Version of Record: https://www.sciencedirect.com/science/article/pii/S2212095521001115
Manuscript_b48dc8e6d684af81acce1c8241d27acc

The Controlling Factors of Urban Heat in Bengaluru, India

Heather S. Sussman[a], Aiguo Dai[a], and Paul E. Roundy[a]

[a]Department of Atmospheric and Environmental Sciences
University at Albany, State University of New York
1400 Washington Ave. Albany, NY 12222, U.S.A.

Corresponding author: Heather S. Sussman (Email: hsussman@albany.edu)

Manuscript submitted to *Urban Climate*
Initial submission: 18 February 2021
First Revision: 3 May 2021

Abstract

Urbanization can induce land cover changes that impact land surface temperature (LST). Many factors can influence the magnitude of urban heat, such as vegetation and aerosols. This work uses linear correlation, composite analysis, multiple linear regression, and random forest to determine the leading controls on urban LST of Bengaluru, India in the dry and wet seasons during daytime and nighttime during 2003–2018 using data from the MODerate Resolution Imaging Spectroradiometer and the European Centre for Medium-Range Forecasts ERA5 reanalysis. Results show that for the dry and wet season daytime, vegetation was the leading factor (linear correlation R = –0.74 and R = –0.34 with urban LST) since reduced vegetation limits evaporative cooling. For the dry season nighttime, vegetation was the leading factor (R = –0.52). Limited evaporative cooling during daytime can increase surface heat retention at night. For the wet season nighttime, specific humidity was the leading factor (R = 0.21) since increased water vapor enhances downward longwave radiation and warms the surface. Therefore, urban heat is primarily controlled by vegetation in Bengaluru. However, since vegetation and specific humidity are related, mitigation strategies that increase vegetation must not increase water vapor substantially, otherwise urban heat may amplify during the wet season nighttime.

Keywords: Bengaluru, India, multiple linear regression, random forest, urban heat island (UHI), urbanization

1. Introduction

More than half of the world's population lives in cities and this urban population continues to increase (Kim and Baik, 2005; Grimm et al., 2008). While urbanization may have positive impacts, such as allowing for cities to be cultural and economic hubs, it can also negatively affect the natural environment. Urbanization can deteriorate air quality due to more concentrated emissions from manufacturing and vehicular traffic (Ramachandran et al., 2012). The local and regional climate can also be impacted by urbanization. For example, Kishtawal et al. (2010) found urbanization to increase the frequency of heavy rainfall over cities in India during the monsoon season. Additionally, urbanization can reduce natural vegetation, which can impact the ability of residents to see and enjoy nature (Andersson, 2006). Reduced vegetation and increased buildings decrease latent heat fluxes and increase surface roughness, which impact the planetary boundary layer (PBL) (Garratt, 1994). As vegetation decreases, surface temperatures can rise under daytime solar heating due to lack of transpiration, leading to higher Bowen ratios. This causes urban surfaces to warm faster than suburban and non-urban areas during the daytime (Taha, 1997), and can contribute to greater heat storage (i.e., higher temperatures) at nighttime (Kim and Baik, 2002). High urban nighttime temperatures can also be amplified by less radiative cooling due to trapping of longwave radiation by urban street canyons (Theeuwes et al., 2014). This observed phenomenon of higher surface temperatures over a city compared to its surroundings is referred to as the urban heat island (UHI) effect. Of all the consequences of urbanization, increased urban heat is of particular concern because it can increase the risk of heat and respiratory related illnesses for a city's inhabitants (Filho et al., 2018).

The magnitude of the UHI effect is often termed the UHI intensity and is measured as the temperature difference between urban and nearby non-urban areas. There are many factors that can influence UHI intensity, such as the season, time of day, local climate, geographic location, amount of vegetation, urban surface properties, population, building material, and anthropogenic factors such as aerosol emissions from factories and vehicles (Kim and Baik, 2005). For example, Peng et al. (2012) analyzed the UHI intensity for 419 cities around the world using land surface temperature (LST) data from 2003–2008 derived from the MODerate Resolution Imaging Spectroradiometer (MODIS) satellite and showed that 64% of the cities had their highest annual-mean UHI intensity in the daytime (mean value of 1.5 K). The same analysis also revealed that UHI intensity is typically highest during the summer daytime (1.9 K), followed by winter daytime (1.1 K), and summer and winter nighttime are similar with a mean of 1.0 K (Peng et al., 2012). Peng et al. (2012) further showed that the magnitudes and the ordering of the strongest to weakest UHI intensity values differ by region of the world. For example, for the 209 Asian cities examined, they found that UHI intensity was highest during summer daytime (1.5 K), followed by winter nighttime (1.2 K), summer nighttime (1.0 K), and lastly winter daytime (0.9 K). Therefore, UHI intensity is observed to vary seasonally, diurnally, and regionally.

Surface properties also impact UHI intensity. As cities continue to build and decrease vegetation, surface evaporative cooling and latent heat fluxes decrease, which increases daytime temperatures, and can amplify the UHI effect (Zhou et al., 2004, 2007; Peng et al., 2012). This warming can be further augmented as PBL mixing and near-surface wind speed, which is strongest during the day, is weakened due to increased urban surface roughness from more buildings

(Garratt, 1994; Miao et al., 2009). Additionally, urban surfaces tend to have a lower surface albedo than vegetation, which can allow for greater absorption of shortwave radiation, and thus warmer temperatures. A related quantity to the latent heat flux is atmospheric water vapor, which can also influence the UHI intensity magnitude. Increased water vapor can enhance downward longwave radiation, allowing the surface to maintain a warmer temperature at night (Dai et al., 1999). Another related quantity is soil moisture, as dry soils can limit evapotranspiration (Dai et al., 1999). Therefore, vegetation, albedo, latent heat, near-surface wind speed, specific humidity, and soil moisture may affect urban heat.

Aerosols are another control on UHI intensity. For example, a high aerosol optical depth (AOD) in Beijing can reduce surface absorption of sunlight by 40–100 Wm^{-2} and decrease urban LST by 1–2 K compared to the city's non-urban surroundings (Jin et al., 2010). Urban LST can decrease due to aerosols' ability to absorb and scatter visible and near-infrared radiation. Typically, more pollutants are concentrated in urban areas than non-urban areas (Tie and Cao, 2009; Kanakidou et al., 2011), and they are usually in the form of black carbon that are generated from the incomplete combustion of vehicular and industrial fuels (Koelmans et al., 2006). Black carbon aerosols are strong absorbers of solar radiation (Jacobson, 2001; Ramachandran and Kedia, 2010). Therefore, black carbon can cool the surface, but warm the atmosphere during daytime (Lacis and Mishchenko, 1995; Cusack et al., 1998; Qian et al., 2003, 2006). This reduction of surface solar heating at daytime due to a high AOD can cancel or exceed the UHI effect and mainly occurs during the dry season when aerosols are not removed as easily by wet deposition (Mitchell et al., 1995). Additionally, this aerosol-induced decrease in surface shortwave radiation is partly compensated by an increase in downward longwave radiation due to the ability of aerosols to scatter longwave radiation (Dufresne et al., 2002). The impact of aerosols may be responsible for extreme variations of UHI intensity. For example, Pandey et al. (2012, 2014) documented a nocturnal urban heat island throughout the year and during the daytime of the monsoon season in New Delhi, India. However, a negative UHI intensity (i.e., urban cool island) was observed for the dry season daytime, and was hypothesized to be due to increased AOD. This example further highlights how UHI intensity can vary seasonally and diurnally, and how unique observations are found when analyzing UHI intensity at the local-scale.

Overall, there are many factors that can influence urban heat and these factors themselves can interact with each other. It is crucial to quantify the relationships between these factors and understand which best explain a city's urban heat so that the UHI formation mechanisms are understood, and proper mitigation and urban planning methods can be developed (Yang et al., 2019, 2020a, 2020b, 2021; Zhang et al., 2017). For instance, if the most important factor is the latent heat flux, it would be wise for urban planners to develop mitigation strategies that increase vegetation to promote more evaporative cooling. However, if water vapor is also important for a given city, a mitigation strategy of increasing vegetation should not increase lower-tropospheric water vapor substantially because water vapor can enhance downward longwave radiation, which could counteract the effort to decrease urban heat. This highlights just one of the many intertwined relationships amongst the possible controlling factors and the items for consideration when developing mitigation strategies.

Some studies have attempted to determine the leading controlling factors of urban heat for a given city through multiple linear regression (MLR) analysis. MLR is used to statistically predict one variable (predictand) from multiple other variables (predictors) when the expected relationship is linear. Assumptions of MLR include that the data are normally distributed with equal variance and that each variable is independent of one another (Vittinghoff et al., 2005). Some studies used MLR to determine the leading controls on UHI intensity. For example, Kim and Baik (2002) analyzed weather station data from Seoul, South Korea and found that the previous-day maximum UHI intensity is the most important factor in determining maximum daytime and nighttime UHI intensity on a given day; Kolokotroni and Giridharan (2008) examined weather station data from London, England and found that surface albedo is the strongest control on day and night UHI intensity; and Zhou et al. (2011) investigated Beijing, China using satellite and weather station data and found relative humidity and AOD to be the most important controls on the maximum nighttime UHI intensity. Some studies have also used MLR to investigate the controlling factors of urban air temperature. For example, Ho et al. (2014) analyzed satellite data over Vancouver, Canada and found LST and incoming solar radiation to be most important in determining peak daytime air temperature in summer, while Makido et al. (2016) examined satellite data over Doha, Qatar and found the most important variable for urban air temperature to be the distance to the coast. Recently, Guo et al. (2020) used MLR and spatial analysis methods (i.e., the spatial lag model and the spatial error model) to investigate explanatory variables in the categories of architectural form, land type, landscape index, social economy, topography, and remote sensing index on controlling urban LST in Dalian City, China. They found variables related to land type, landscape index, and remote sensing index to best explain urban LST. Some of these previous studies have also applied machine learning methods to find the most important statistical associations with UHI intensity or urban temperature (e.g., Kim and Baik, 2002; Zhou et al., 2011; Ho et al., 2014; Makido et al., 2016). All of these studies found the machine learning methods to have higher accuracy compared to the their MLR analyses, likely since the assumptions of MLR may not always be valid since an environmental or climate system can be nonlinear (Smith et al., 2013).

