181–193.

Boardman, M., & Trappenberg, T. (2006). A heuristic for free parameter optimization with support vector machines. Paper presented at the the 2006 IEEE international joint conference on neural network proceedings.

Bosart, L. F., Sprigg, W. A., & Council, N. R (1998). *The Meteorological Buoy and Coastal Marine Automated Network for the United States*. Washington, DC: National Academies Press.

Bramer, M. (2006). *Artificial intelligence in theory and practice: IFIP 19th world computer congress, TC 12: IFIP AI 2006 stream, august 21-24, 2006, Santiago, Chile* (Vol. 217). New York: Springer Science & Business Media.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chou, J.-S., & Pham, A.-D. (2013). Enhanced artificial intelligence for ensemble approach to predicting high performance concrete compressive strength. *Construction and Building Materials*, *49*, 554–563.

Collins, F. C. (1995). A Comparison of Spatial Interpolation Techniques in Temperature Estimation. Virginia Tech,

Crone, S. F., Guajardo, J., & Weber, R. (2006). A Study on the Ability of Support Vector Regression and Neural Networks to Forecast Basic Time Series Patterns. Paper presented at the IFIP International Conference on Artificial Intelligence in Theory and Practice.

Das, S. P., & Padhy, S. (2018). A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. *International Journal of Machine Learning and Cybernetics*, *9*(1), 97–111.

de Carvalho, J. R. P., Assad, E. D., & Pinto, H. S. (2011). Kalman filter and correction of the temperatures estimated by PRECIS model. *Atmospheric Research*, *102*(1–2), 218–226.

Eccel, E., Ghielmi, L., Granitto, P., Barbiero, R., Grazzini, F., & Cesari, D. (2007). Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models. *Nonlinear Processes in Geophysics*, *14*(3), 211–222.

Flake, G. W., & Lawrence, S. (2002). Efficient SVM regression training with SMO. *Machine Learning*, *46*(1–3), 271–290.

Forkuor, G., Dimobe, K., Serme, I., & Tondoh, J. (2018). Landsat-8 vs. Sentinel-2: Examining the added value of sentinel-2's red-edge bands to land-use and land-cover mapping in Burkina Faso. *GIScience and Remote Sensing*, *55*, 331–354.

