



Article

Machine Learning-Based Modeling of Air Temperature in the Complex Environment of Yerevan City, Armenia

Garegin Tepanosyan¹, Shushanik Asmaryan¹ , Vahagn Muradyan¹, Rima Avetisyan¹, Azatuhi Hovsepyan¹, Anahit Khlghatyan¹ , Grigor Ayvazyan¹ and Fabio Dell'Acqua^{2,*}

¹ Centre for Ecological-Noosphere Studies, National Academy of Sciences of Armenia, Abovyan Street 68, Yerevan 0025, Armenia; garegin.tepanosyan@cens.am (G.T.); shushanik.asmaryan@cens.am (S.A.); vahagn.muradyan@cens.am (V.M.); rima.avetisyan@cens.am (R.A.); aza.hovsepyan@cens.am (A.H.); anahit.khlghatyan@cens.am (A.K.); grigor.ayvazyan@cens.am (G.A.)

² Department of Electrical, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia, Italy

* Correspondence: fabio.dellacqua@unipv.it; Tel.: +39-0382-985664

Abstract: Machine learning (ML) was used to assess and predict urban air temperature (T_{air}) considering the complexity of the terrain features in Yerevan (Armenia). The estimation was performed based on the Partial Least-Squares Regression (PLSR) model with a high number (30) of input variables. The relevant parameters include a newly purposed modification of spectral index IBI-SAVI, which turned out to strongly impact T_{air} prediction together with land surface temperature (LST). Cross-validation analysis on temperature predictions across a station-centered 1000 m circular area revealed quite a high correlation ($R^2_{\text{Val}} = 0.77$, $\text{RMSE}_{\text{Val}} = 1.58$) between the predicted and measured T_{air} from the test set. It was concluded the remote sensing is an effective tool to estimate T_{air} distribution where a dense network of weather stations is not available. However, further developments will include incorporation of additional weather parameters from the weather stations, such as precipitation and wind speed, as well as the use of non-parametric ML techniques.

Keywords: urban air temperature; land surface temperature; multiple independent variables; urban heat; remote sensing data; machine learning (ML); ML-driven partial least-squares regression (PLSR)



Citation: Tepanosyan, G.; Asmaryan, S.; Muradyan, V.; Avetisyan, R.; Hovsepyan, A.; Khlghatyan, A.; Ayvazyan, G.; Dell'Acqua, F. Machine Learning-Based Modeling of Air Temperature in the Complex Environment of Yerevan City, Armenia. *Remote Sens.* **2023**, *15*, 2795. <https://doi.org/10.3390/rs15112795>

Academic Editor: Federico Karagulian

Received: 1 April 2023

Revised: 22 May 2023

Accepted: 24 May 2023

Published: 27 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Air temperature (T_{air}) is a climate variable describing the energy and thermal balance in a very special zone of the Earth–atmosphere system, namely the surface of the Earth, which is, at the same time, the very bottom layer of the atmosphere [1]. It is a useful factor in tracking the climate change associated with human activities, especially in urban areas, which represent the “peaks” of anthropogenic activities and are where the climate change impact is expressed and sensed the most. T_{air} is extremely relevant as an input to models determining the urban heat island (UHI) effect, i.e., the higher air temperatures recorded in urban areas with respect to the surrounding countryside. UHIs can directly affect the health and well-being of urban residents. Not only do higher temperatures obviously cause increased discomfort during heat waves, but high UHI intensity is generally correlated with increased concentrations of air pollutants. Mapping T_{air} is thus important to understand the dynamics of urban microclimates; however, T_{air} is usually recorded by weather stations, which measure T_{air} between 1.5 and 2 m above ground and are distributed sparsely, thus failing to provide synoptic spatial coverage [2,3]. Moreover, urban areas are more heterogeneous than rural areas; hence, the effective coverage of weather stations providing long-term observational data tends to be narrow [4,5], leaving large swathes of urban areas unobserved. This is a reason to use remote sensing data for T_{air} prediction; remote sensing offers a possibility to track its seasonal behavior and especially its spatial distribution. T_{air} cannot be observed directly from space but thermal infrared sensors enable the derivation

of land surface temperature (LST), which is a limiting condition for the energy balance and is widely used to assess the spatial distribution of T_{air} [6,7].

The methodologies and approaches to assess and predict T_{air} via remote sensing are different. They are mainly based on a hybrid methodology, which combines GIS and re-remote sensing data. For example, in 2008, Cristobel J et al. applied this methodology combining geographical variables (e.g., altitude, latitude, continentality and solar radiation) and remote sensing predictors related with T_{air} , such as albedo, LST and the vegetation index NDVI obtained from Landsat sensors NOAA and TERRA (MODIS), and they used multiple regression analysis and spatial interpolation techniques for data processing. The authors support that this combined approach underpins the best T_{air} models, and NDVI and LST are the most powerful remote sensing (RS)-based predictors in T_{air} modeling [8].

It is worth recalling that LST is a basic physical parameter to describe ecological, hydrological and atmospheric processes and has a strong relationship with near surface T_{air} . However, these two parameters have different responses to atmospheric conditions, and their link becomes even more complex in mountainous complex terrain and/or where weather stations are scarce. In 2020, Mutiibwa D et al. analyzed the benefits and some limitations of using LST as a variable for predicting T_{air} in two case study regions of Nevada that are well-known for complex mountainous topography. Though with some limitations and complexities, the relationship between LST and T_{air} was found to be consistent [9].

In 2020, Nikoloudakis N et al. applied the hybrid methodology to predict T_{air} in urban areas without LST. They developed and used predictive models, which are based on urban morphological peculiarities, such as land cover and terrain, as well as in situ T_{air} measurements from urban weather stations [10].

The modeling of urban T_{air} is much more complicated because of the heterogeneity, where the scarcity of weather stations hinders accurate spatial representation of T_{air} [11]. It is further complicated in mountainous urban areas [12]. The use of machine learning (ML) approaches increases the accuracy of the estimations of T_{air} [13]. A wide variety of statistical and ML models have been used so far. Among them, the most popular ones still remain multiple regression and ANN models [1,11,12,14–16], with differences in results mainly due to the selection of input variables.

A most important component of regression models is the number of variables, which can range from one to tens of variables. For example, up to 7 variables (skin temperature (LST), elevation, the Normalized Difference Water Index (NDWI), Sky View Factor (SVF), incident solar radiation, distance from the ocean and atmospheric water vapor) were used by Hung Chak [17], 24 variables by Modeste Meliho (Aqua day clear-sky coverage (ACDC), Aqua night clear-sky coverage (ACNC), Aqua day view angle (ADVA), Aqua day view time (ADVT), Aqua Band 31 Emissivity (AE31), Aqua Band 32 Emissivity (AE32), Aqua day land surface temperature (ALSTD), Aqua night land surface temperature (ALSTN), Aqua night view angle (ANVA), Aqua night view time (ANVT), Terra day clear-sky coverage (TCDC), Terra night clear-sky coverage (TCNC), Terra day view angle (TDVA), Terra day view time (TDVT), Terra Band 31 Emissivity (TE31), Terra Band 32 Emissivity (TE32), Terra day land surface temperature (TLSTD), Terra night land surface temperature (TLSTN), Terra night view angle (TNVA), Terra night view time (TNVT), elevation (DEM), Hillshade (HSD), sky view and slope) [1], 10 variables by Hanna Meyer (LST, DEM, slope, aspect, sky view, month, season, time, sensor and ice) [18], 11 variables by Yongming Xu (monthly daytime T_s , monthly nighttime T_s , monthly percent of cloudy days, monthly percent of cloudy nights, monthly NDVI, monthly NDSI, monthly albedo, monthly solar radiation, annual landcover, elevation and TI) [19] and 6 variables by Phan Thanh Noi (AQUA daytime, AQUA nighttime, TERRA daytime and TERRA nighttime), as well as two additional auxiliary datasets (elevation and Julian day)), including 5 variables by Long Li (daytime LST, nighttime LST, NDVI, nighttime light and DEM) [20], 9 variables by Yongming Xu (land surface temperature (LST), normalized difference vegetation index (NDVI), modified normalized difference water index (MNDWI), latitude, longitude, distance to ocean, altitude, albedo and solar radiation) [19], up to 10 variables by Munkhdulam Otgonbayar

(daytime and nighttime LST data (LSTd and LSTn), quality information (QCd and QCn), observation information (DvA, NvA, DvT and NvT), emissivity data (Em31 and Em32), clear-sky coverage (CsD and CsN), elevation, slope, aspect, latitude and longitude) [21], 7 variables by Chunling Wang (digital elevation model (DEM), LST, downward shortwave radiation (DSR), Normalized Difference Vegetation Index (NDVI), land cover (LC), latitude (LAT) and declination of the sun) [22].

