The LST predict of Delaware River Basin: An OLS based scenario analysis on climatic-socioeconomic policies

Weilai Xu, Yuewei Shi, Yajie Zeng, Chelsea Liu

Abstract

In the context of global warming, projecting urban surface temperature has become essential for assessing future risks. This study aims to predict land surface temperature (LST) in the Delaware River Basin for various scenarios in the coming century. OLS and Random Forest models are compared, with OLS demonstrating better LST modeling performance. Heat exposure assessment is conducted based on the predictions, indicating a potential 1.66°C increase and a doubled population exposed to heat under SSP3.

1 Introduction

Under the effects of global warming and urbanization, Land surface temperatures are increasing, causing health risks, energy demands, and environmental stress. To reduce the hazards of high temperatures, we need to understand the influencing factors and development trends of land surface temperature.

In addressing these concerns, our study aims to predict LST accurately through advanced modeling techniques. These models consider various factors, including land use patterns, population density, terrain features, and temporal changes. The Delaware River Basin is chosen as the study area, benefiting from comprehensive data gathered through remote sensing, population statistics, and digital elevation models.

The crux of this research is to develop and refine predictive models for LST, which is crucial for mitigating the negative impacts of urban heat. By applying methodologies such as Ordinary Least Squares (OLS) regression and Random Forest analysis, our work seeks to enhance the understanding of the multifaceted factors that contribute to higher LST in urban environments. This is not just an academic exercise but a step towards practical solutions for urban climate challenges in the era of global warming.

Previous studies indicate that the deterioration of urban thermal environments has become one of the most prominent features of global climate change.[1] Urban populations are now under threat. [2]It is necessary to investigate the factors of urban heat increment.

Numerous studies have confirmed that land surface temperature is influenced by the composition of the land cover and its spatial patterns.[3] Brans et al.[4] and Geng et al.[5] found water body and green space have profound cooling effects. While urban impervious surface can heat the city.[3]Both non-linear and linear model is used in fitting land surface temperature with geographical parameters. [6, 7]GWR is widely used for the spatial overflow effect of LST. OLS and some machine-learning models such as Random Forest and Cubist are also used. Phan et al. proved that cubist and RF perform better in mountainous areas.[6]

Climate change as well as urban development is exacerbating the deterioration of the urban heat environment. Several researches focus on the impact of climate change on heat exposure but only on large continent scales.[8, 9] Result shows that urbanization and population growth are less influential than climate policy. However, a recent study in Singapore also emphasizes the importance of urbanization and population growth on a smaller scale.[10]

To mitigate the adverse impacts of increasing urban heat, there is a need to develop accurate predictive models for land surface temperature. By employing models such as Ordinary Least Squares (OLS) regression and Random Forest, the research endeavors to fill the existing gap in understanding the complex interplay of factors contributing to elevated land surface temperature in urban environments.

2 Materials and methods

2.1 Study area

The Delaware River Basin was selected for this study due to its critical role as an economic hub and a key regional water supply. This area is particularly relevant for examining the impact of 21st-century climate change, as it is projected to experience significant warming, with temperatures rising 1–5 °C by the century's end, especially in winter and summer.[11]

2.2 Data Collection

All data sources are presented in Appendix A. A 1-kilometer fishnet within Delaware River Basin is created for zonal statistical analysis and producing a dataframe including all features later.

The historical land use data is derived from the National Land Cover Database from

EROS. The predicted land use data under SSPs are from Dornbierer et al.(2021)[12] The original data is of 30-meter resolution. Data is transformed to projected coordinates and reclassified into 8 classes specified in Table 1. Then, using the fishnet as feature zone input and land use data as feature class input to perform a tabulate area. The area of each land use class within grids of the fishnet is calculated and exported as a table. Since we mainly study urban area data, we remove places where Developed Low Intensity and Developed Med-High Intensity areas account for less than 40%.

Table 1: Reclassification of Land Use

Class Number	**Land Use Type**
1	Cropland
2	Open Water
3	Developed Low Intensity
4	Developed Med-High Intensity
5	Barren
6	Forest
7	Grassland
8	Wetlands

Historical population count data is from WorldPop. The predicted population count data under various Shared Socioeconomic Pathways (SSPs) published by Li et al.[13] is produced based on WorldPop data. The images are projected and processed through zonal statistics as a table using ArcGIS Pro. The result table shows the total population count within a 1-kilometer grid.

3DEP DEM dataset and land surface temperature dataset are downloaded via Google Earth Engine. They are reprojected and the average elevation and total population in each grid are calculated through zonal statistics as a table.

2.3 Data Preprocessing

In this part, the paper mainly conducts multicollinearity removal and data normalization. Among them, multicollinearity removal is to prevent inaccurate coefficient estimation and instability of the model. Specifically speaking, this paper uses Pearson coefficient and VIF to test the multicollinearity of the dependent variable while using Min-Max Scaling to conduct data normalization.