Gorissen, D., Dhaene, T., & Turck, F. D. (2009). Evolutionary model type selection for global surrogate modeling. *Journal of Machine Learning Research*, *10*(Sep), 2039–2078.
Han, L., Sun, J., Zhang, W., Xiu, Y., Feng, H., & Lin, Y. (2017). A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, *122*, 4038–4051. https://doi.org/10.1002/2016JD025783
Hart, K. A., Steenburgh, W. J., Onton, D. J., & Siffert, A. J. (2004). An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and Paralympic winter games. *Weather and Forecasting*, *19*(2), 200–218.
Healey, S. P., Cohen, W. B., Yang, Z., Kenneth Brewer, C., Brooks, E. B., Gorelick, N., et al. (2018). Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment*, *204*, 717–728. https://doi.org/10.1016/j.rse.2017.09.029
Ho, H. C., Knudby, A., Sirovyak, P., Xu, Y., Hodul, M., & Henderson, S. B. (2014). Mapping maximum urban air temperature on hot summer days. *Remote Sensing of Environment*, *154*, 38–45.
Horning, N. (2013). Introduction to decision trees and random forests. *American Museum of Natural History*, *2*, 1–27.
Hulme, M., Conway, D., Jones, P., Jiang, T., Barrow, E., & Turney, C. (1995). Construction of a 1961–1990 European climatology for climate change modelling and impact applications. *International Journal of Climatology*, *15*(12), 1333–1363.
Im, J., Park, S., Rhee, J., Baik, J., & Choi, M. (2016). Downscaling of AMSR-E soil moisture with MODIS products using machine learning approaches. *Environmental Earth Sciences*, *75*(15), 1120.
Isaksson, R. (2018). Reduction of Temperature Forecast Errors with Deep Neural Networks. Uppsala University, LUVAL. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-350264
Ishida, T., & Kawashima, S. (1993). Use of cokriging to estimate surface air temperature from elevation. *Theoretical and Applied Climatology*, *47*(3), 147–157.
Kecman, V. (2001). *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. Cambridge: MIT Press.
Kim, M., Im, J., Han, H., Kim, J., Lee, S., Shin, M., & Kim, H. (2015). Landfast Sea ice monitoring using multisensory fusion in the Antarctic. *GIScience and Remote Sensing*, *52*, 239–256.
Kim, M., Im, J., Park, H., Park, S., Lee, M.-I., & Ahn, M.-H. (2017). Detection of tropical overshooting cloud tops using himawari-8 imagery. *Remote Sensing*, *9*(7), 685.
Klinenberg, E. (2015). *Heat Wave: A Social Autopsy of Disaster in Chicago*. Chicago: University of Chicago Press.
Kühnlein, M., Appelhans, T., Thies, B., & Nauss, T. (2014). Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. *Remote Sensing of Environment*, *141*, 129–143.
Lee, A., Geem, Z., & Suh, K.-D. (2016). Determination of optimal initial weights of an artificial neural network by using the harmony search algorithm: Application to breakwater armor stones. *Applied Sciences*, *6*(6), 164.
Lee, J., Im, J., Kim, K., & Quackenbush, L. (2018). Machine learning approaches for estimating Forest stand height using plot-based observations and airborne LiDAR data. *Forests*, *9*(5), 268.
Lee, S., Han, H., Im, J., Jang, E., & Lee, M.-I. (2017). Detection of deterministic and probabilistic convection initiation using Himawari-8 advanced Himawari imager data. *Atmospheric Measurement Techniques*, *10*(5), 1859–1874.
Lee, S., Im, J., Kim, J., Kim, M., Shin, M., Kim, H., & Quackenbush, L. (2016). Arctic Sea ice thickness estimation from CryoSat-2 satellite data using machine learning-based lead detection. *Remote Sensing*, *8*, 698.