In this study, a Partial Least-Squares Regression (PLSR) model was used to assess and predict urban T_{air} considering more than 30 variables. To the best of our knowledge, statistical regression models starting from 30 variables have not been presented elsewhere in the existing literature.

One more difference with the urban study cases described above lies in the complexity of the mountain features with respect to the size of the considered area, and the sparse configuration of weather stations. This study focuses on the city of Yerevan, Armenia, which stands out for unique geographical parameters, in particular for the high variation range of absolute elevation, exceeding 500 m (>1600 ft) on a small area covering about 220 sq km. The area has a dry climate, and only three weather stations are operating, which limits the information on spatial variation of T_{air} . In other case studies from literature, like Athens and Heraklion (Greece), Los Angeles (CA, USA), Seoul (South Korea), Vancouver (Canada) and Erbil (Iraq), where T_{air} predictions were performed using remote sensing data [10,12,16,23], such a combination was never found.

Hence, the objectives of the current study are as follows:

1. Considering the complexity of the terrain configuration in Yerevan, to assess for the first time the feasibility of estimating urban T_{air} based on remote sensing data alone.
2. Estimate the Urban T_{air} of the city of Yerevan using the PLSR model with a high (30) number of input variables.

To the best of our knowledge, so far, no investigations have been performed that could answer these questions.

2. Materials and Methods

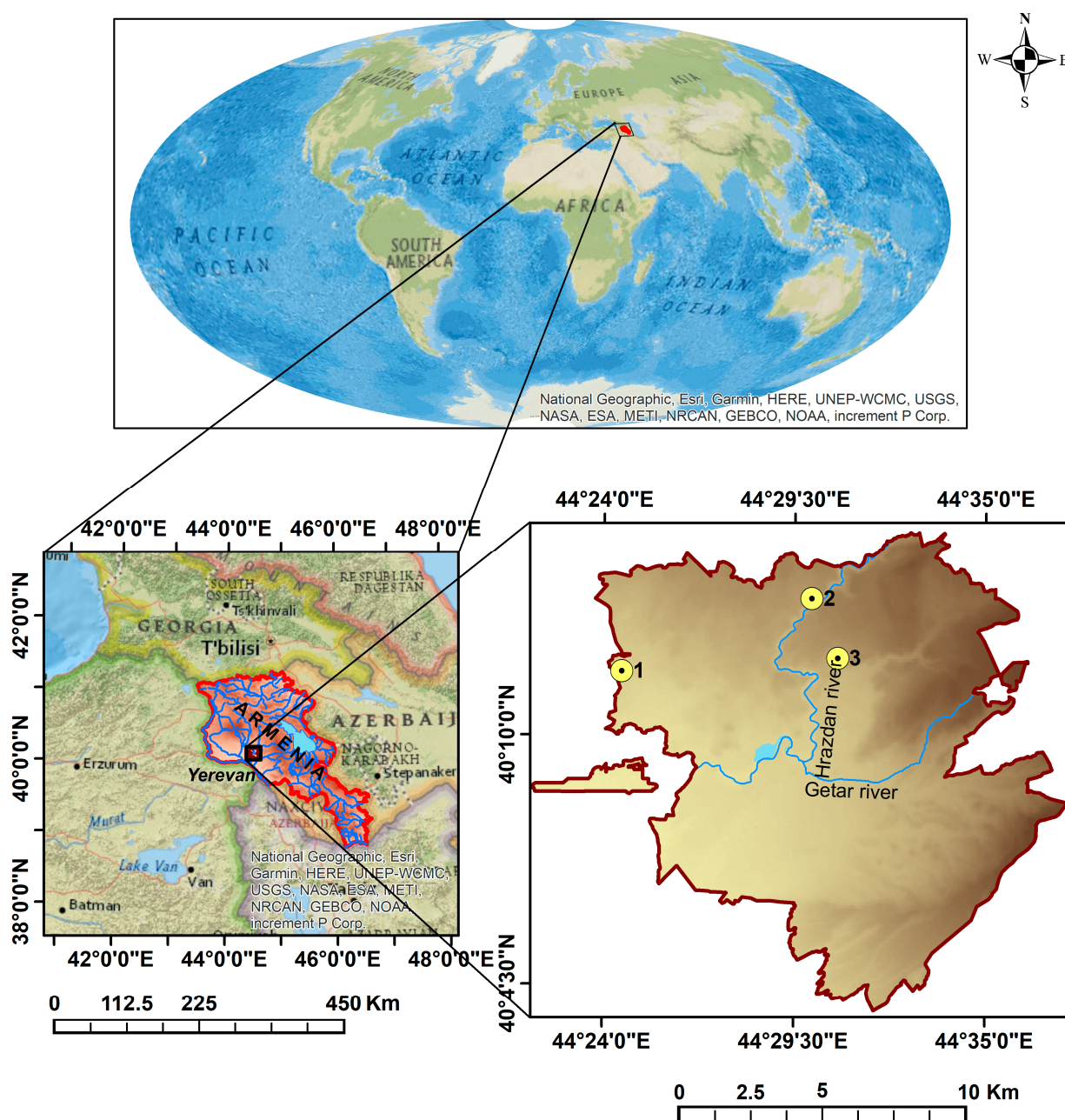
2.1. Test Site, Description and Terrain/Climate Features

Yerevan is the capital of Armenia, covering an area of approximately 220 km² with 1.1 million inhabitants, which represents 36% of the total population and 56% of the urban population. The density of population in Yerevan exceeds 4900 inhabitants/km² [24].

Yerevan lies on a plain on the edge of the Ararat Valley at altitudes of 850–1400 m (Figure 1). It has a dry continental climate. The average annual T_{air} is between 9.1 °C and 12.1 °C, with a seasonal fluctuation of 27 °C between average summer and winter temperatures. Winters are cold with a lot of snowfall and average temperatures in January ranging between −5 °C and −2.5 °C, with the absolute minimum T_{air} between −21 °C and −31 °C. Springs are brief, characterized by volatile weather. Summers are long, hot and dry, with an average T_{air} between 22.1 and 25.4 °C. The absolute maximums of T_{air} registered in July are between 40 °C and 43 °C [25,26]. The study area is located in the dry subtropical climate zone. Thus, climate change is especially expressed in increasing amplitude of urban T_{air} swings, as discussed later.

During summer, winds blowing from the mountains (north-east) to the valley (south-west) sometimes reach a speed of 15–20 m/s. The duration of the heating season is between 137 and 161 days. Annual rainfall is 286–440 mm, peaking in November, while the highest share of rainy days is in May. Yerevan also enjoys a lot of sunshine. The annual average of sunshine is 2578 h. Hours of sunshine per day will vary from an average of 7 h in winter to 13 h in summer [26].

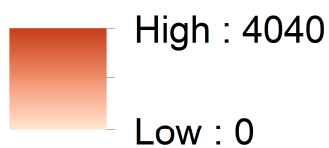
Figure 1 shows the geographical location of Yerevan on the territory of Armenia reflecting the mountainous character of the surface. The figure also shows an unequal distribution of weather stations, which are located in the north and west of the city on different altitudes (see Table 1). As mentioned above, this pattern limits the possibility to observe spatial variations of T_{air} .



LEGEND

- Meteorological stations
- Rivers
- Lakes
- Armenia
- Yerevan

Digital Elevation
Model of Armenia (m)



Digital Elevation
Model of Yerevan (m)

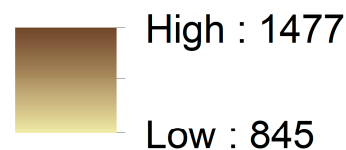


Figure 1. Geographical location and hypsometry of Armenia and Yerevan and the distribution of the weather stations on territory of Yerevan: (1) Yerevan_agro; (2) Yerevan_aerologia; (3) Arabkir.

Table 1. Geographical coordinates and altitude of the weather stations operating on the territory of Yerevan.

N	Name of Station	Latitude	Longitude	Height a. s. l. (m)
1.	Yerevan_agro	40°11'19"N	44°23'55"E	942
2.	Yerevan_aerologia	40°13'2"N	44°29'59"E	1134
3.	Arabkir	40°11'43"N	44°30'44"E	1113

Since the late 1980s, land cover in Yerevan has been changing rapidly, which results in potential sharpening of the natural conditions of the city on the recent climate changes. The last national communication on climate change shows that, from 1981 to 2013, the summer heat wave in Yerevan has increased about 40 days on average [27,28]. Some studies were conducted to investigate the reasons. For instance, Tepanosyan et al. investigated the influence of spatial-temporal changes of land cover on the territory of Yerevan city on the surface urban heat, using time series of remote sensing data (Landsat TM/ETM+/OLI-TIRS images) [26]. However, no studies have been conducted so far on developing approaches to enhance visualization and to monitor spatial-temporal variation of Tair using remote sensing data and technologies for this area.