In contrast to MLR, Vittinghoff et al. (2005) stated that modeling using regression trees (i.e., decision trees) with recursive partitioning does not make assumptions about the distribution of the data and the interactions between predictor variables are incorporated into the regression tree model. Recursive partitioning involves the subdivision of a sample into groups that are as similar as possible by minimizing the variance within the group in order to determine a numerical response variable. Overall, regression tree modeling makes nonlinear interactions between predictor variables easier to consider (Hastie et al., 2009; Ismail et al., 2010; Vincenzi et al., 2011), which is a major advantage over MLR. The concept of regression trees was furthered by Breiman et al. (1984) who stated that by using the power of computers, it is possible to generate thousands of regression trees. This idea became known as the random forest (RF) model in which each regression tree is grown from a resampled version of the beginning training dataset and a different random subset of input variables is evaluated for inclusion into the tree at each branch in the growing process (Breiman, 2001). Together, these sources of randomness ensure the independence of each tree. The use of the RF model has greatly enhanced determination of the leading controls and prediction accuracy for situations with multiple contributing factors in many fields, including land cover classification (Pal, 2005; Gislason et al., 2006), ecology (Prasad et al., 2006), remote sensing applications (Ismail and Mutanga, 2010; Mutanga et al., 2012), weather forecasting (McGovern et al., 2014; Williams, 2014; Gagne et al., 2017), among others, but may not necessarily be an improvement in neuroscience (Smith et al., 2013).

While many studies, e.g., Kim and Baik, 2002; Zhou et al., 2011; Ho et al., 2014; Makido et al., 2016, have done novel work using regression and machine learning methods to determine controlling factors of urban heat for Seoul, Beijing, Vancouver, and Doha, individual cities are unique in how the variables might interact and influence urban heat due to their varying magnitudes and importance within the local climate. Furthermore, each city can have a different set of potential controlling factors that may need to be considered. Therefore, the findings from these previous studies may not be applicable to other cities.

One particular city, Bengaluru, India, the third most populous city of India, is a prime example of rapid urbanization. For the time between 2001 to 2011, Bengaluru experienced a 47.18% increase in its population to approximately 10 million (Census of India, 2011). Additionally, Bengaluru was once known as the *"Garden City"* of India, but is now known as the *"Silicon City"* (Sudhira et al., 2007) due to the increased presence of the information technology industry and near-depletion of its natural vegetation. Bengaluru (city center: 12.97°N, 77.59°E) is centrally located in southern India on the Deccan Plateau at an elevation of 900 m. Bengaluru is characterized by a tropical savanna climate (Peel et al., 2007), in which distinct dry and wet seasons are observed. Sussman et al. (2019) recently analyzed the UHI intensity of Bengaluru seasonally and diurnally using MODIS LST data from 2003–2018. Their results showed that the highest mean UHI intensity occurred during the dry season nighttime (1.43 K), followed by the wet season daytime (1.14 K), wet season nighttime (1.02 K), and lastly the dry season daytime (–0.60 K). There are many possible mechanisms for the UHI formation in Bengaluru and the lack of an UHI during the dry season daytime. Sussman et al. (2019) hypothesized that the urban cool island observed during the dry season daytime could be due to the observed increasing trend in AOD, similar to the hypothesis of Pandey et al. (2012, 2014) for New Delhi. However, since there are multiple driving factors that all occur simultaneously, more work needs to be done in evaluating the leading controls of urban heat in Bengaluru so that optimal mitigation strategies can be developed. Since Sussman et al. (2019) found that LST trends from 2003–2018 were mainly concentrated over urban areas (i.e., Figs. 4 and 5 in Sussman et al., 2019), and thus urban LST trends are contributing to trends in UHI intensity for Bengaluru, the goal of this study is to answer the following questions:

1) How strongly related are the potential drivers over the urban surface in Bengaluru?
2) Using the potential drivers over the urban surface, which best determine urban LST in Bengaluru?
3) Using linear correlation, composite analysis, MLR, and the RF algorithms to answer Question 2, how do these methods compare?

2. Data and methods

2.1 Study region

This study focused on a 50 km × 50 km region surrounding the Bengaluru city center, which is the same region analyzed by Sussman et al. (2019). This study region was chosen since it is large enough to capture all of Bengaluru and surrounding non-urban areas, yet small enough to exclude other urbanizing cities within the area.

2.2 Datasets

Table 1 summarizes the datasets used in this study, including their spatiotemporal resolutions and weblink. All data were analyzed for 2003–2018 since MODIS instruments Terra and Aqua have data available since March 2000 and July 2002, thus forcing the analysis to begin in 2003 in order to use data from both instruments. Terra and Aqua obtain data in 36 spectral bands that have a wavelength range from 0.4–14.4 μm and image the entire Earth's surface every 1–2 days. The sun synchronous orbital characteristics of Terra and Aqua have a daytime equatorial crossing time of approximately 10:30 and 13:30 local solar time, and a nighttime equatorial crossing time of approximately 22:30 and 01:30 local solar time. Four variables were obtained from MODIS Collection 6, which include AOD, enhanced vegetation index (EVI), land cover, and LST. For EVI and LST, in which the data are separated by instrument, the Terra and Aqua measurements were averaged. Additionally, LST data is measured at day and night, so by averaging the Terra and Aqua measurements, a daytime average is made at 12:00 local solar time and a nighttime average at 00:00 local solar time. The AOD, EVI, and land cover data were re-sized to a 1 km resolution using nearest neighbor interpolation in order to match the spatial resolution of the LST data. The AOD and EVI data were also converted to 8-day composites to match the temporal resolution of the LST data. This was done since the LST data is the dependent variable for this study. Land cover did not have its temporal resolution changed since land cover is not expected to have as much short-term variability as the other variables. Lastly, the AOD data is measured in two wavelengths, namely 0.47 μm (blue band) and 0.55 μm (green band). Sussman et al. (2019) showed that these AOD measurements are approximately the same in both bands, therefore the wavelengths were averaged into a single measurement of AOD throughout this study.

Albedo, latent heat, soil moisture, and 10-m zonal (u) and meridional (v) wind components were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5-Land reanalysis. This data was obtained and averaged for daytime at 05:00 and 08:00 UTC (10:30 and 13:30 Indian Standard Time; IST) and 17:00 and 20:00 UTC (22:30 and 01:30 IST) for nighttime. This corresponds to the timing of the MODIS LST observations. The soil moisture was measured from the surface (0 cm) to a depth of 7 cm below the surface. The 10-m wind speed was calculated using the vector components. Specific humidity data was obtained from the ECMWF ERA5 reanalysis on pressure levels. Data was obtained for daytime and nighttime in a similar way to the ERA5-Land variables, and from 850–1000 hPa in order to calculate the lower-tropospheric specific humidity since near-surface water vapor should have the greatest influence on surface temperature. All ERA5 variables were converted to 8-day composites in the same way as MODIS.

2.3 Season and urban classification

Sussman et al. (2019) performed their analysis from 2003–2018 for daytime and nighttime averaged over the dry (December-January-February; DJF) and wet (August-September-October; ASO) seasons. Precipitation was found to be at its minimum during DJF and at its maximum during ASO (i.e., Fig. 1 in Sussman et al., 2019). Urban area was distinguished using the MODIS land cover dataset, and urban LST was determined by matching the urban pixels with the LST data, which were on the same spatial resolution. The calculated 8-day composite urban LST values from Sussman et al. (2019) are used in this study as the predictand in the MLR and RF analyses, together with the updated 2018 land cover data that were unavailable to Sussman et al. (2019).

All variables were classified into their urban component prior to analysis. For the MODIS variables of AOD and EVI, urban values were distinguished using the MODIS land cover dataset, similar to how urban LST was determined. To determine the urban ERA5 grids, the percentage of MODIS urban land cover pixels was calculated within each ERA5 grid. If 70% or more of the pixels were urban, the ERA5 grid was classified as urban. This threshold was chosen so that majority of a grid is urban land. If this threshold is increased to 75%, the ERA5 data would have one less urban grid, and if it were increased to 80%, the ERA5 data would have two less urban grids. No further changes occur if the threshold is increased beyond 80%. Repeating the linear correlation analysis (see Section 2.4) for the 75% and 80% thresholds revealed that while the correlation values do change, the main conclusions in terms of which variable has the highest correlation for each season and time of day do not change, likely since majority of the urban grids with the >70% threshold are still used. Therefore, the major conclusions for all analyses done in this study are likely not sensitive to this threshold.

2.4 Analysis

This work assessed the associations of AOD, albedo, EVI, latent heat, soil moisture, near-surface specific humidity, and 10-m wind speed with urban LST. These factors were chosen based on the literature of the key physical drivers of urban heat (e.g., Taha, 1997; Dai et al., 1999; Kim and Baik, 2005; Jin et al., 2010; Peng et al., 2012). These factors were also chosen due to their relevance within Bengaluru. It is important to assess variables related to moisture, vegetation, and aerosols, and how they may be transported by wind given Bengaluru's tropical location that used to be well known for its natural vegetation, but is now characterized by its increasing pollution.