Li, Y., Zhao, C., Zhang, T., Wang, W., Duan, H., Liu, Y., et al. (2018). Impacts of land-use data on the simulation of surface air temperature in Northwest China. *Journal of Meteorological Research*, *32*(6), 896–908. https://doi.org/10.1007/s13351-018-7151-5
Libonati, R., Trigo, I., & DaCamara, C. C. (2008). Correction of 2 m-temperature forecasts using Kalman filtering technique. *Atmospheric Research*, *87*(2), 183–197.
Liu, T., Abd-Elrahman, A., Morton, J., & Wilhelm, V. L. (2018). Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system. *GIScience & Remote Sensing*, *55*(2), 243–264.
Liu, T., Im, J., & Quackenbush, L. (2015). A novel transferable individual tree crown delineation model based on fishing net dragging and boundary classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *110*, 34–47.
Mansaray, L., Yang, L., Kabba, V., Kanu, A., Huang, J., & Wang, F. (2019). Optimising rice mapping in cloud-prone environments by combining quad-source optical with sentinel-1A microwave satellite imagery. *GIScience and Remote Sensing*, *56*, 1333–1354.
Marzban, C. (2003). Neural networks for postprocessing model output: ARPS. *Monthly Weather Review*, *131*(6), 1103–1111.
Moghim, S., & Bras, R. L. (2017). Bias correction of climate modeled temperature and precipitation using artificial neural networks. *Journal of Hydrometeorology*, *18*(7), 1867–1884.
Nguyen, M. Q., Atkinson, P. M., & Lewis, H. G. (2006). Superresolution mapping using a Hopfield neural network with fused images. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(3), 736–749.
Noi, P., Degener, J., & Kappas, M. (2017). Comparison of multiple linear regression, cubist regression, and random forest algorithms to estimate daily air surface temperature from dynamic combinations of MODIS LST data. *Remote Sensing*, *9*(5), 398.
Omrani, H., Tayyebi, A., & Pijanowski, B. (2017). Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based land transformation model: An integrated ML-CA-LTM modeling framework. *GIScience & Remote Sensing*, *54*(3), 283–304.
Orr, A., Phillips, T., Webster, S., Elvidge, A., Weeks, M., Hosking, S., & Turner, J. (2014). Met Office unified model high-resolution simulations of a strong wind event in Antarctica. *Quarterly Journal of the Royal Meteorological Society*, *140*(684), 2287–2297.
Özçelik, R., Diamantopoulou, M. J., Brooks, J. R., & Wiant, H. V. Jr. (2010). Estimating tree bole volume using artificial neural network models for four species in Turkey. *Journal of Environmental Management*, *91*(3), 742–753. https://doi.org/10.1016/j.jenvman.2009.10.002
Park, S., Im, J., Park, S., Yoo, C., Han, H., & Rhee, J. (2018). Classification and mapping of paddy rice by combining Landsat and SAR time series data. *Remote Sensing*, *10*(3), 447.
Perkins, S., Alexander, L., & Nairn, J. (2012). Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters*, *39*, L20714. https://doi.org/10.1029/2012GL053361
Pham, T., Yoshino, K., & Bui, D. (2017). Biomass estimation of *Sonneratia caseolaris* (*l.*) Engler at a coastal area of Hai Phong city (Vietnam) using ALOS-2 PALSAR imagery and GIS-based multi-layer perceptron neural networks. *GIScience and Remote Sensing*, *54*, 329–353.
Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Retrieved from Technical report:
Price, D. T., McKenney, D. W., Nalder, I. A., Hutchinson, M. F., & Kesteven, J. L. (2000). A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agricultural and Forest Meteorology*, *101*(2–3), 81–94.