2.2. Preparation of Input Data

2.2.1. Satellite Data

Figure 2 shows the steps of the study. The input data consist of open-source remote sensing (RS) surface reflectance products from LANDSAT 4-5TM, 7ETM and 8 OLI/TIRS, covering the season from June to August for years between 1984 and 2020 and obtained and directly processed in Google Earth Engine (GEE). Two gaps can be observed in the time sequence in 2003 and 2012; the corresponding data were discarded due to low quality. Cloud mask and cloud filtering were implemented using the CFMASK algorithm, as well as a per-pixel saturation mask [29]. Spectral indices also were calculated to complement the list of input data, which consisted of Normalized Difference Vegetation Index—NDVI, Normalized Difference Water Index—NDWI and the Index-Based Built-Up Index and Soil-Adjusted Vegetation Index (IBI-SAVI) (1).

$$\text{IBI-SAVI} = (((\text{NDBI} + 1) - ((\text{SAVI} + 1) + (\text{MNDWI} + 1))/2))/((\text{NDBI} + 1) + ((\text{SAVI} + 1) + (\text{MNDWI} + 1))/2) \quad (1)$$

where

$$\text{NDVI} = (\text{NIR} - \text{RED})/(\text{NIR} + \text{RED}) \quad (2)$$

NIR—Spectral reflectance in near infrared region

RED—Spectral reflectance in the red region

$$\text{SAVI} = ((\text{NIR} - \text{R})/(\text{NIR} + \text{RED} + \text{L})) * (1 + \text{L}) \quad (3)$$

NIR—Spectral reflectance in near infrared region

RED—Spectral reflectance in the red region

L—Soil brightness correction factor

$$\text{MNDWI} = (\text{GREEN} - \text{SWIR})/(\text{GREEN} + \text{SWIR}) \quad (4)$$

GREEN—Spectral reflectance in green region

SWIR—Spectral reflectance in the short-wave infrared region

$$\text{NDBI} = (\text{SWIR} - \text{NIR})/(\text{SWIR} + \text{NIR}) \quad (5)$$

SWIR—spectral reflectance in the short-wave infrared region

NIR—spectral reflectance in near infrared region

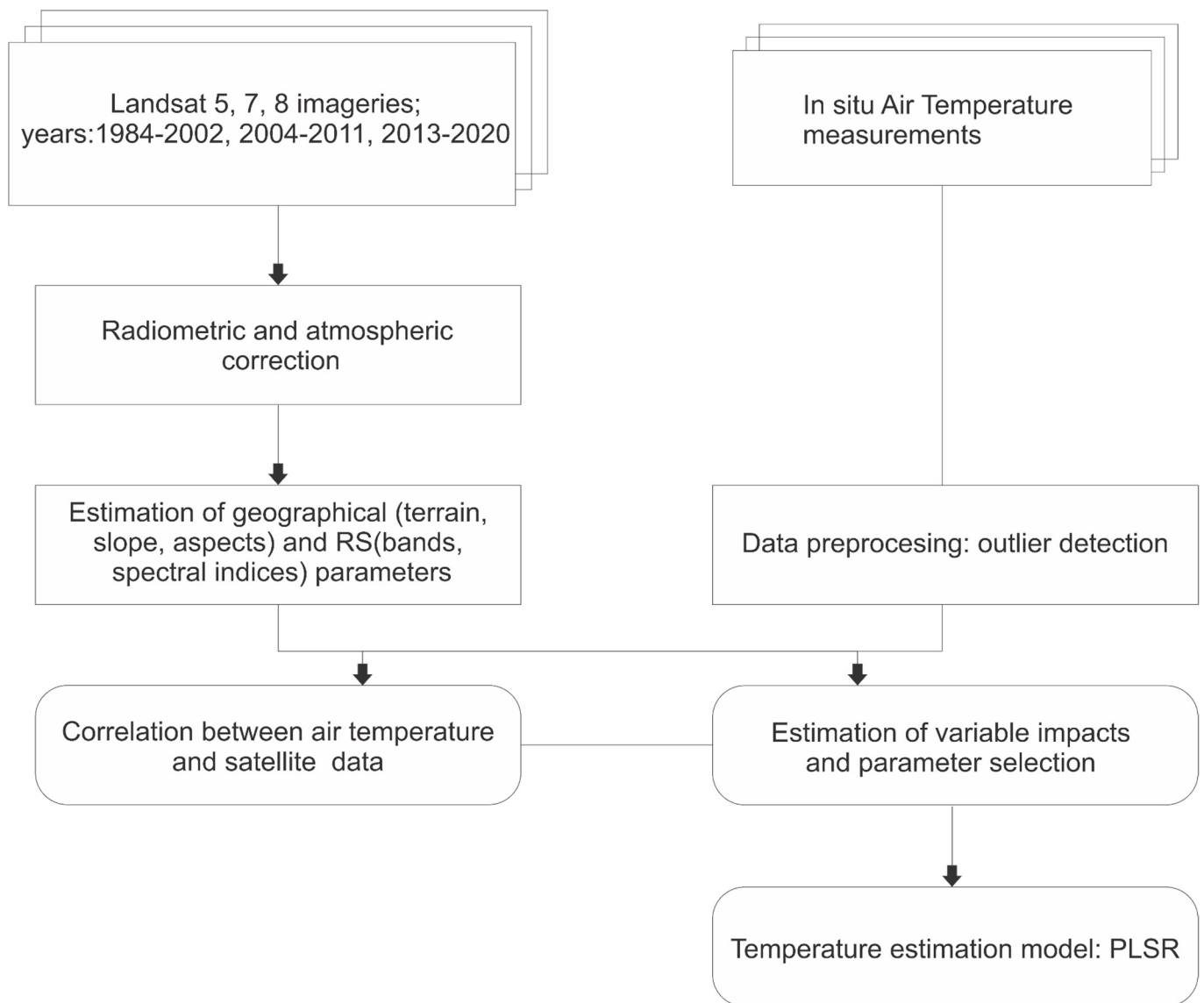


Figure 2. Methodological flowchart of the study.

The IBI-SAVI is a combined indicator, which we propose for use in this context. It is calculated from Normalized Difference Built-Up Index—NDBI, SAVI and Modified Normalized Difference Water Index—MNDWI. The formula was modified partly following suggestions by Xu [30], and introducing further changes in rescaling SAVI to match the other rescaled indexes.

The LST for each year as an input data was calculated using Landsat LST Web Application. This online web application provides fast and easy access to the global scale LST from the Landsat archives based on the single channel (SC) algorithm [26,31–33]. Input data also contain geographical data such as elevation and its derivatives (aspects, slopes). Solar radiation was also modeled from DEM and the sun location using the area solar radiation toolset of ArcMap. The elevation ruggedness is calculated as Terrain Ruggedness Index, which is used to characterize the elevation difference of the DEM’s adjacent cells [34]. Several single bands (RGB, NIR, SWIR1/2) of RS data complement the list of the input data.

Weather data (T_{air} , cloud cover, dew point, wind direction, wind speed, solar radiation) were acquired from Armenia State Hydromet Service for three weather stations (Figure 1). We decided not to include precipitation as it is reputed to have a delayed effect on temperature, and it would not introduce additional relevant information. The dates of

measurements were selected to match the dates of RS data acquisition as closely as possible. However, only two stations cover the whole considered time period (1984–2020); the third station (Yerevan-Aerology) covers only years between 2011 and 2020.

As Tepanosyan et al. stated, several approaches are possible when studying the relationship between RS data and climatic factors [26]. According to the first approach, maps of the several climatic parameters should be produced using interpolation methods, which requires data from a sufficient number of weather stations. In our case, the terrain shape and very few weather stations make this approach not viable [35–38]. The second approach implies studying relationships between climatic data from weather stations and average spectral index values obtained from pixels surrounding the respective stations [39,40]. In this case study, the second approach was used, and mean NDVI, NDWI and IBI-SAVI values across a station-centered circular area were extracted; using multitemporal data, a time series was formed for each station. Many studies accepted a 3×3 pixel window size as the optimal one for deriving average values of the spectral indices such as NDVI [39,40]; in this study, the averaged values were instead calculated over differently sized circular windows around the weather stations. The rationale was to investigate the spatial dependence of the parameter impact; the selected sizes were 30 m, 100 m, 200 m, 300 m, 400 m, 500 m, 600 m, 700 m, 800 m, 900 m and 1000 m. A further increase of the radius results in an overlap of the areas between the two weather stations.

2.2.2. Weather Data

As explained in Section 2, we considered weather data from three weather stations located in or around Yerevan. Weather data were provided by the “Hydrometeorology and monitoring center” State Non-Commercial Organization (SNCO) of the Ministry of Environment of Armenia. Three “MicroStep-MIS” stations working with a full program provide following data on daily and hourly basis: horizontal visibility, cloud cover, atmospheric phenomena, soil temperature at the surface and at different depths, air temperature, air humidity, atmospheric pressure, wind direction and intensity, precipitation, sunshine duration, wind direction and intensity, dew point, solar radiation, etc. In this paper, we limited ourselves to considering T_{air} investigations about other possible weather variables, which will be the subject of the future work.