First, to understand the relationships between the prospective controlling factors, which could help explain underlying physical relationships between a factor and urban LST, the linear correlation (R) between each pair of predictor variables was computed for the 8-day composite data. Significance of the correlation was assessed at the 5% and 10% levels using the Student's t-test. Next, a check on whether the prospective controlling factors are inter-correlated (i.e., multicollinearity) was done. Multicollinearity is problematic because it can result in inaccurate estimates of variable importance since an independent variable can be explained by other variables used in the statistical models. The variance inflation factor (VIF) was computed for each pair of the independent variables, which is a measurement of the extent to which an independent variable can be explained by all the other independent variables in the model. In general, if VIF $\geq$ 10, a high degree of multicollinearity is present (Belsley et al., 1980), and thus the greater the VIF, the greater the multicollinearity. A VIF of 1 would indicate no multicollinearity. The VIF of a given variable is determined by:

$$\mathrm{VIF} = \frac{1}{1-\mathrm{R}^2} \qquad \text{Eq. (1)}$$

In Eq. (1), R^2 is the coefficient of determination of the regression of the given variable on all other independent variables (i.e., the fraction of the variance explained by the other variables). If the VIF of a variable exceeds 10, the variable will be removed from the given analysis in order to reduce multicollinearity, starting with the variable that produced the highest VIF. The new VIF for the remaining variables will then be re-computed. This process will continue until all VIF values are less than 10. This is similar to other regression analyses such as Vu et al. (2015) in which the VIF was used to select an appropriate subset of climate variables for prediction of electricity demand.

Using all prospective controlling factors not removed from the multicollinearity check, a linear correlation of the 8-day composite data was calculated between each independent variable and urban LST. A composite analysis of the 8-day composite data followed in which for each independent variable, the values associated with the bottom (≤10th) and top (≥90th) percentiles were derived and the corresponding urban LST at those cases were found. At both thresholds, the urban LST was averaged and the significance between these two composite means of urban LST was assessed by the Student's t-test. The composite mean difference was calculated as the cases at the top percentiles minus the bottom percentiles. This composite analysis was done to understand which of the prospective controlling factors are best associated with urban LST at extreme values. For the above calculations, significance was assessed at the 5% and 10% levels.

Next, the MLR and RF statistical analyses were performed. The prospective controlling factors were first standardized by converting the 8-day composite data into anomalies relative to their 2003–2018 mean and in units of their standard deviation for each season and time of day. The standardized variables were then inputted to the MLR model to compute the standardized regression coefficients, which were used to assess variable importance. Significance of the standardized regression coefficients was tested using the Student's t-test. The variables without standardization were then used to calculate urban LST for each season and time of day in the MLR model. The root mean square error (RMSE) and mean absolute error (MAE) were computed to compare the calculation of urban LST by MLR to observations (i.e., the direct calculation of 8-day composite urban LST). The total variance explained by each MLR model (R^2) was also calculated along with its significance in an effort to evaluate model performance.

After the MLR analysis, the RF was used to assess variable importance. For each RF simulation, 1000 trees were used, in which each tree determined a decision on the value of urban LST. One-thousand trees were chosen in order to guarantee model stabilization (Jiang et al., 2020) and since the mean squared error calculated for the regression trees plateaued around the 900–1000 tree range (not shown). Therefore, adding more than 1000 trees would not necessarily improve the results. Predictor importance was assessed by randomly permuting each predictor value and determining how much it changes the model's prediction (Breiman, 2001). If the change is large, the predictor is likely important. Similar to the MLR analysis, the RMSE, MAE, and R^2 were computed to compare the prediction of urban LST by the RF to observations.

The MLR and RF results were compared for their similarities and differences in terms of their error values and explained variance. Additionally, since a model can become overfit to the data when using multiple predictors by noise in the model predictors fitting to noise in the predictand, and thus a lower error value and higher variance explained can be obtained, a test for overfitting was done. A k-fold cross-validation (Geisser, 1975) procedure was applied to each MLR and RF model for each season and time of day. This process included the following steps: 1) holding back one year's worth of data from the training, 2) developing the models on the other data, 3) using the data from the held back year to test the model's fit, and 4) repeating this process until all years in the dataset have been held back. K-fold cross validation obtains an independently predicted dataset that has no influence of overfitting. The R^2, RMSE, and MAE were averaged over all 16 folds (i.e., 16 years from 2003–2018) and compared to the model that used all the data in fitting. Similar values indicate that the influence of overfitting is likely minimal; however, if the cross-validation values are much lower than the regular model, overfitting is likely present.

For DJF daytime, it is hypothesized that AOD will be the leading control since Sussman et al. (2019) found an urban cool island at this time along with a significant, decreasing trend in LST and a significant, increasing trend in AOD over the city. For ASO daytime, it is hypothesized that EVI will be the leading control since vegetation is most abundant at this time and can therefore impact variables related to moisture and evaporative cooling. For both DJF and ASO nighttime, it is also hypothesized that the leading control is EVI. For DJF, since vegetation is decreasing (Sussman et al., 2019), causing the already dry urban surface to likely have a lesser latent heat flux, this will increase heat retention at night. A similar reasoning exists for ASO; however, the urban surface would not be as dry.

3. Results

3.1 Relationships among the controlling factors

Figure 1 shows the linear correlation matrix between each controlling factor pair measured over the urban surface for each season and time of day. For DJF daytime and nighttime, all variable pairs between latent heat, soil moisture, EVI, and specific humidity have positive correlation coefficients that are significant at the 5% level. These four variables are all related to moisture, i.e., if there is a high amount of vegetation and soil moisture, that will increase the latent heat flux and near-surface water vapor, thus they are all directly related. In contrast, for ASO daytime and nighttime, positive correlations exist between EVI and latent heat, EVI and soil moisture, and EVI and specific humidity that are significant at the 5% level. No significant correlation is found between soil moisture and latent heat, and a negative relationship is found between latent heat and specific humidity that is significant at the 5% level. From the ERA5 data, the 2003–2018 mean values of specific humidity are 9 g kg^{-1} in DJF daytime, 13.7 g kg^{-1} in ASO daytime, 10 g kg^{-1} in DJF nighttime, and 14.2 g kg^{-1} in ASO nighttime. Therefore, since ASO is characterized by greater near-surface moisture than in DJF, evapotranspiration in ASO may be limited by higher air saturation rather than vegetation or soil moisture, thus causing the negative relationship between latent heat and specific humidity.

Another relationship shown in Fig. 1 includes the inverse correlation between EVI and wind speed in ASO daytime (R= –0.13), DJF nighttime (R= –0.36), and ASO nighttime (R= –0.16). These correlations are all significant at the 5% level except in ASO daytime, where it is significant at the 10% level. Reduced urban vegetation is equivalent to increased urban land cover with more buildings, which can increase surface roughness and cause more drag, resulting in slower surface winds. In DJF daytime, the relationship is significant at the 5%, but is positive (R= 0.20). The decreased amount of vegetation, and thereby warmer daytime urban surface, can increase PBL height and cause more vertical mixing and downward momentum transport, thus leading to a stronger 10-m wind speed (Dai and Deser, 1999).

A common relationship that is significant at the 5% level for both seasons and time of day includes the direct relationship between AOD and specific humidity (R= 0.36 in DJF daytime, R= 0.21 in ASO daytime, R= 0.35 in DJF nighttime, and R=0.24 in ASO nighttime). Most urban aerosols are typically in the form of black carbon, which can become hydrophilic in the atmosphere (McMeeking et al., 2011), and thereby increase in volume when a large amount of water vapor is present (Guo et al., 2014). A significant, inverse relationship is observed between AOD and latent heat in ASO daytime and nighttime since high moisture fluxes can induce cloud formation and precipitation, which would cause wet deposition of aerosols. The relationship between AOD and latent heat is not significant in DJF daytime nor nighttime possibly due to less vegetation and drier soils, and thereby a smaller latent heat flux compared to ASO. Another relationship with AOD includes a significant, positive correlation with albedo in ASO daytime and nighttime (R= 0.37 for both), yet a significant, negative correlation with albedo is shown for DJF daytime and nighttime (R= –0.15 for both). Since albedo is measured as the fraction of incident shortwave radiation that the surface reflects and during the dry season aerosols are abundant due to little wet deposition, the amount of shortwave radiation incident at the surface may be less and the longwave radiation reflected by the surface will be low due to a reduction of urban LST. This would result in a lower surface albedo. In contrast, in ASO, when aerosols can be washed out of the atmosphere, more shortwave radiation may be able to reach the surface and the reflection is likely higher than in DJF due to a warmer LST, which would result in a higher surface albedo. Another relationship with albedo occurs with EVI. A negative correlation that is significant at the 5% level is found between these factors in DJF daytime and nighttime (R= –0.75 for both). During DJF, the surface is dry with little vegetation, causing the surface to warm quickly and emit more thermal radiation. This would increase the albedo of the surface given an unchanged amount of incoming solar radiation. Related to this previous relationship, significant, negative correlations exist with albedo and latent heat, soil moisture, and specific humidity in DJF daytime and nighttime since these factors are directly linked to EVI. For ASO daytime and nighttime, a significant, negative relationship also exists between albedo and latent heat, but a significant, positive relationship exists between albedo and specific humidity. This direct relationship between albedo and specific humidity could be due to how increased water vapor can enhance downward longwave radiation, thus warming the surface and increasing thermal radiation emitted, which under a constant amount of incoming solar radiation would increase surface albedo.

For the multicollinearity check, all VIF values calculated were less than 10 (not shown). Therefore, despite several significant correlations found in Fig. 1, multicollinearity appeared to be minimal, and thus none of the prospective controlling factors needed to be removed prior to analysis for factor importance. As a consequence, this alleviates a potential concern that the multicollinearity check could have removed a variable that physically could be important, but statistically was expressed by the other independent variables.