2.3.1 Pearson correlation coefficients Analysis

Table 2: Pearson correlation coefficients

Variables	Correlation coefficient
Percentage of Cropland	-0.238727
Percentage of Open Water	-0.083048
Percentage of Developed Low Intensity	-0.034189
Percentage of Developed Med-High Intensity	0.518109
Percentage of Barren	-0.002616
Percentage of Forest	-0.495540
Percentage of Grassland	0.008410
Percentage of Wetlands	-0.063019
Average elevation	-0.591892
Population count	0.456317
Year	-0.056269

Our analysis revealed complex relationships between LST and various factors:

- LULC_4 and LST: Strong positive correlation (0.518), indicating higher temperatures in urban areas like commercial or industrial zones. POP and LST: Positive correlation (0.448), reflecting heat accumulation in densely populated areas.
- LULC_1 and LST: Negative correlation (-0.239), suggesting cooling effects of green spaces.
- DEM and LST: Significant negative correlation (-0.593), showing terrain's impact on surface temperatures.
- LULC_6: Excluded due to severe multicollinearity. Other Variables: Except for LULC_5 and LULC_7, all other variables show a significant correlation with LST ($p < 0.05$).

2.3.2 VIF Analysis

After Pearson analysis, we removed the LULC_6 factor and then performed VIF analysis on the model.

Table 3: Table of VIF results

Variables	VIF
Percentage of Cropland	1.6855
Percentage of Open Water	1.1452
Percentage of Developed Low Intensity	2.9619
Percentage of Developed Med-High Intensity	3.8295
Percentage of Barren	1.0371
Percentage of Grassland	1.0948
Percentage of Wetlands	1.4582

Variables	VIF
Average elevation	1.3297
Population count	1.6766
Year	1.0147

From the above results, it can be seen that the VIF of all factors at present are small, ranging from 1 to 4, indicating that there is no strong collinearity between these factors, so all the above factors can be selected as independent variables.

2.3.3 Data normalization

Min-Max Scaling: Applied to ensure uniform scales and ranges for different variables in the multivariable regression model. This prevents variables with extensive value ranges from disproportionately influencing the model. By scaling values between 0 and 1, each variable contributes equally to the model's outcomes.

$$NormalizedValue = \frac{Value - Min}{Max - min} \tag{1}$$

- The original data were normalized to ensure comparability and analysis across different variables. The normalized ratio of LULC, DEM, POP, and Year were chosen as explanatory variables to predict LST.
- This selection was based on a prior understanding of factors affecting surface temperatures, such as terrain, population density, and changes in land use.

3 Result

Firstly, the paper uses Pearson analysis and VIF to screen independent variables with multicollinearity are eliminated. Then, the paper utilized data from the years 2010, 2015, and 2020, employing Ordinary Least Squares (OLS) regression and Random Forest models for constructing our models. Finally, the model is applied to the data of 2005, the prediction results of the two models are compared, and the model that fits better will be selected. Then we use that model to simulate and predict the LST in future years.

We utilized data from the years 2010, 2015, and 2020, employing Ordinary Least Squares (OLS) regression and Random Forest models for constructing our models. Finally, the model is applied to the data of 2005, the prediction results of the two models are compared, and the model that fits better will be selected. Then we use that model to simulate and predict the LST in future years.

3.1 OLS Regression

Our OLS regression analysis reveals significant relationships between land surface temperature (LST) and several key variables. POP_scaled, with a coefficient of 4.3447

and a near-zero p-value, demonstrates a strong positive impact on LST, suggesting a notable rise in surface temperature with increased normalized population size, likely due to urban heat factors. The LULC_scaled variables show diverse effects: LULC_1_scaled has a slight negative impact on LST (coefficient -0.4488), while LULC_4_scaled significantly increases LST (coefficient 7.6255), reflecting the different impacts of various land uses and covers, such as vegetation and urban development. DEM_scaled, with a coefficient of -9.8579, indicates a significant negative effect of elevation on LST, in line with cooler temperatures at higher altitudes. The year_scaled variable, with a coefficient of -1.2567, points to a decrease in surface temperature over time, possibly due to environmental policies or temporal changes.

Regarding model performance, the R-squared value of 0.595 suggests the model explains 59.5% of the variability in LST. However, when applied to 2005 data, the model's performance metrics indicate lower accuracy: MSE of 5.8574, RMSE of 2.4202, R^2 of 0.3752, and MAE of 1.9237. These figures show a substantial deviation in predictive accuracy for 2005, with the model explaining only about 37.5% of the LST variance, significantly lower than the overall model. POP_scaled remains the most important predictor, with terrain (DEM_scaled) and time (year_scaled) being significant factors.