Initially, we considered all measurements, whereas at a successive stage, outliers were removed before reprocessing.

In order to detect potential outliers among data, the boxplot technique was used: parameter values that were outside the (Quartile 1– $1.5 \times$ Interquartile range) – (Quartile 3 + $1.5 \times$ Interquartile) range were considered as outlier candidates. However, only acquisitions labeled as outlier candidates on all parameters were finally considered outliers and removed. All other candidates, i.e., outside the range but not on all variables, were still retained. This was done to ensure that only data items that could be considered outliers with a high degree of confidence were removed.

3. Statistical Analysis and Modeling

To study the relationships between T_{air} and satellite data, a Pearson correlation analysis including significance estimation was first carried out to identify the best candidates to contribute to T_{air} estimation. For this purpose, the calculated mean and standard deviation (SD) values of the components/variables were input. At this level, however, no actual parameter selection was completed, leaving it to an automated process to be implemented at a later stage.

All statistical analyses were conducted using the Python programming language using Jupyter. Linear regression (LR) was implemented by the Scikit learn algorithm “LinearRegression” [41]. T_{air} prediction was treated as a supervised regression problem. Therefore, the Partial Least-Squares Regression (PLSR) was selected in this research as the statistical approach for evaluation. For all models, the input dataset was randomly split into training (75%) and testing (25%) sets.

PLSR regression was run considering various possible combinations of the input parameters, starting from a single variable and progressively increasing the number of variables. The Variable Importance in Projection (VIP) scores were used to prioritize selections of input variables. VIP scores estimate the importance of each variable in the projection used in a PLSR model and are often used for variable selection. A variable with a VIP score close to or greater than 1 (one) can be considered important in a given model [21,42].

Each combination was tested for mean square errors in predicting the training set. It was found that the optimal MSE values were obtained for 10 variables.

Further expansion of the input set did not lead to any improvement in MSE, which actually worsened. The input set to the prediction process was then set to the list above, and predictions were compared with the test set.

4. Results and Discussions

Pearson values are reported in Table 2. The reported figures suggest that LST-mean has the most significant influence ($r = 0.79$; $p < 0.001$) on T_{air} . All other components, such as IBI-SAVI-mean ($r = 0.35$; $p < 0.001$), SWIR1-mean ($r \approx 0.3$; $p < 0.001$), SWIR2I-mean ($r \approx 0.3$; $p < 0.001$) and Red-mean ($r \approx 0.3$; $p < 0.001$), show a significant positive correlation. Green-mean, Blue-means and Aspect_SD at ($r \approx 0.2$; $p < 0.001$), ($r \approx 0.14$; $p < 0.01$) and ($r \approx 0.11$; $p < 0.01$), respectively, also show a significant positive correlation with T_{air} . Some other components, such as NDWI_mean ($r = -0.35$; $p < 0.001$), NDVI_mean ($r = -0.25$; $p < 0.001$), NDWI_SD ($r = -0.24$; $p < 0.001$), NDVI_SD ($r = -0.26$; $p < 0.001$), IBI SAVI_SD and SWIR2_SD ($r = -0.13$; $p < 0.001$), show a significant negative correlation with T_{air} . In fact, two components contribute the most information to the temperature prediction, and one of them is the modified IBI-SAVI index. This will be compared with the results of the automated selection, as explained in the following.

Table 2. Pearson correlations between T_{air} and all components. Bold indicates the strongest correlation found among all analyzed variables.

N	Variables	Correlation Coefficient (r)	p_Value
1.	Blue_mean	0.14	2×10^{-3}
2.	Green_mean	0.16	4×10^{-4}
3.	Red_mean	0.26	3×10^{-9}
4.	NIR_mean	0.01	8×10^{-1}
5.	SWIR1_mean	0.29	7×10^{-11}
6.	SWIR2_mean	0.30	9×10^{-12}
7.	NDVI_mean	-0.25	1×10^{-8}
8.	NDWI_mean	-0.35	2×10^{-15}
9.	IBI SAVI_mean	0.35	1×10^{-15}
10.	LST_mean	0.79	1×10^{-15}
11.	Aspect_mean	-0.07	1×10^{-1}
12.	Slope_mean	-0.06	2×10^{-1}
13.	Elev_mean	-0.08	7×10^{-2}
14.	Rugged_mean	-0.06	2×10^{-1}
15.	Sol_rad_mean	-0.19	1×10^{-5}
16.	Blue_SD	-0.10	3×10^{-2}
17.	Green_SD	-0.09	4×10^{-2}
18.	Red_SD	-0.07	1×10^{-1}
19.	NIR_SD	-0.07	1×10^{-1}
20.	SWIR1_SD	-0.09	4×10^{-2}
21.	SWIR2_SD	-0.13	3×10^{-3}
22.	NDVI_SD	-0.26	2×10^{-9}
23.	NDWI_SD	-0.24	5×10^{-8}
24.	IBI SAVI_SD	-0.12	5×10^{-3}
25.	LST_SD	-0.01	9×10^{-1}

Table 2. Cont.

N	Variables	Correlation Coefficient (r)	p_Value
26.	Aspect_SD	0.11	1×10^{-2}
27.	Slope_SD	−0.03	6×10^{-1}
28.	Elev_SD	−0.10	3×10^{-2}
29.	Rugged_SD	−0.02	6×10^{-1}
30.	Sol_rad_SD	0.04	4×10^{-1}

As mentioned above, the spatial dependence of the parameter impact was investigated for the selected areas sized 30 m, 100 m, 200 m, 300 m, 400 m, 500 m, 600 m, 700 m, 800 m, 900 m and 1000 m through PLSR estimation. The table below shows the estimation results for the all the sized areas. Prior to the PLSR run, the estimation of variable impacts and the selection of the parameters (VIP) were conducted.

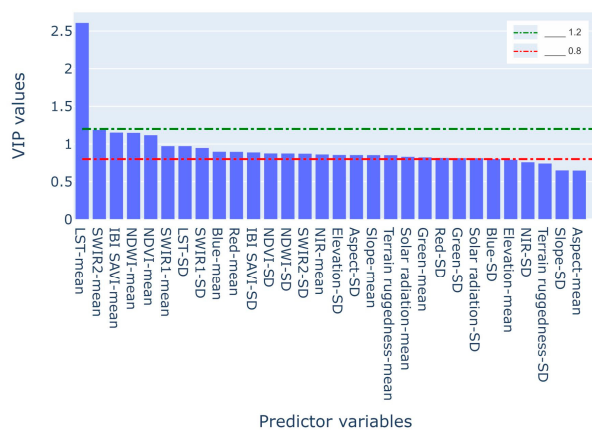
The estimated importance of the 30 predictor variables in the PLS regression model for all the sized buffer zones is shown in the Figure 3.

As mentioned above, the selection of the sizes was stopped at 1000 m because the further increase of the radius results in an overlap of the areas between two weather stations.

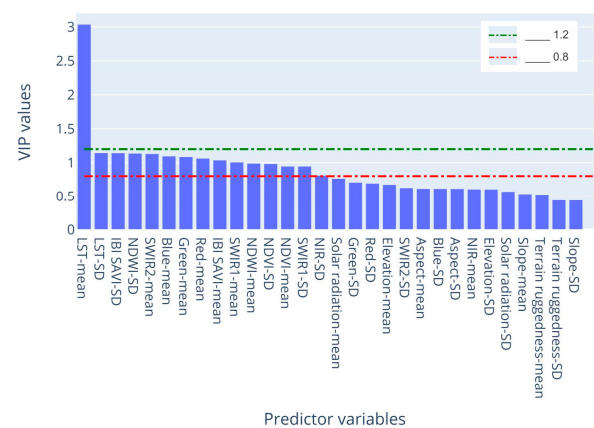
As seen in Table 3, the number of VIP components varies when increasing the radius of the circles around the weather stations. Table 3 shows that the quantity of the components (predictor variables) stabilized.

Table 3. The PLSR descriptive for the different sized areas with VIP scores.