3.2 Importance of the controlling factors

For the linear correlation analysis between each prospective controlling factor and urban LST (Fig. 2a), the highest magnitude correlation was found with EVI in DJF daytime (R= –0.74) and was significant at the 5% level. Therefore, as urban EVI decreases, urban LST increases, likely due to less moisture content to produce a high latent flux that would otherwise cool the surface. Similarly, the other variables related to moisture, i.e., latent heat, soil moisture, and specific humidity, also exhibit negative correlations with urban LST that are significant at the 5% level in DJF daytime. A negative relationship is shown with AOD (R= –0.19) that is significant at the 5% level, indicating that as urban aerosols increase, urban LST decreases, which is likely due to increased absorption of solar radiation by aerosols. A positive correlation is shown with albedo

(R= 0.58) that is significant at the 5% level during DJF daytime, which could be caused by the reflection of solar radiation and emission of terrestrial radiation that are scattered by aerosols, and sent back towards the surface, and thus induces warming. For ASO daytime, the highest magnitude correlations are found with EVI and soil moisture (R= –0.34 for both), which are significant at the 5% level. Similar to DJF daytime, all variables related to moisture have significant, negative correlations for the same physical reasons. A significant, positive correlation was found with albedo (R= 0.19), which is also similar to DJF daytime. For DJF nighttime, the highest magnitude correlation was found with EVI (R= –0.52), which is significant at the 5% level. Even though the latent heat flux is partially driven by solar radiation, and thus is at a minimum at nighttime, if it is low during the day due to low vegetation, the surface can retain more heat at night. Therefore, a significant, negative correlation was also found with latent heat (R= –0.31). Significant, positive correlations were found with albedo (R= 0.42) and specific humidity (R= 0.18) during DJF nighttime. If the aforementioned hypothesis about albedo during the daytime is correct, then this will cause a higher retention of heat at night. For specific humidity, a high amount of water vapor can enhance downward longwave radiation, and thus warm the surface. For ASO nighttime, the highest magnitude correlation was found with specific humidity (R=0.21), which is significant at the 5% level and is likely due to an enhancement of downward longwave radiation by water vapor. A significant, positive correlation was also found with AOD (R= 0.12), which could be due to how aerosols can enhance longwave radiation and warm the atmosphere during daytime as well as that water vapor can increase the volume of aerosols as determined from the relationship shown in Fig. 1. A significant, negative correlation was found with wind speed in ASO nighttime (R= –0.16), indicating that a high urban wind speed can reduce urban LST (i.e., temperature advection) and is likely due to increased urban surface roughness.

For the composite analysis (Fig. 2b), the largest magnitude composite mean difference for DJF daytime was found for EVI (–7.33 K), and is significant at the 5% level. Therefore, the days with EVI values greater than or equal to the 90^{th} percentile of EVI are associated with an urban LST that is significantly lower than days with EVI values less than or equal to the 10^{th} percentile value of EVI. This matches the correlation analysis, and similarly, all variables related to moisture show significant, negative composite mean differences in DJF daytime. Wind speed also shows a significant, negative composite mean difference (–1.98 K) due to how a high urban wind speed can advect urban heat, therefore the windiest days cause an urban LST that is significantly lower than calm conditions. A significant, positive composite mean difference is found for albedo (6.77 K), which may be related to how a high AOD can decrease surface albedo (Fig. 1), and with a lesser albedo, more heat can be retained by the surface. Similar to DJF daytime, the highest magnitude composite mean difference for ASO daytime is for EVI (–2.66 K), which is significant at the 5% level. In the correlation analysis, EVI and soil moisture had the same coefficients, but for the composite analysis, soil moisture has a mean composite difference of –1.79 K. A significant, negative composite mean difference is also found for specific humidity in ASO daytime (–0.89 K). For DJF nighttime, the highest magnitude composite mean difference was found for albedo (2.73 K); however, this is only slightly higher in magnitude than that of EVI (–2.53 K), which was the variable with the highest correlation. Significant, negative composite mean differences were found for latent heat (–1.56 K) and soil moisture (–1.09 K) in DJF nighttime since these factors are also related to vegetation. A significant, positive composite mean difference was found for specific humidity (0.92 K), therefore nights with extremely high values of water vapor are associated with a significantly greater urban LST compared to nights with extremely low values of water vapor. This is likely a result of how water vapor can enhance downward longwave radiation and increase LST. For ASO nighttime, the only significant composite mean difference was found for wind speed (–0.94 K), which does not match the correlation analysis.

Table 2 shows the standardized regression coefficients from the MLR analysis. According to these results, the most important variable for DJF daytime is EVI, which matches the correlation and composite analyses. EVI is also shown to be most important for ASO daytime, which matches the composite analysis and somewhat matches the correlation analysis given that soil moisture had the same correlation as EVI. For DJF nighttime, the most important variable is EVI, which only matches the correlation analysis. For ASO nighttime, near-surface specific humidity is shown to be most important, which only matches the correlation analysis. Note that in DJF daytime, the standardized regression coefficient for soil moisture is positive, despite soil moisture having a negative correlation coefficient with urban LST in Fig. 2a. In linear regression, it can be expected that the regression and correlation coefficients will have the same sign given only one predictor variable. However, since multiple predictors are used here, confounding can occur in which a predictor variable can influence the dependent variable and other independent variables, thereby resulting in a sign change for the regression coefficient compared to the correlation coefficient and possibly less reliable results (Graham, 2003). A physically unexpected sign of the regression coefficient also occurred for soil moisture in DJF nighttime. Since Fig. 1c shows that soil moisture and EVI are directly related, it would be expected that soil moisture would have a negative regression coefficient similar to EVI. Confounding also occurred for the latent heat flux regression coefficient in ASO nighttime. Since Fig. 1d shows a direct relationship between EVI and latent heat, it would be expected that latent heat would have a negative regression coefficient similar to EVI. Note that if these variables that experienced confounding are removed from the MLR analysis, the major results in terms of which variable has the highest magnitude standardized regression coefficient do not change (not shown). For the RF analysis (Fig. 3), the results continue to be consistent for DJF and ASO daytime compared to the previous methods, in which EVI is shown to be the leading control. For DJF nighttime, EVI is shown to be the most important, which matches the correlation and MLR analyses, but not the composite analysis. For ASO nighttime, albedo appears to be the leading control, which does not match any of the previous analyses.

Comparing the MLR and RF methods (Fig. 4), the RMSE and MAE values are the lowest for all seasons and time of day cases for the RF model. In terms of the variance explained, for MLR, DJF daytime performed best (R^2=0.65), followed by DJF nighttime (R^2=0.57), ASO daytime (R^2=0.26), and lastly ASO nighttime (R^2=0.15). The variance explained for each RF model improved over its respective MLR model. The highest explained variance was found for DJF daytime (R^2=0.92), followed by DJF nighttime (R^2=0.87), ASO daytime (R^2=0.79), and lastly ASO nighttime (R^2=0.77). However, these improvements in error values and explained variance in the RF model may be the result of overfitting. While the MLR analysis showed to have little influence of overfitting after performing the k-fold cross-validation (not shown), the RF did have a degree of overfitting (Table 3). The overfitting is shown to be most noticeable for ASO daytime and nighttime due to large differences in the explained variance values. The range in values of urban LST is less for ASO daytime and nighttime compared to DJF daytime and nighttime (Fig. 4), therefore, even if the predictors vary in magnitude, they can reach similar decisions for the predictand, likely resulting in overfitting.

Overall, since all four methods show EVI to have the highest respective value for DJF daytime, it is likely the leading control on urban LST at this time. It was expected AOD would be the leading control, and while its association is statistically significant in some of the methods, results suggest that the effect of aerosols are smaller. Despite UHI intensity being negative during DJF daytime, its magnitude is smallest during DJF daytime and it does not have a significant trend

(Sussman et al., 2019). Since Bengaluru was once known as the *"Garden City"* of India and now as the *"Silicon City"* (Sudhira et al., 2007), surface changes in terms of vegetation are noteworthy qualitatively and quantitatively. These results suggest the impacts of aerosols are enough to cancel out a high UHI that would occur if only vegetation were at play during DJF daytime. For ASO daytime, while the correlation analysis produced the same value for EVI and soil moisture, all subsequent analyses showed EVI to have the highest values. Therefore, EVI is likely the leading control on urban LST during ASO daytime. This matches the hypothesis that since vegetation is most abundant during the wet season, and thus most impactful on the amount of evaporative cooling, it will likely largely influence surface temperature. For DJF nighttime, the correlation, MLR, and RF analyses show EVI to lead, while the composite analysis shows albedo to lead slightly over EVI. Therefore, three of the four methods agree. From a physical perspective, since there is great confidence that EVI is the leading control during DJF daytime, there would likely be a high retention of heat at nighttime due to a low EVI. While impacts of albedo also appear to control during DJF daytime, these impacts are not nearly as strong as EVI. Therefore, it may be more likely that EVI is the leading control during DJF nighttime, which matches the hypothesis. For ASO nighttime, the correlation and MLR analyses show specific humidity to lead, the composite analysis shows wind speed to have the only significant composite mean difference, and the RF analysis shows albedo to lead. Therefore, two methods have consensus on specific humidity while the other methods diverge. From a physical standpoint, lower-tropospheric water vapor is abundant during the wet season and is expected to be most influential at nighttime. A high water vapor content can enhance downward longwave radiation, which is the primary type of radiation at nighttime. Therefore, it may be likely that specific humidity is the leading control during ASO nighttime. EVI was hypothesized to be the leading control during ASO nighttime, and while its association is significant in some of the methods, since ASO is also characterized by high amounts of water vapor, which can be influential on surface temperature at nighttime, it amplifies the impacts of heat retention by the surface due to a low EVI.

A further comparison of the MLR and RF statistical models shows that all season and time of day cases had consensus for what the leading control is, with the exception of ASO nighttime in which MLR suggested specific humidity and the RF suggested albedo. Analysis of the linear correlation (Fig. 2a) for ASO nighttime shows that correlations are weakest in magnitude compared to the other season and time of day cases, and thus are not as linear in nature. To further illustrate these weakly linear relationships amongst the predictor variables in ASO nighttime, Fig. 5 shows the relationship between urban LST and albedo. Comparing Figs. 5a–b, while the best relationship for the data is unknown, it is clear that the nonlinear curves fit the observed data better than the linear line. Comparing Figs. 5c–d, in which the predicted urban LST is plotted against albedo for MLR and RF, the nonlinearity of the relationship is best preserved by the RF predicted result, despite this prediction having the histories of the other predictors. There was consensus on specific humidity for the correlation and MLR analyses for ASO nighttime, which physically makes sense. While albedo could be important for the retention of heat at nighttime, the influences of water vapor are expected to be most impactful during ASO nighttime. Perhaps the complexity of this time, i.e., wet days and dry days both occur, water vapor is abundant, and vegetation is at its peak, yet is decreasing, is not understood by the RF. Therefore, while the RF method may be able to better capture nonlinear relationships, this machine learning method may not yet understand the physics of the data, and therefore are correct in some cases but misleading in others.