Rakkiyappan, R., & Balasubramaniam, P. (2008). Delay-dependent asymptotic stability for stochastic delayed recurrent neural networks with time varying delays. *Applied Mathematics and Computation*, *198*(2), 526–533.
Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, *11*(1), 41–53.
Richardson, H., Hill, D., Denesiuk, D., & Fraser, L. (2017). A comparison of geographic datasets and field measurements to model soil carbon using random forests and stepwise regressions (British Columbia, Canada). *GIScience and Remote Sensing*, *54*, 573–591.
Russo, S., Sillmann, J., Sippel, S., Barcikowska, M. J., Ghisetti, C., Smid, M., & O'Neill, B. (2019). Half a degree and rapid socioeconomic development matter for heatwave risk. *Nature Communications*, *10*(1), 136. https://doi.org/10.1038/s41467-018-08070-4
Salcedo-Sanz, S., Deo, R., Carro-Calvo, L., & Saavedra-Moreno, B. (2016). Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. *Theoretical and Applied Climatology*, *125*(1–2), 13–25.
Schulze, H., & Langenberg, H. (2014). Climate science: Urban heat. *Nature Geoscience*, *7*(8), 553.
Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, *179*(6), 764–774. https://doi.org/10.1093/aje/kwt312
Shamshirband, S., Mohammadi, K., Chen, H.-L., Samy, G. N., Petković, D., & Ma, C. (2015). Daily global solar radiation prediction from air temperatures using kernel extreme learning machine: A case study for Iran. *Journal of Atmospheric and Solar-Terrestrial Physics*, *134*, 109–117.
Shao, Y., & Lunetta, R. S. (2012). Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. *ISPRS Journal of Photogrammetry and Remote Sensing*, *70*, 78–87.
Sim, S., Im, J., Park, S., Park, H., Ahn, M., & Chan, P. W. (2018). Icing detection over East Asia from geostationary satellite data using machine learning approaches. *Remote Sensing*, *10*(4), 631.
Sonobe, R., Yamaya, Y., Tani, H., Wang, X., Kobayashi, N., & Mochizuki, K. (2017). Assessing the suitability of data from sentinel-1A and 2A for crop classification. *GIScience and Remote Sensing*, *54*, 918–938.
Stensrud, D. J., & Yussouf, N. (2003). Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Monthly Weather Review*, *131*(10), 2510–2524.
Tiryaki, S., & Aydın, A. (2014). An artificial neural network model for predicting compression strength of heat treated woods and comparison with a multiple linear regression model. *Construction and Building Materials*, *62*, 102–108.
Ustaoglu, B., Cigizoglu, H. K., & Karaca, M. (2008). Forecast of daily mean, maximum and minimum temperature time series by three artificial neural network methods. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *15*(4), 431–445.
van Wezel, M., & Potharst, R. (2007). Improved customer choice predictions using ensemble methods. *European Journal of Operational Research*, *181*(1), 436–452.
Vashani, S., Azadi, M., & Hajjam, S. (2010). Comparative evaluation of different post processing methods for numerical prediction of temperature forecasts over Iran. *Research Journal of Environmental Sciences*, *4*(3), 305–316.
Wallace, J. M., Tibaldi, S., & Simmons, A. J. (1983). Reduction of systematic forecast errors in the ECMWF model through the introduction of an envelope orography. *Quarterly Journal of the Royal Meteorological Society*, *109*(462), 683–717.
Wang, J., Wu, X., & Zhang, C. (2005). Support vector machines based on K-means clustering for real-time business intelligence systems. *International Journal of Business Intelligence and Data Mining*, *1*(1), 54–64.
Webster, S., Brown, A., Cameron, D., & Jones, C. (2003). Improvements to the representation of orography in the met Office unified model. *Quarterly Journal of the Royal Meteorological Society*, *129*(591), 1989–2010.
Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Cambridge: Academic Press.
Wylie, B. K., Pastick, N. J., Picotte, J. J., & Deering, C. A. (2019). Geospatial data mining for digital raster mapping. *GIScience & Remote Sensing*, *56*(3), 406–429.
Xu, Y., Knudby, A., & Ho, H. C. (2014). Estimating daily maximum air temperature from MODIS in British Columbia, Canada. *International Journal of Remote Sensing*, *35*(24), 8108–8121.
Yang, J., Guo, A., Li, Y., Zhang, Y., & Li, X. (2019). Simulation of landscape spatial layout evolution in rural-urban fringe areas: A case study of Ganjingzi District. *GIScience & Remote Sensing*, *56*(3), 388–405.
Yi, C., Shin, Y., & Roh, J.-W. (2018). Development of an urban high-resolution air temperature forecast system for local weather information services based on statistical downscaling. *Atmosphere*, *9*(5), 164.
Yoo, C., Han, D., Im, J., & Benjamin, B. (2019). Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *157*, 155–170.
Yoo, C., Im, J., Park, S., & Quackenbush, L. J. (2018). Estimation of daily maximum and minimum air temperatures in urban landscapes using MODIS time series satellite data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *137*, 149–162.
Yoo, S., Im, J., & Wagner, J. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*, 293–306.
Yun, J. I., Choi, J.-Y., & Ahn, J.-H. (2001). Seasonal trend of elevation effect on daily air temperature in Korea. *Korean Journal of Agricultural and Forest Meteorology*, *3*(2), 96–104.
Zhang, J., Cha, D.-H., & Lee, D.-K. (2009). Investigating the role of MODIS leaf area index and vegetation-climate interaction in regional climate simulations over Asia. *Terrestrial, Atmospheric and Oceanic Sciences*, *20*(2). https://doi.org/10.3319/TAO.2008.04.03.01(A)
Zheng, W., Ek, M., Mitchell, K., Wei, H., & Meng, J. (2017). Improving the stable surface layer in the NCEP global forecast system. *Monthly Weather Review*, *145*(10), 3969–3987.
Zjavka, L. (2016). Numerical weather prediction revisions using the locally trained differential polynomial network. *Expert Systems with Applications*, *44*, 265–274.