PLSR Descriptive	30 m	100 m	200 m	300 m	400 m	500 m	600 m	700 m	800 m	900 m	1000 m
R^2_{Train}	0.72	0.73	0.73	0.75	0.75	0.74	0.75	0.75	0.75	0.76	0.76
$\text{RMSE}_{\text{Train}}$	1.68	1.66	1.65	1.58	1.58	1.61	1.60	1.59	1.59	1.57	1.56
R^2_{CV}	0.68	0.68	0.68	0.71	0.71	0.70	0.71	0.70	0.70	0.71	0.72
RMSE_{CV}	1.80	1.79	1.80	1.70	1.70	1.73	1.73	1.74	1.74	1.71	1.67
R^2_{Test}	0.70	0.71	0.71	0.73	0.72	0.73	0.72	0.74	0.74	0.75	0.77
$\text{RMSE}_{\text{Test}}$	1.78	1.73	1.75	1.69	1.71	1.69	1.72	1.65	1.65	1.61	1.58
N of VIP components	14	14	10	14	14	13	14	10	10	10	10

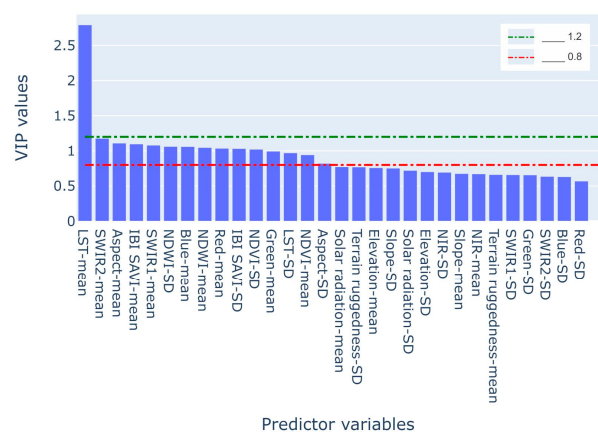


(a)

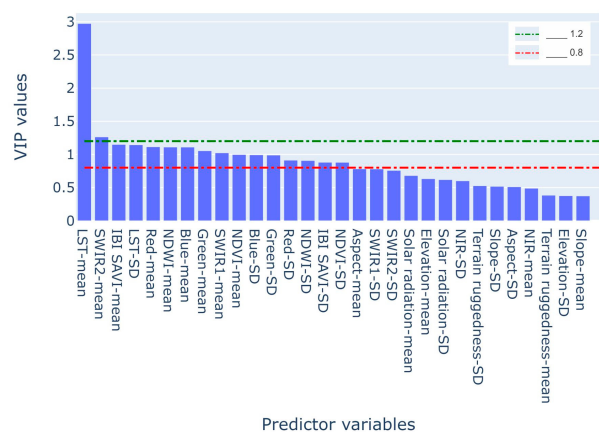


(b)

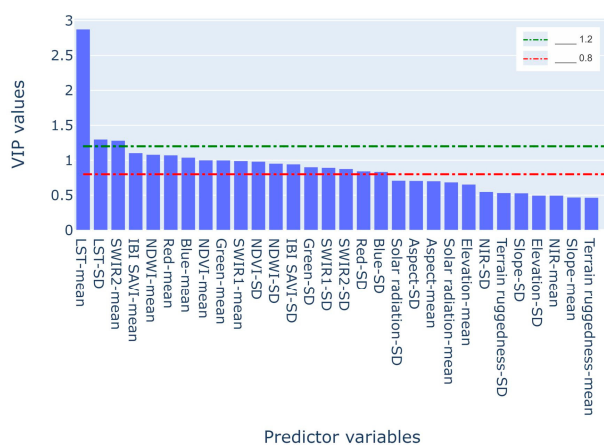
Figure 3. Cont.



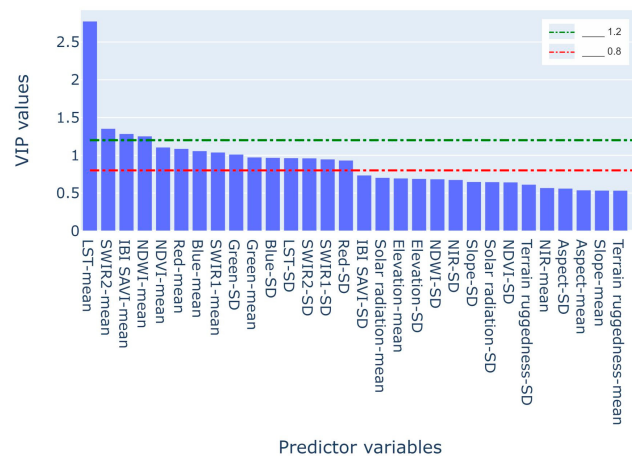
(c)



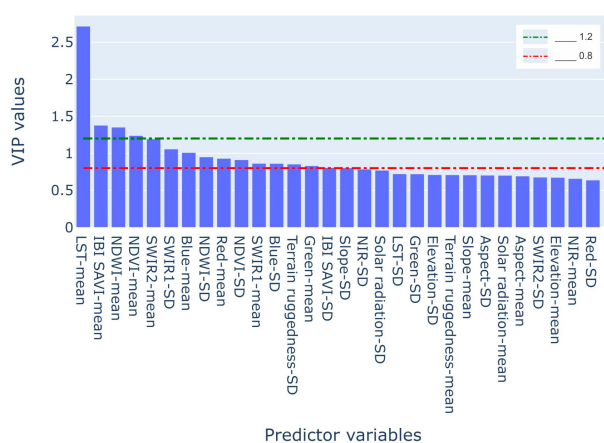
(d)



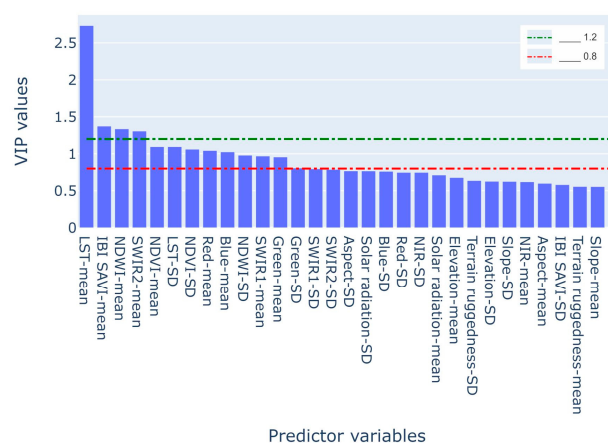
(e)



(f)



(g)



(h)

Figure 3. Cont.

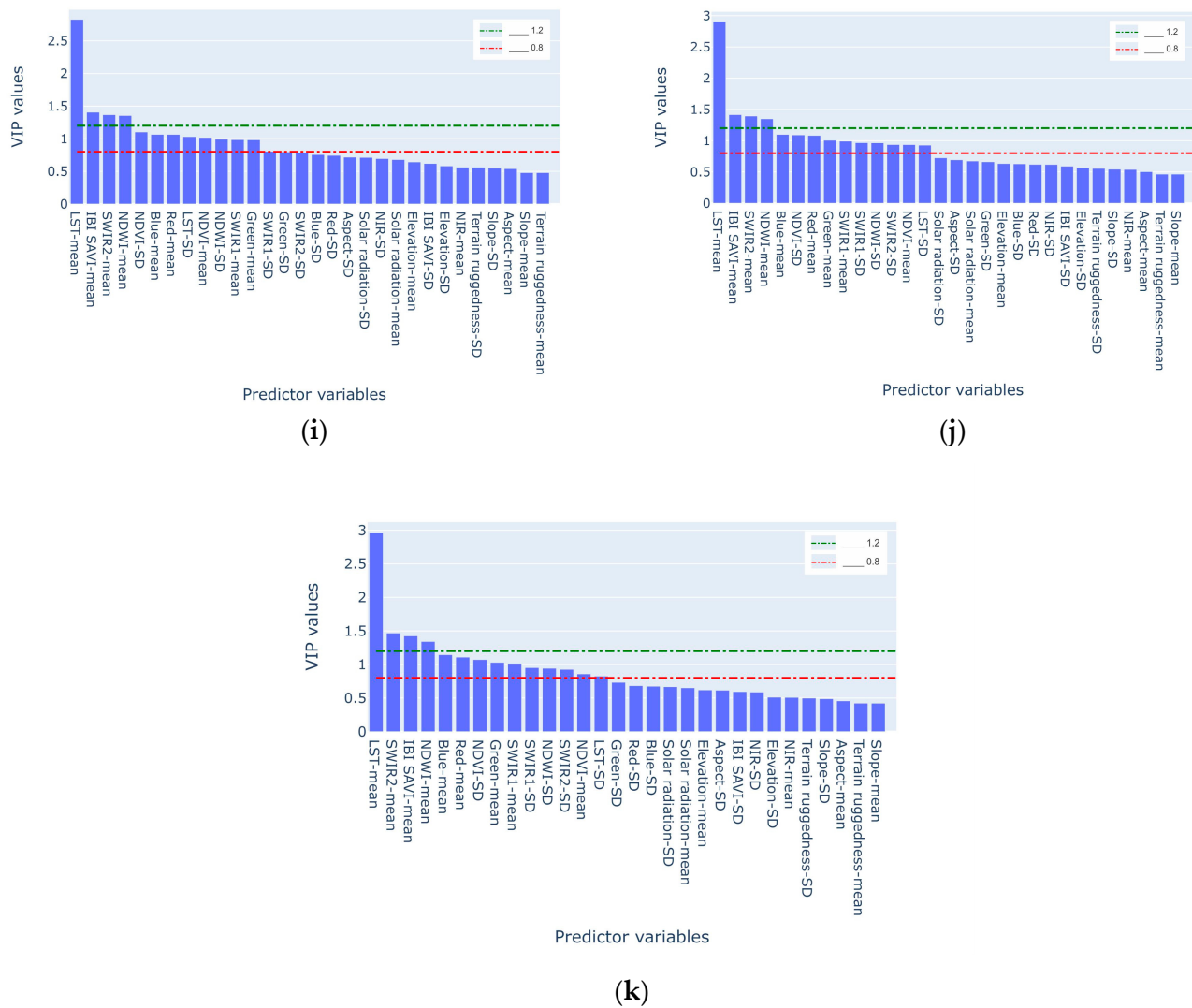


Figure 3. PLSR variable importance for each circled zone: (a) 30 m; (b) 100 m; (c) 200 m; (d) 300 m; (e) 400 m; (f) 500 m; (g) 600 m; (h) 700 m; (i) 800 m; (j) 900 m; (k) 1000 m.