4. Summary and Discussion

There are multiple environmental factors that can influence a city's UHI and these factors often are related to each other. Previous studies have tried to determine the leading controlling factors using a variety of regression and machine learning methods for different cities (e.g., Kim and Baik, 2002; Zhou et al., 2011; Ho et al., 2014; Makido et al., 2016). However, this is a research question that must be addressed at the local-scale since each city is unique in terms of which environmental factors may be important within its microclimate and how the factors interact with each other. This study built on previous work by applying a similar framework, but using variables specific to Bengaluru. The MLR and RF methods, as well as linear correlation and a composite analysis, were applied to Bengaluru to assess variable importance for urban LST from 2003–2018 using MODIS and ERA5 data during the dry and wet seasons.

Results showed that for both the dry and wet seasons during the daytime, as well as the dry season nighttime, EVI was the leading control. For the wet season nighttime, specific humidity was shown to be the leading control. Therefore, urban heat is primarily controlled by vegetation in Bengaluru, and thus urban heat can be reduced by increasing vegetation within the city. However, vegetation and specific humidity are related (Fig. 1). At daytime, increased specific humidity is due to an increase in the latent heat flux brought about by a high amount of vegetation.

Therefore, specific humidity and urban LST are inversely related at daytime (Fig. 2a). In contrast, specific humidity can increase urban LST at nighttime due to its ability to enhance downward longwave radiation, and thus specific humidity is directly related to urban LST at nighttime (Fig. 2a). Therefore, mitigation strategies that increase vegetation must not increase water vapor substantially, otherwise urban heat may amplify during the wet season nighttime since specific humidity is the controlling factor at that time. In terms of the statistical models, results showed the RF model to have lower RMSE and MAE values and higher explained variance values compared to MLR. However, while the error values may be less and the model may perform better, it may be more accurate for the wrong reasons from a physical standpoint as well as due to overfitting.

A limitation of this work includes that many different sets of controlling factors could have been chosen and assessed in different ways. For example, Makido et al. (2016) used mean albedo within a certain radius, percentage of urban area within a certain radius, percentage of vegetation within a certain radius, and distance to the coast as the controlling factors to analyze urban heat in Doha, Qatar. Therefore, even though their work also used albedo and vegetation, the variables were assessed as distance effects on urban heat. To increase robustness of the results shown here, future work should conduct modeling experiments to understand how perturbations to the leading control influences urban heat and other related variables. For example, an experiment that decreases vegetation over the urban surface, while keeping all other parameters constant, should result in an urban LST increase and a decrease in the latent heat flux, soil moisture, and specific humidity during the daytime, and vice versa for an increase in vegetation, according to the results of this study. Additional modeling experiments can also be done for the factors that were deemed to be less important to evaluate if those factors produce smaller changes in urban LST than the leading factors. Overall, modeling experiments of this nature will help to understand how much urban heat can be reduced as well as amplified if perturbing the controlling factors in both directions and better understand mechanisms.

Acknowledgments

HS acknowledges the funding support from the Science, Mathematics, and Research for Transformation (SMART) fellowship. HS also acknowledges the technical assistance of Dr. Ajay Raghavendra for using the random forest algorithm. AD acknowledges funding support from the funding from the U.S. National Science Foundation (Grant Nos. AGS-2015780 and OISE-

1743738) and the U.S. National Oceanic and Atmospheric Administration (Award No. NA18OAR4310425). PR acknowledges funding support from the U.S. National Science Foundation grant number AGS1757342. The authors also thank the two anonymous reviewers for their constructive comments that have improved the quality of this manuscript and Dr. Peter J. Marcotullio for serving as Editor of this manuscript.

Author contributions

HS conceptualized the study. AD and PR helped refine the methods. HS performed the calculations and made the figures. HS wrote the first draft of this manuscript. AD and PR helped revise this manuscript and provided feedback.

References

Andersson, E., 2006. Urban landscapes and sustainable cities. Ecol. Soc. 11, 34.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley, New York.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees (1st Ed.), CRC Press, Boca Raton.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Census of India: Karnataka District Census Handbook, Bangalore, 2011. http://censusindia.gov.in/2011census/dchb/2918_PART_A_DCHB_BANGALORE.pdf. (accessed 2 May 2019).

Cusack, S., Slingo, A., Edwards, J.M., Wild, M., 1998. The radiative impact of a simple aerosol climatology on the Hadley Centre atmospheric GCM. Quart. J. Roy. Meteor. Soc. 124, 2517–2526.

Dai, A., Deser, C., 1999. Diurnal and semidiurnal variations in global surface wind and divergence fields. J. Geophys. Res. 104, 31109–31125.

Dai, A., Trenberth, K., Karl, T.R., 1999. Effects of clouds, soil moisture, precipitation, and water vapor on diurnal temperature range. J. Clim. 12, 2451–2473.

Dufresne, J.L., Gautier, C., Ricchiazzi, P., 2002. Long wave scattering effects of mineral aerosols. J. Atmos. Sci. 59, 1959–1966.

Garrat, J.R., 1994. Review: the atmospheric boundary layer. Earth Sci. Rev. 37, 89–134.

Gagne, D.J., McGovern, A., Haupt, S., Sobash, R., Williams, J., Xue, M., 2017. Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. Wea. Forecasting. 32, 1819–1840.

Geisser, S., 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70, 320–328.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognit. Lett. 27, 294–300.

Graham, M.H., 2003. Confronting multicollinearity in ecological multiple regression. Ecology. 84, 2809–2815.

Grimm, N.B., Faeth, S.H., Golubiewski, N.E., Redman, C.L., Wu, J., Bai, X., Briggs, J.M., 2008. Global change and the ecology of cities. Science 319, 756–760.
Guo, A., Yang, J., Xiao, X., Xia, J., Jin, C., Li, X., 2020. Influences of urban spatial form on urban heat island effects at the community level in China. Sustain. Cities Soc. 53, 101972.
Guo, S., Hu, M., Zamora, M.L., Peng, J., Shang, D., Zheng, J., Du, Z., Wu, Z., Shao, M., Zeng, L., 2014. Elucidating severe urban haze formation in China. Proc. Natl. Acad. Sci. USA 111, 17373–17378.
Filho, W.L., Icaza, L.E., Neht, A., Klavins, M., Morgan, E.A., 2018. Coping with the impacts of urban heat islands. A literature based study on understanding urban heat vulnerability and the need for resilience in cities in a global climate change context. J. Clean. Prod. 171, 1140–1149.
Hastie, T., Tibshirani, R., Friedman, J., 2009. Elements of statistical learning: data mining, inference and prediction (2nd Ed.), Springer Verlag, Heidelberg.
Ho, H., Knudby, A., Sirovyak, P., Xu, Y., Hodul, M., Henderson, S., 2014. Mapping maximum urban air temperature on hot summer days. Remote Sens. Environ. 154, 38–45.
Ismail, R., Mutanga, O., Kumar, L., 2010. Modeling the potential distribution of pine forests susceptible to sirex noctillo infestations in Mpumalanga, South Africa. T. GIS. 14, 709–726.
Ismail, R., Mutanga, O., 2010. A comparison of regression tree ensembles: predicting Sirex noctillo induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa. Int. J. Appl. Earth Obs. 12, S45–S51.
Jacobson, M.Z., 2001. Strong radiative heating due to the mixing state of black carbon on atmospheric aerosols. Nature. 409, 695–697.
Jiang, Y., Zhou, L., Raghavendra, A., 2020. Observed changes in fire patterns and possible drivers over Central Africa. Environ. Res. Lett. 15, 0940b8.
Kanakidou, M., Mihalopoulos, N., Kindap, T., Im, U., Vrekoussis, M., Gerasopoulos, E., Dermitzaki, E., Unal, A., Koçak, M., Markakis, K., Melas, D., Kouvarakis, G., Youssef, A.F., Richter, A., Hatzianastassiou, N., Hilboll, A., Ebojie, F., Wittrock, F., von Savigny, C., Burrows, J.P., Ladstaetter-Weissenmayer, A., Moubasher, H., 2011. Megacities as hot spots of air pollution in the East Mediterranean. Atmos. Environ. 45, 1223–1235.
Kim, Y.H., Baik, J.J., 2002. Maximum urban heat island intensity in Seoul. J. Appl. Meteorol. Climatol. 41, 651–659.
Kim, Y.H., Baik, J.J., 2005. Spatial and temporal structure of the urban heat island in Seoul. J. Appl. Meteorol. Climatol. 44, 591–605.
Kishtawal, C.M., Niyogi, D., Tewari, M., Pielke, R.A., Shepherd, J.M., 2010. Urbanization signature in the observed heavy rainfall climatology over India. Int. J. Climatol. 30, 1908–1916.
Koelmans, A.A., Jonker, M.T.O., Cornelissen, G., Bucheli, T.D., van Noort, P.C.M., Gustafsson, Ö., 2006. Black carbon: The reverse of its dark side. Chemosphere. 63, 365–377.
Kolokotroni, M., Giridharan, R., 2008. Urban heat island intensity in London: An investigation of the impact of physical characteristics on changes in outdoor air temperature during summer. Sol Energy 82, 986–998.
Lacis, A.A., Mishchenko, M.I., 1995: Climate forcing, sensitivity, and response, in: Charlson, R.J., Heintzenberg, J. (Eds.), Aerosol Forcing of Climate. John Wiley, Chichester, UK, pp. 11–42.
Makido, Y., Shandas, V., Ferwati, S., Sailor, D.J., 2016. Daytime variation of urban Heat Islands: the case study of Doha, Qatar. Climate 4, 32.