The VIP scores for the 1000 m buffer zone are shown in Table 4. The LST-mean has the highest VIP score (2.77). According to Table 4, the following variables feature VIP scores greater than 1 (one): SWIR2_mean (1.42), IBI SAVI-mean (1.29) and NDWI-mean (1.23). Blue-mean, red-mean and SWIR1_mean show scores of approximately 1.1, and NDVI_SD has a VIP score of 1.0. All the others are close to or below 1.0.

Table 4. The list of the variables with VIP scores.

N	Predictor Variables	VIP Scores
1.	Blue-mean	1.113
2.	Green-mean	1.019
3.	Red-mean	1.098
4.	NIR-mean	0.683
5.	SWIR1-mean	1.067
6.	SWIR2-mean	1.415
7.	NDVI-mean	0.8985
8.	NDWI-mean	1.225
9.	IBI SAVI-mean	1.285
10.	LST-mean	2.772

Table 4. Cont.

N	Predictor Variables	VIP Scores
11.	Aspect-mean	0.661
12.	Slope-mean	0.667
13.	Elevation-mean	0.643
14.	Terrain ruggedness-mean	0.666
15.	Solar radiation-mean	0.661
16.	Blue-SD	0.730
17.	Green-SD	0.692
18.	Red-SD	0.658
19.	NIR-SD	0.712
20.	SWIR1-SD	0.910
21.	SWIR2-SD	0.894
22.	NDVI-SD	1.011
23.	NDWI-SD	0.923
24.	IBI SAVI-SD	0.586
25.	LST-SD	0.960
26.	Aspect-SD	0.746
27.	Slope-SD	0.651
28.	Elevation-SD	0.710
29.	Terrain ruggedness-SD	0.650
30.	Solar radiation-SD	0.688

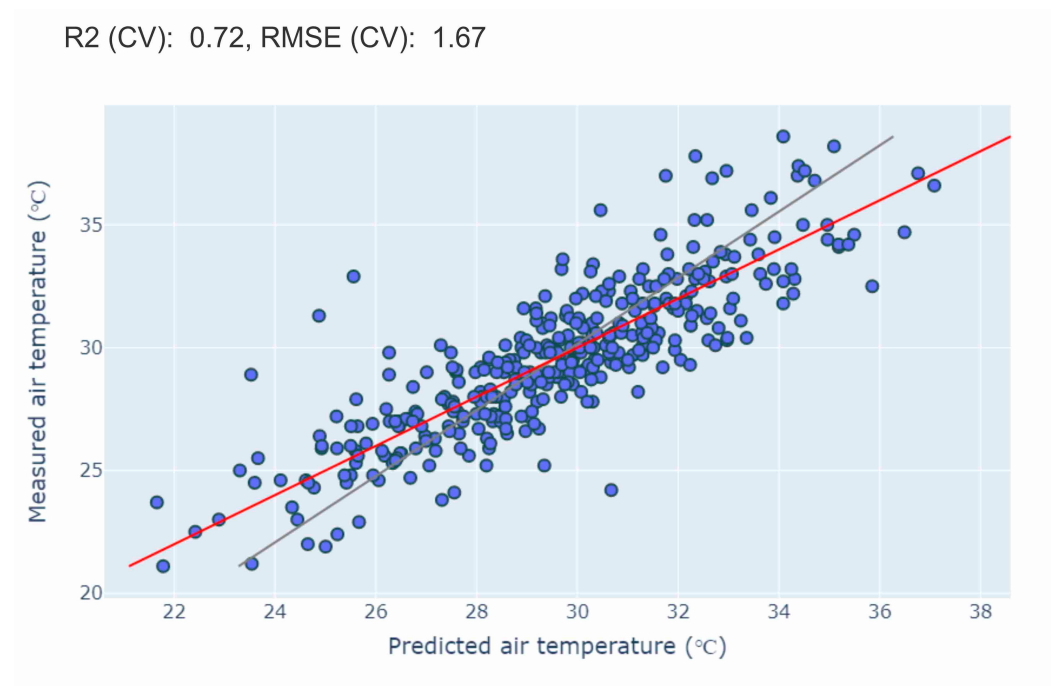
The results of the PLSR model received for the 1000 m buffer zone are shown in Figure 4. As it can be seen, the PLSR model provides satisfactory results both for calibration ($R^2_{\text{Cal}} = 0.72$, $\text{RMSE}_{\text{Cal}} = 1.67$) and validation ($R^2_{\text{Val}} = 0.77$, $\text{RMSE}_{\text{Val}} = 1.58$).

For comparison, it should be noted that in a number of other studies the errors in the daily T_{air} estimation generally fall in the range between 2 and 3 °C [43]. In particular, where PLSR was applied with the Leave-One-Out Cross-Validation method, $\text{RMSE} = 2.71$ °C, $R^2 = 0.83$ [1] and $\text{RMSE} = 2.1\text{--}3.6$ °C [44]. The results by Otgonbayar et al. in 2019 show that R^2 varies between 0.74 and 0.87 and RMSE varies from 1.20 °C to 2.19 °C [21].

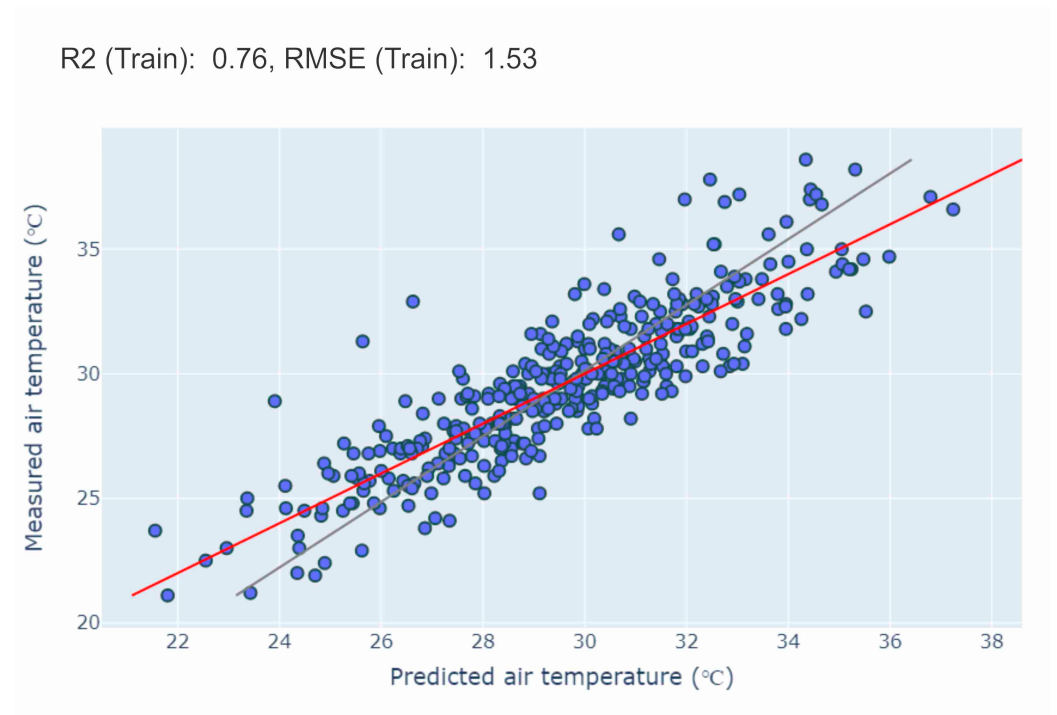
In this work, using PLSR driven by a wide range of predictor variables (30), values of T_{air} on an area with complex terrain features such as Yerevan (Armenia) were predicted with a high accuracy of $\text{RMSE}_{\text{Val}} = 1.58$ °C. In the process, it was noticed that 5 predictor variables of the selected 10 with high VIP scores also feature comparatively high (LST-mean: $r = 0.79$; $p < 0.001$) correlation coefficients (IBI-SAVI-mean: $r = 0.35$; $p < 0.001$; SWIR1-mean and Red-mean: $r \approx 0.3$; $p < 0.001$). However, this shows that, among these five variables, Landsat-derived land surface temperature plays a key role in modeling T_{air} , with all other variables having a significantly smaller impact. The studies of Otgonbayar et al. concluded that PLSR even represents seasonal and spatial variations in T_{air} when the time series of LST was included as the predictor variables [21].

The results of the importance analysis highlighted a pool of parameters, which have the biggest impact on T_{air} . The heterogeneity of the area makes it particularly difficult to venture guesses on the reasons for the composition of such pool of variables.

Previous studies on estimating urban T_{air} from remote sensing data were performed using more advanced ML models, such as Random Forest, Cubist and Support Vector Machine (SVM), as well as neural networks to estimate urban air temperature [12,15,44]. Though we saw the great potential of the remote sensing data to estimate the T_{air} on Yerevan's territory, there is still a strong need to continue the studies using above-mentioned advanced ML models.



(a)



(b)

Figure 4. *Cont.*

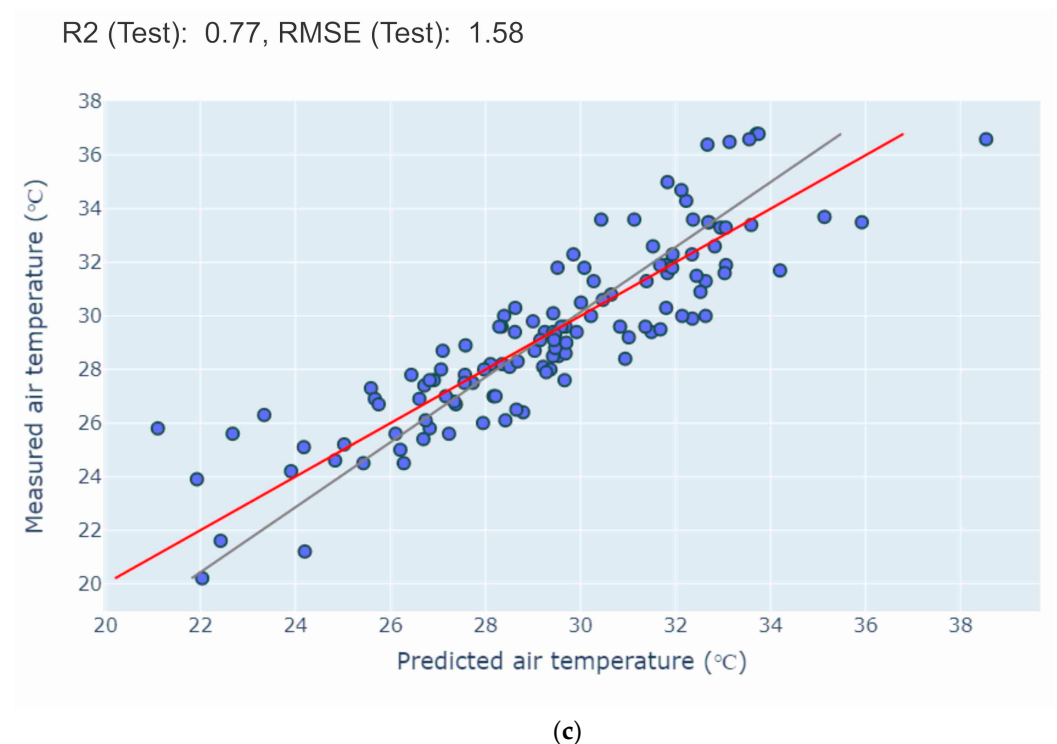


Figure 4. Scatter plot of predicted vs. measured T_{air} contents when validating the PLSR model ($n = 500$): (a) cross validation; (b) training (75%); (c) testing (25%).

5. Conclusions

The main purpose of the research described in this paper was to assess the feasibility of estimating urban T_{air} , in a complex terrain configuration, based on remote sensing data alone using the PLSR model with a high amount of input variables. The novelty of this study includes the features of the considered area, which is complex and with a broad distribution of different elevations, and the high number of environmental parameters considered, exceeding 30. The key findings are outlined below:

1. Of the 30 parameters considered, 10 can be identified as relevant and can be used alone in the prediction; adding more parameters will not improve prediction, but will require more computational resources.
2. The relevant parameters include a newly proposed modification of index IBI-SAVI, which turned out to strongly impact T_{air} prediction.
3. Cross-validation analysis on temperature predictions across a station-centered 1000 m circular area revealed quite a high correlation ($R^2_{\text{Val}} = 0.77$, $\text{RMSE}_{\text{Val}} = 1.58$) between the predicted and measured T_{air} from the test set.
4. In light of the above, we may estimate that remote sensing is an effective tool to estimate T_{air} distribution where a dense network of weather stations is not available.

Future developments will include incorporation of additional weather parameters from the weather stations, such as precipitation and wind speed, and the use of non-parametric machine learning (ML) techniques, whose structure may be more suitable to represent the complex link between observables and target parameters in a complex environment like the one considered in this study.

Author Contributions: Conceptualization, S.A., G.T., V.M. and F.D.; methodology, S.A., G.T., V.M. and F.D.; data analysis and visualization, G.T., R.A., A.K. and A.H.; writing—review and editing, S.A., V.M., G.A. and F.D. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the Science Committee of the Ministry of Education Science Culture and Sport of RA, in the framework of research project No. 20TTTCG-1E009. The work was also partly funded by the project NODES, which received funding from the MUR-M4C2 1.5 of PNRR with grant agreement no. ECS00000036.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge Department of GIS and Remote Sensing of the Center for Ecological-Noosphere Studies of the National Academy of Sciences (Armenia) and for the existing facilities to conduct this research and to the Department of Electrical, Computer and Biomedical Engineering, University of Pavia. I particular, the authors thank Fabio Dell’Acqua for the dedicated collaboration on the exchange of the experience when conducting the research and for support when preparing the manuscript. This study was fully supported by the Science Committee of the Ministry of Education, Science, Sport and Culture of RA in the frames of the research project No. 20TTTCG-1E009.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Meliho, M.; Khattabi, A.; Zejli, D.; Orlando, C.A.; Dansou, C.E. Artificial Intelligence and Remote Sensing for Spatial Prediction of Daily Air Temperature: Case Study of Souss Watershed of Morocco. *Geo-Spat. Inf. Sci.* **2022**, *25*, 244–258. [\[CrossRef\]](#)
2. Ding, L.; Zhou, J.; Zhang, X.; Liu, S.; Cao, R. Downscaling of Surface Air Temperature over the Tibetan Plateau Based on DEM. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *73*, 136–147. [\[CrossRef\]](#)
3. Shah, D.B.; Pandya, M.R.; Trivedi, H.J.; Jani, A.R. Estimating Minimum and Maximum Air Temperature Using MODIS Data over Indo-Gangetic Plain. *J. Earth Syst. Sci.* **2013**, *122*, 1593–1605. [\[CrossRef\]](#)
4. Nichol, J.E.; To, P.H. Temporal Characteristics of Thermal Satellite Images for Urban Heat Stress and Heat Island Mapping. *ISPRS J. Photogramm. Remote Sens.* **2012**, *74*, 153–162. [\[CrossRef\]](#)
5. Fu, P.; Weng, Q. Variability in Annual Temperature Cycle in the Urban Areas of the United States as Revealed by MODIS Imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 65–73. [\[CrossRef\]](#)
6. Vogt, J.V.; Viau, A.A.; Paquet, F. Mapping Regional Air Temperature Fields Using Satellite-Derived Surface Skin Temperatures. *Int. J. Climatol.* **1997**, *17*, 1559–1579. [\[CrossRef\]](#)
7. Zakšek, K.; Schroedter-Homscheidt, M. Parameterization of Air Temperature in High Temporal and Spatial Resolution from a Combination of the SEVIRI and MODIS Instruments. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 414–421. [\[CrossRef\]](#)
8. Cristóbal, J.; Ninyerola, M.; Pons, X. Modeling Air Temperature through a Combination of Remote Sensing and GIS Data. *J. Geophys. Res.* **2008**, *113*, D13106. [\[CrossRef\]](#)
9. Mutiibwa, D.; Strachan, S.; Albright, T. Land Surface Temperature and Surface Air Temperature in Complex Terrain. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4762–4774. [\[CrossRef\]](#)
10. Nikoloudakis, N.; Stagakis, S.; Mitraka, Z.; Kamarianakis, Y.; Chrysoulakis, N. Spatial Interpolation of Urban Air Temperatures Using Satellite-Derived Predictors. *Appl. Clim.* **2020**, *141*, 657–672. [\[CrossRef\]](#)
11. Orellana-Samaniego, M.L.; Ballari, D.; Guzman, P.; Ospina, J.E. Estimating Monthly Air Temperature Using Remote Sensing on a Region with Highly Variable Topography and Scarce Monitoring in the Southern Ecuadorian Andes. *Appl. Clim.* **2021**, *144*, 949–966. [\[CrossRef\]](#)
12. Yoo, C.; Im, J.; Park, S.; Quackenbush, L.J. Estimation of Daily Maximum and Minimum Air Temperatures in Urban Landscapes Using MODIS Time Series Satellite Data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *137*, 149–162. [\[CrossRef\]](#)
13. Cifuentes, J.; Marulanda, G.; Bello, A.; Reneses, J. Air Temperature Forecasting Using Machine Learning Techniques: A Review. *Energies* **2020**, *13*, 4215. [\[CrossRef\]](#)
14. Bechtel, B.; Zakšek, K.; Oßenbrügge, J.; Kaveckis, G.; Böhner, J. Towards a Satellite Based Monitoring of Urban Air Temperatures. *Sustain. Cities Soc.* **2017**, *34*, 22–31. [\[CrossRef\]](#)
15. Ho, H.C.; Knudby, A.; Sirovyak, P.; Xu, Y.; Hodul, M.; Henderson, S.B. Mapping Maximum Urban Air Temperature on Hot Summer Days. *Remote Sens. Environ.* **2014**, *154*, 38–45. [\[CrossRef\]](#)
16. Agathangelidis, I.; Cartalis, C.; Santamouris, M. Estimation of Air Temperatures for the Urban Agglomeration of Athens with the Use of Satellite Data. *Geoinform. Geostat. Overv.* **2016**, *4*, 1–7. [\[CrossRef\]](#)
17. Ho, H.C.; Knudby, A.; Xu, Y.; Hodul, M.; Aminipouri, M. A Comparison of Urban Heat Islands Mapped Using Skin Temperature, Air Temperature, and Apparent Temperature (Humidex), for the Greater Vancouver Area. *Sci. Total Environ.* **2016**, *544*, 929–938. [\[CrossRef\]](#)
18. Meyer, H.; Pebesma, E. Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods Ecol. Evol.* **2021**, *12*, 1620–1633. [\[CrossRef\]](#)
19. Xu, Y.; Shen, Y. Reconstruction of the Land Surface Temperature Time Series Using Harmonic Analysis. *Comput. Geosci.* **2013**, *61*, 126–132. [\[CrossRef\]](#)