McGovern, A., Gagne, D.J., Williams, J., Brown, R., Basara, J., 2014. Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. Mach. Learn. 95, 27–50.

McMeeking, G.R., Good, N., Petters, M.D., McFiggans, G., Coe, H., 2011. Influences on the fraction of hydrophobic and hydrophilic black carbon in the atmosphere. Atmos. Chem. Phys. 11, 5099–5112.

Miao, S., Chen, F., LeMone, M., Tewari, M., Li, Q., Wang, Y., 2009. An observational and modeling study of characteristics of urban heat island and boundary layer structures in Beijing. J. Appl. Meteorol. Climatol. 48, 484-501.

Mitchell, J.F.B., Davis, R.A., Ingram, W.J., Senior, C.A., 1995. On surface temperature, greenhouse gasses, and aerosols: models and observations. J. Clim. 8, 2364–2386.

Mutanga, O., Adam, E., Cho, M.A., 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int. J. Appl. Earth Obs. 18, 399–406.

Pal, M. 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222.

Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the K öppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644.

Peng, S., Piao, S., Ciais, P., Friedlingstein, P., Ottle, C., Bréon, F.M.., Nan, H., Zhou, L., Myneni, R.B., 2012. Surface urban heat island across 419 global big cities. Environ. Sci. Technol. 46, 696–703.

Prasad, A., Iverson, L., Liaw, A., Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems. 9, 181–199.

Qian, Y., Leung, R., Ghan, S.J., Giorgi, F., 2003. Regional climate effects of aerosols over China: Modeling and observations. Tellus B. 55, 914–934.

Qian, Y., Kaiser, D.P., Leung, L.R., Xu, M., 2006. More frequency cloud-free sky and less surface solar radiation in China from 1955 to 2000. Geophys. Res. Lett. 33, L01812.

Ramachandran, S., Kedia, S., 2010. Black carbon aerosols over an urban region: Radiative forcing and climate impact. J. Geophys. Res. 115, D10202.

Ramachandran, S., Kedia, S., Srivastava, R., 2012. Aerosol optical depth trends over different regions of India. Atmos. Environ. 49, 338–347.

Smith, P.F., Ganesh, S., Liu, P., 2013. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. J. Neurosci. Methods 220, 85–91.

Sudhira, H.S., Ramachandra, T.V., Bala Subrahmanya, M.H., 2007. City profile: Bangalore. Cities 24, 379–390.

Sussman, H.S., Raghavendra, A., Zhou, L., 2019. Impacts of increased urbanization on surface temperature, vegetation, and aerosols over Bengaluru, India. Remote Sens. Appl.: Soc. Environ. 16, 100261.

Taha, H., 1997. Urban climates and heat islands: albedo, evapotranspiration, and anthropogenic heat. Energ. Buildings. 25, 99–103.

Theeuwes, N.E., Steeneveld, G.J., Ronda, R.J., Heusinkveld, L.W., van Hove, W.A., Holtslag, A.A.M., 2014. Seasonal dependence of the urban heat island on the street canyon aspect ratio. Q. J. R. Meteorol. Soc. 140, 2197–2210.

Tie. W., Cao, J., 2009. Aerosol pollution in China: Present and future impact on environment. Particuology. 7, 426–431.

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A., Torricelli, P.,

2011. Application of a random forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecol. Model. 222, 1471–1478.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C., McCulloch, C.E., 2005. Regression methods in statistics: linear, logistic, survival and repeated measures models. Springer, New York.

Vu, D.H., Muttaqi, K.M., Agalgaonkar, A.P., 2015. A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. Appl. Energy. 140, 385–394.

Williams, J., 2014. Using random forests to diagnose aviation turbulence. Mach. Learn. 95, 51–70.

Yang, J., Jin, S., Xiao, X., Jin, C., Xia, J., Li, X., Wang, S., 2019. Local climate zone ventilation and urban land surface temperatures: towards a performance-based and wind-sensitive planning proposal in megacities. Sustain. Cities Soc. 47, 101487.

Yang, J., Wang, Y., Xiu, C., Xiao, X., Xia, J., Jin, C., 2020a. Optimizing local climate zones to mitigate urban heat island effect in human settlements. J. Clean. Prod. 275, 123767.

Yang, J., Zhan, Y., Xiao, X., Xia, J., Sun, W., Li, X., 2020b. Investigating the diversity of land surface temperature characteristics in different scale cities based on local climate zones. Urban Clim. 34, 100700.

Yang, J., Ren, J., Sun, D., Xiao, X., Xia, J., Jin, C., Li, X., 2021. Understanding land surface temperature impact factors based on local climate zones. Sustain. Cities Soc. 69, 102818.

Zhang, X., Estoque, R.C., Murayama, Y., 2017. An urban heat island study in Nanchang City, China based on land surface temperature and social-ecological variables. Sustain. Cities Soc. 32, 557–568.

Zhou, J., Chen, Y., Wang, J., Zhan, W., 2011. Maximum nighttime urban heat island (UHI) intensity simulation by integrating remotely sensed data and meteorological observations. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 4, 138–146.

Zhou, L., Dickinson, R.E., Tian, Y.H., Fang, J.Y., Li, Q.X., Kaufmann, R.K., 2004. Evidence for a significant urbanization effect on climate in China. Proc. Natl. Acad. Sci. 101, 9540–9544.

Zhou, L., Dickinson, R.E., Tian, Y., Vose, R.S., 2007. Impact of vegetation removal and soil aridation on diurnal temperature range in a semiarid region – Application to the Sahel. Proc. Natl. Acad. Sci. USA. 104, 17937–17942.

Tables and Figures

Table 1. Summary of all datasets used in this study including their spatial resolution, temporal frequency, and weblink for more information. All datasets were obtained for 2003–2018.

Variable	Dataset	Spatial/Temporal Resolution	Weblink
Aerosol optical depth (AOD)	MODIS Terra and Aqua combined product (MCD19A2)	1 km/daily	https://lpdaac.usgs.gov/products/mcd19a2v006/
Enhanced vegetation index (EVI)	MODIS Aqua (MYD13A2) and Terra (MOD13A2)	1 km/16-day composites	Aqua: https://lpdaac.usgs.gov/products/myd13a2v006/ Terra: https://lpdaac.usgs.gov/products/mod13a2v006/
Land cover	MODIS Terra and Aqua combined land cover (MCD12Q1)	500 m/annual	https://lpdaac.usgs.gov/products/mcd12q1v006/
Land surface temperature (LST)	MODIS Aqua (MYD11A2) and Terra (MOD11A2)	1 km/8-day composites	Aqua: https://lpdaac.usgs.gov/products/myd11a2v006/ Terra: https://lpdaac.usgs.gov/products/mod11a2v006/
Albedo	ERA5-Land	0.1°/hourly	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview
Latent heat			
Soil moisture			
10-m u and v wind			
Specific humidity from 850–1000 hPa	ERA5-Pressure levels	0.25°/hourly	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview

Table 2. Summary of the standardized regression coefficients for all possible controlling factors of urban LST in Bengaluru, India as determined by multiple linear regression. Bold indicates the coefficient is significant at the 10% level. Bold and underlined indicates the coefficient is significant at the 5% level.

	DJF Daytime	ASO Daytime	DJF Nighttime	ASO Nighttime
Albedo	**0.22**	**0.29**	**0.39**	0.06
AOD	**–0.11**	–0.02	0.06	0.07
EVI	**–0.79**	**–0.37**	**–0.64**	**–0.14**
Latent heat	–0.02	0.04	–0.01	**0.24**
Soil moisture	**0.51**	**–0.19**	**0.40**	–0.14
Specific humidity	**–0.18**	–0.01	**0.36**	**0.35**
Wind speed	–0.03	**–0.18**	–0.06	–0.14

Table 3. Comparison of the variance explained (R^2), root mean square error (RMSE), and mean absolute error (MAE) between the urban LST random forest models with the training set only (i.e., values from Fig. 4) and with cross-validation. The error values are in units of K.

	Training Set Only			Cross-validation		
	R^2	RMSE	MAE	R^2	RMSE	MAE
DJF Daytime	0.90	1.09	0.78	0.71	1.69	1.25
ASO Daytime	0.79	0.95	0.75	0.26	1.37	1.13
DJF Nighttime	0.87	0.61	0.48	0.54	0.96	0.77
ASO Nighttime	0.77	0.57	0.42	0.14	0.79	0.60

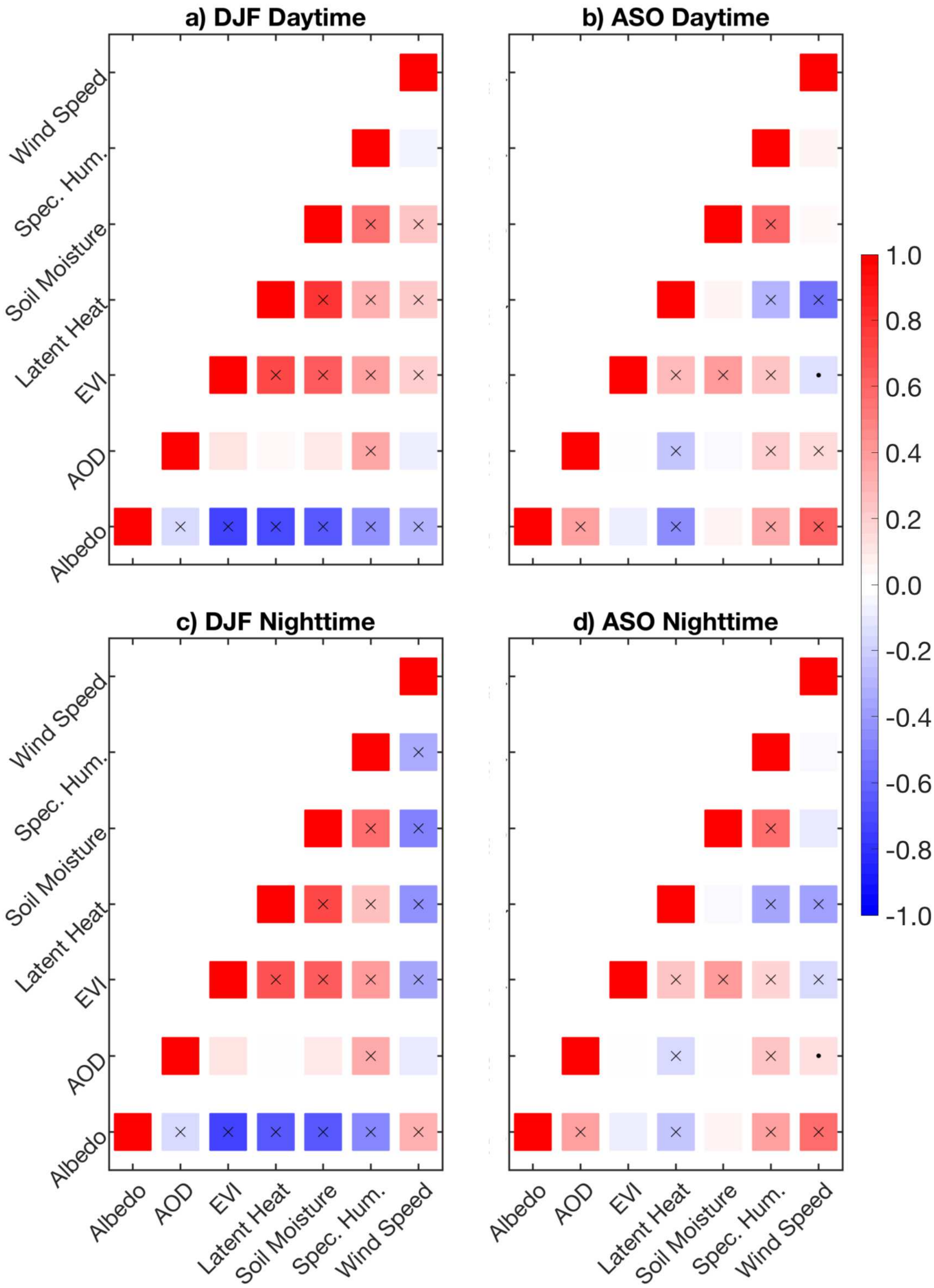

Figure 1. The matrix of linear correlation coefficients among the prospective controlling factors of urban LST in Bengaluru, India for a) DJF daytime, b) ASO daytime, c) DJF nighttime, and d) ASO nighttime. A center dot (cross) indicates that the correlation is statistically significant at the 10% (5%) level. The abbreviations are: aerosol optical depth (AOD), enhanced vegetation index (EVI), and specific humidity (spec. hum.).

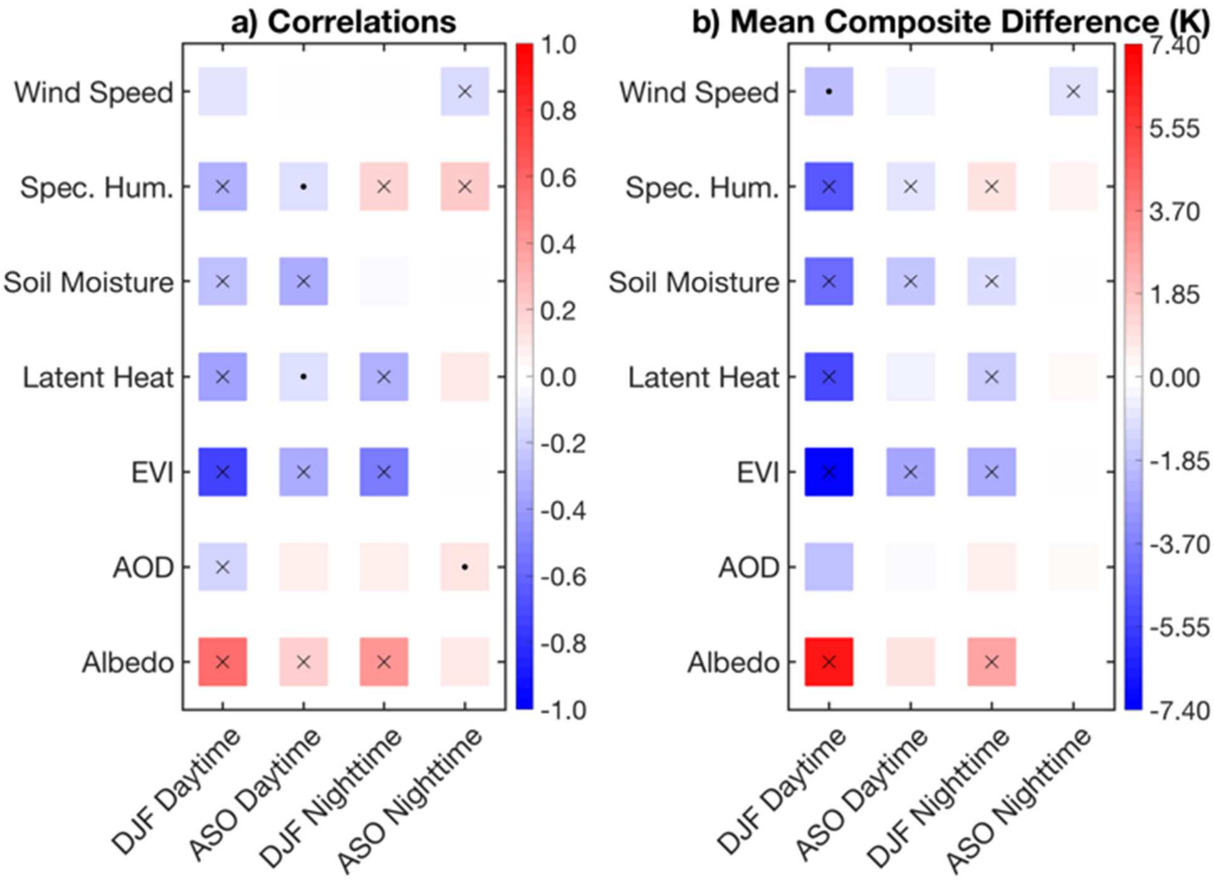

Figure 2. a) Temporal linear correlation coefficients of the prospective controlling factors with urban LST in Bengaluru, India during 2003–2018 for each season and time of day. b) The urban LST composite mean difference (K) in Bengaluru, India for the events corresponding ≥90th percentile minus ≤10th percentile of each prospective controlling factor. A center dot (cross) indicates that the correlation or composite mean difference is statistically significant at the 10% (5%) level. The abbreviations are: aerosol optical depth (AOD), enhanced vegetation index (EVI), and specific humidity (spec. hum.).

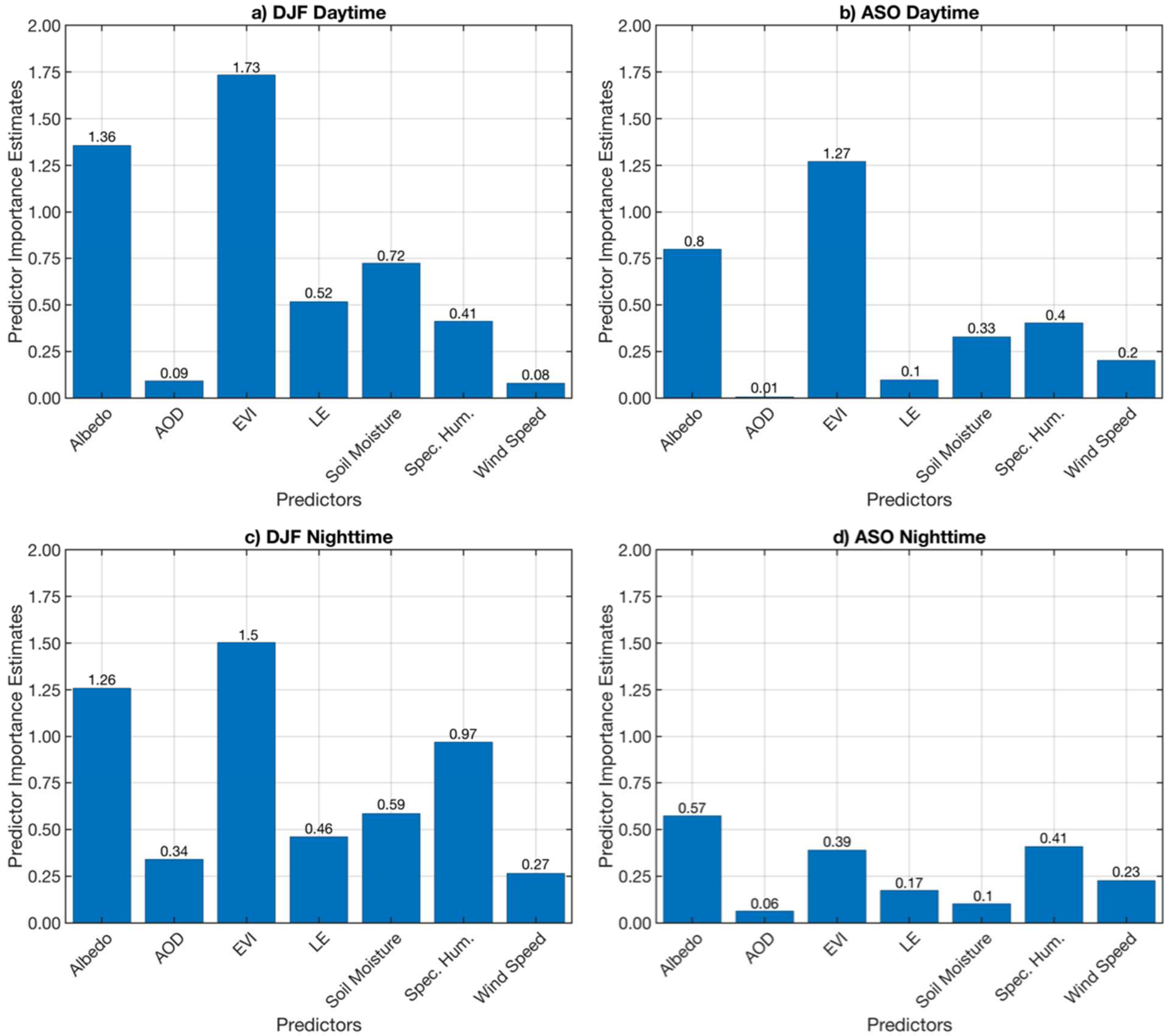

Figure 3. The predictor importance estimates for urban LST in Bengaluru, India based on the random forest (RF) model for each season and time of day. The abbreviations are: aerosol optical depth (AOD), enhanced vegetation index (EVI), latent heat (LE), and specific humidity (spec. hum.).