20. Noi, P.T.; Degener, J.; Kappas, M. 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 Sens.* **2017**, *9*, 398. [CrossRef]
21. Otgonbayar, M.; Atzberger, C.; Mattiuzzi, M.; Erdenedalai, A. Estimation of Climatologies of Average Monthly Air Temperature over Mongolia Using MODIS Land Surface Temperature (LST) Time Series and Machine Learning Techniques. *Remote Sens.* **2019**, *11*, 2588. [CrossRef]
22. Wang, C.; Bi, X.; Luan, Q.; Li, Z. Estimation of Daily and Instantaneous Near-Surface Air Temperature from MODIS Data Using Machine Learning Methods in the Jingjinji Area of China. *Remote Sens.* **2022**, *14*, 1916. [CrossRef]
23. Rasul, A.; Balzter, H.; Smith, C. Applying a Normalized Ratio Scale Technique to Assess Influences of Urban Expansion on Land Surface Temperature of the Semi-Arid City of Erbil. *Int. J. Remote Sens.* **2017**, *38*, 3960–3980. [CrossRef]
24. Statistical Committee of the Republic of Armenia. Available online: <https://www.armstat.am/en/> (accessed on 13 March 2023).
25. Yerevan Green City Action Plan. Available online: <https://www.yerevan.am/en/yerevan-green-city-action-plan/> (accessed on 13 March 2023).
26. Tepanosyan, G.; Muradyan, V.; Hovsepyan, A.; Pinigin, G.; Medvedev, A.; Asmaryan, S. Studying Spatial-Temporal Changes and Relationship of Land Cover and Surface Urban Heat Island Derived through Remote Sensing in Yerevan, Armenia. *Build. Environ.* **2021**, *187*, 107390. [CrossRef]
27. Climate Change Information Center. Available online: <http://www.nature-ic.am/en> (accessed on 13 March 2023).
28. Third National Communication on Climate Change: Under the United Nations Framework Convention on Climate Change; “Lusabats” Publishing House, Yerevan, 2015. Available online: <https://unfccc.int/resource/docs/natc/armnc3.pdf> (accessed on 13 March 2023).
29. Foga, S.; Scaramuzza, P.L.; Guo, S.; Zhu, Z.; Dilley, R.D.; Beckmann, T.; Schmidt, G.L.; Dwyer, J.L.; Joseph Hughes, M.; Laue, B. Cloud Detection Algorithm Comparison and Validation for Operational Landsat Data Products. *Remote Sens. Environ.* **2017**, *194*, 379–390. [CrossRef]
30. Xu, H. A New Index for Delineating Built-up Land Features in Satellite Imagery. *Int. J. Remote Sens.* **2008**, *29*, 4269–4276. [CrossRef]
31. Parastatidis, D.; Mitraka, Z.; Chrysoulakis, N.; Abrams, M. Online Global Land Surface Temperature Estimation from Landsat. *Remote Sens.* **2017**, *9*, 1208. [CrossRef]
32. Jimenez-Munoz, J.C.; Cristobal, J.; Sobrino, J.A.; Soria, G.; Ninyerola, M.; Pons, X.; Pons, X. Revision of the Single-Channel Algorithm for Land Surface Temperature Retrieval from Landsat Thermal-Infrared Data. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 339–349. [CrossRef]
33. Jimenez-Munoz, J.C.; Sobrino, J.A.; Skokovic, D.; Mattar, C.; Cristobal, J. Land Surface Temperature Retrieval Methods From Landsat-8 Thermal Infrared Sensor Data. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1840–1843. [CrossRef]
34. Riley, S.; Degloria, S.; Elliot, S.D. A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity. *Int. J. Sci.* **1999**, *5*, 23–27.
35. Chuai, X.W.; Huang, X.J.; Wang, W.J.; Bao, G. NDVI, Temperature and Precipitation Changes and Their Relationships with Different Vegetation Types during 1998–2007 in Inner Mongolia, China. *Int. J. Climatol.* **2013**, *33*, 1696–1706. [CrossRef]
36. Hou, G.; Zhang, H.; Wang, Y. Vegetation Dynamics and Its Relationship with Climatic Factors in the Changbai Mountain Natural Reserve. *J. Mt. Sci.* **2011**, *8*, 865–875. [CrossRef]
37. Yagoub, Y.E.; Li, Z.; Musa, O.S.; Anjum, M.N.; Wang, F.; Bi, Y.; Zhang, B. Correlation between Climate Factors and Vegetation Cover in Qinghai Province, China. *J. Geogr. Inf. Syst.* **2017**, *9*, 403–419. [CrossRef]
38. Zhao, Z.-Q.; He, B.-J.; Li, L.-G.; Wang, H.-B.; Darko, A. Profile and Concentric Zonal Analysis of Relationships between Land Use/Land Cover and Land Surface Temperature: Case Study of Shenyang, China. *Energy Build.* **2017**, *155*, 282–295. [CrossRef]
39. Cui, L.; Shi, J. Temporal and Spatial Response of Vegetation NDVI to Temperature and Precipitation in Eastern China. *J. Geogr. Sci.* **2010**, *20*, 163–176. [CrossRef]
40. Gitelson, A.A.; Kaufman, Y.J. MODIS NDVI optimization to fit the AVHRR data series—Spectral considerations. *Remote Sens. Environ.* **1998**, *66*, 343–350. [CrossRef]
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2012**, *12*, 2825–2830.
42. Zakharov, V.P.; Bratchenko, I.A.; Artemyev, D.N.; Myakinin, O.O.; Kozlov, S.V.; Moryatov, A.A.; Orlov, A.E. 17-Multimodal Optical Biopsy and Imaging of Skin Cancer. In *Neurophotonics and Biomedical Spectroscopy*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 449–476. ISBN 978-0-323-48067-3.
43. Benali, A.; Carvalho, A.C.; Nunes, J.P.; Carvalhais, N.; Santos, A. Estimating Air Surface Temperature in Portugal Using MODIS LST Data. *Remote Sens. Environ.* **2012**, *124*, 108–121. [CrossRef]
44. Zhang, H.; Zhang, F.; Ye, M.; Che, T.; Zhang, G. Estimating Daily Air Temperatures over the Tibetan Plateau by Dynamically Integrating MODIS LST Data. *JGR Atmos.* **2016**, *121*, 11–425. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.