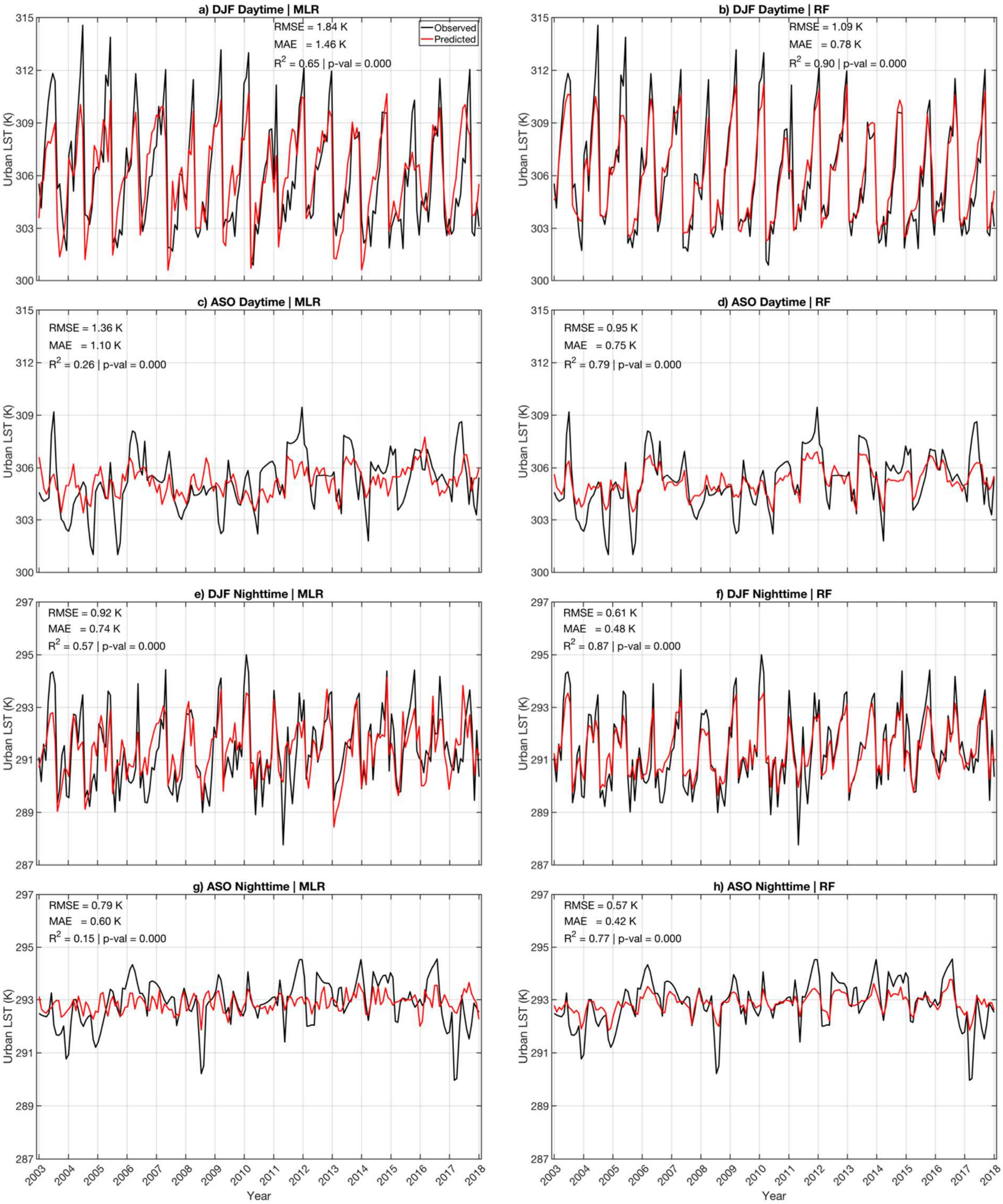

Figure 4. (Left column) The predicted responses (red) of urban LST in Bengaluru, India from 2003–2018 based on the multiple linear regression (MLR) model for each season and time of day compared to the observed (black) time series of urban LST based on MODIS LST and land cover data. The root mean square error (RMSE) and mean absolute error (MAE) of the predicted compared to the observed are reported in each panel. The total variance explained by the MLR model (R^2) and its p-value (p-val) are also shown. (Right column) Same as left column, but based on the random forest (RF) model.

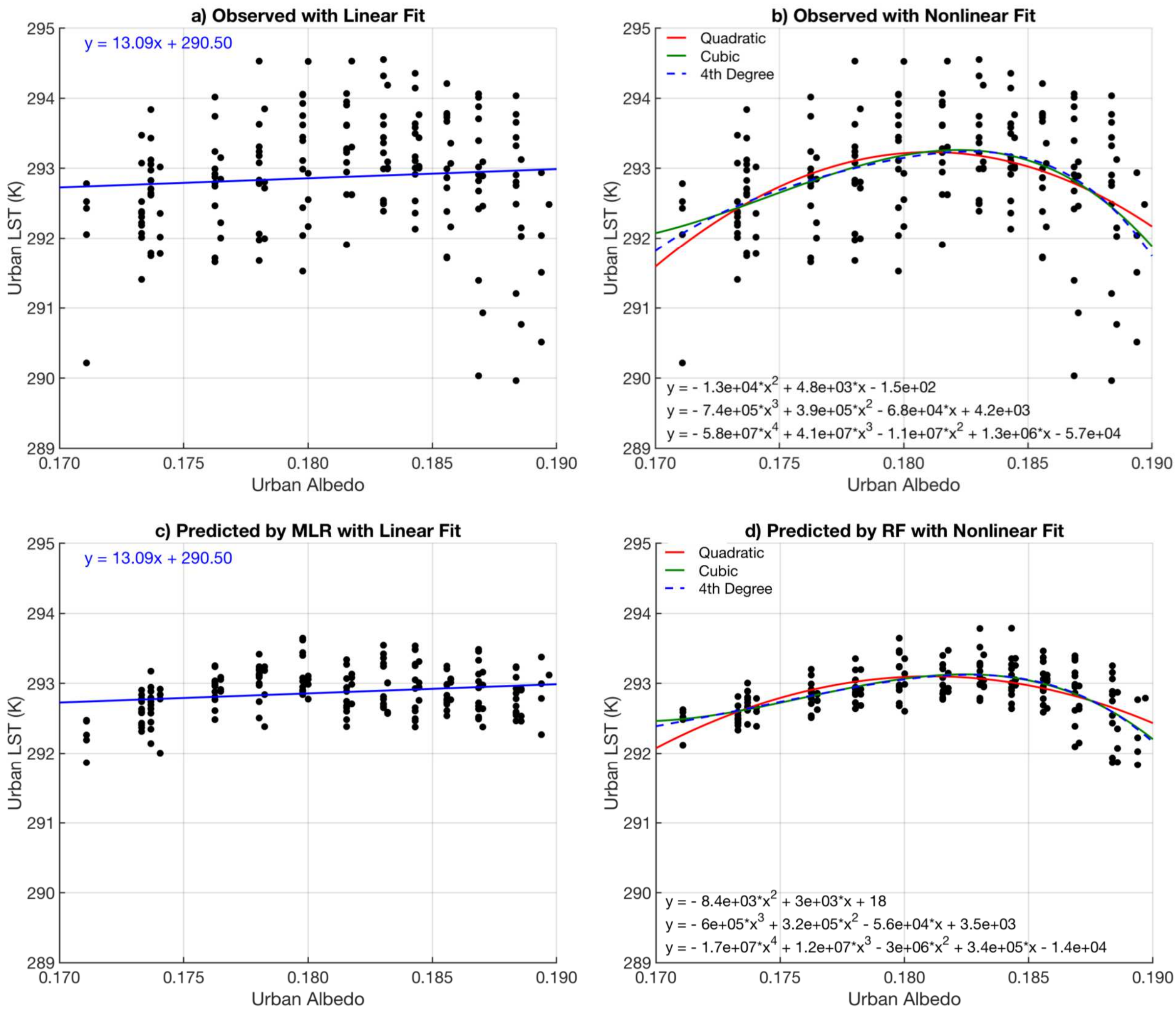

Figure 5. a) The observed relationship in August-September-October (ASO) nighttime between urban LST and albedo in Bengaluru, India during 2003–2018 and the associated linear fit. b) Same as a) but with quadratic, cubic, and 4th degree nonlinear fit lines. c) The predicted urban LST by the multiple linear regression (MLR) model in relation to observed albedo in Bengaluru, India during 2003–2018 for ASO nighttime and the associated linear fit. d) The predicted urban LST by the random forest (RF) model in relation to observed albedo in Bengaluru, India during 2003–2018 for ASO nighttime and the associated quadratic, cubic, and 4th degree nonlinear fit lines. The fit line equations are shown in each panel.