In conclusion, while the model demonstrates substantial impacts of factors like population size, land use, terrain, and time on surface temperature, the 2005 data analysis reveals sensitivity to temporal variations and the need for further research to enhance accuracy, particularly for specific periods or geographic regions.

$$\begin{aligned} LST =& 304.0682 - 0.4488 \times LULC\_1 + -0.5954 \times LULC\_2 \\ &+ 5.0377 \times LULC\_3 + 7.6255 \times LULC\_4 + 2.2075 \times LULC\_5 \\ &+ 3.9130 \times LULC\_7 - 0.3371 \times LULC\_8 - 9.8579 \times DEM \\ &+ 4.3447 \times POP - 1.2567 \times Year \end{aligned} \tag{2}$$

3.2 Random Forest

We employed a random forest model for our analysis. Initially, we partitioned the data from 2010, 2015, and 2020 into training and test sets, allocating 80% of the data for training and the remaining 20

3.3 Model Selection

The R^2 value of OLS is larger than the random forest model. That is to say, OLS yields a better fit to the data, explaining about half of the variability in dependent variables. The comparative results show that OLS exhibits a lower Mean Squared Error (MSE) and root Mean Squared Error (RMSE), signifying that the difference between the predicted value of the model and the actual observed value is small, and the model can fit the data well. Therefore, we choose the OLS model because the model exhibits a better fit.

Table 4: OLS vs RF

	OLS	**RF**
R2	0.4800	0.2509
MAE	1.7514	2.1661
RMSE	2.2017	2.6427
MSE	4.8475	6.9837

3.4 Model Result

The result shows that under the SSP3 scenario, the regional rivalry path will lead to the most significant increase in land surface temperature by 1.25℃ in mid-century and 1.66℃ at the end of the century, also with the most population affected by heat exposure which is almost doubled in 2100. Under the SSP5 fossil-fueled development scenario, the DRB area will be heated by 1.23℃ in mid-century and 1.6℃ by the end. The affected population will increase first to around 2.98 billion, however, followed by a drop to 2.40 billion which is even less than the population affected in 2020. The SSP2 middle-road scenario will also lead to a significant increase in LST rise and a large jump in the affected population to 3.18 billion in 2050, falling to 3.03 billion in 2100 inside the current urban boundary. Considering the change in land use, the LST increase could be slightly milder but still severe. More people in the newly expanded urban area will experience heat exposure, especially under SSP5.

Table 5: LST distribution and population affected by heat exposure in original urban boundary

	LST_mean	**LST_std**	**Affected_pop**
2050-SSP2	307.58	2.52	3.18e+06
2050-SSP3	307.60	2.50	3.47e+06
2050-SSP5	307.58	2.51	2.98e+06
2100-SSP2	308.02	2.61	3.03e+06
2100-SSP3	308.01	2.60	4.29e+06
2100-SSP5	307.95	2.61	2.40e+06
2020 (control)	306.35	3.14	2.46e+06

Table 6: LST distribution and population affected by heat exposure in changed urban boundary

	LST_mean	**LST_std**	**Affected_pop**
2050-SSP2	307.46	2.56	3.18e+06
2050-SSP3	307.47	2.55	3.47e+06
2050-SSP5	307.44	2.56	2.98e+06
2100-SSP2	307.73	2.69	2.97e+06

	LST_mean	LST_std	Affected_pop
2100-SSP3	307.74	2.66	4.29e+06
2100-SSP5	307.70	2.69	2.41e+06
2020 (control)	306.35	3.14	2.46e+06

Regional rivalry path has a great impact on the Delaware River Basin land surface temperature increase and can double the risk of facing heat exposure. Nonetheless, under each scenario, if no action is taken, the land surface temperature in the Delaware River Basin will at least increase by 1.09℃ in 2050, which would lead to severe consequences such as flooding in this water supply entity for the region.

4 Conclusion

Considering the impact of both social and natural factors on land surface temperature in the context of development and climate change, OLS and Random Forest Regression models are compared. OLS model has a better performance in predicting LST in the Delaware River Basin. And the result shows that it will heat by 1.66℃ and double the population exposed to heat under SSP3. We should be aware of the possible consequences and prepare for them.

Appendix A Data Source

Table A1 Dataset Information

Dataset	Year	Format	Resolution	Source
LST-MYD11A2.061	2005, 2010, 2015, 2020; June to August	Raster	1km, 8 day	USGS (https://lpdaac.usgs.gov)
Land Use Prediction under SSPs	2020 to 2100	Raster	30m	USGS (https://www.sciencebase.gov)
Land Use Historical	2006, 2010, 2016, 2021	Raster	30m	USGS NLCD (https://www.usgs.gov), USGS (https://www.sciencebase.gov)
Population Prediction under SSPs	2020 to 2100	Raster	1km	https://www.geosimulation.cn/FPOP.html
Historical Population	2005, 2010, 2015, 2020	Raster	100m	WorldPop (https://www.worldpop.org)
DEM	1998 to 2020	Raster	10.2m	USGS_3DEP_10m (https://developers.google.com)
Delaware River Basin Boundary	-	Vector	-	https://www.nj.gov

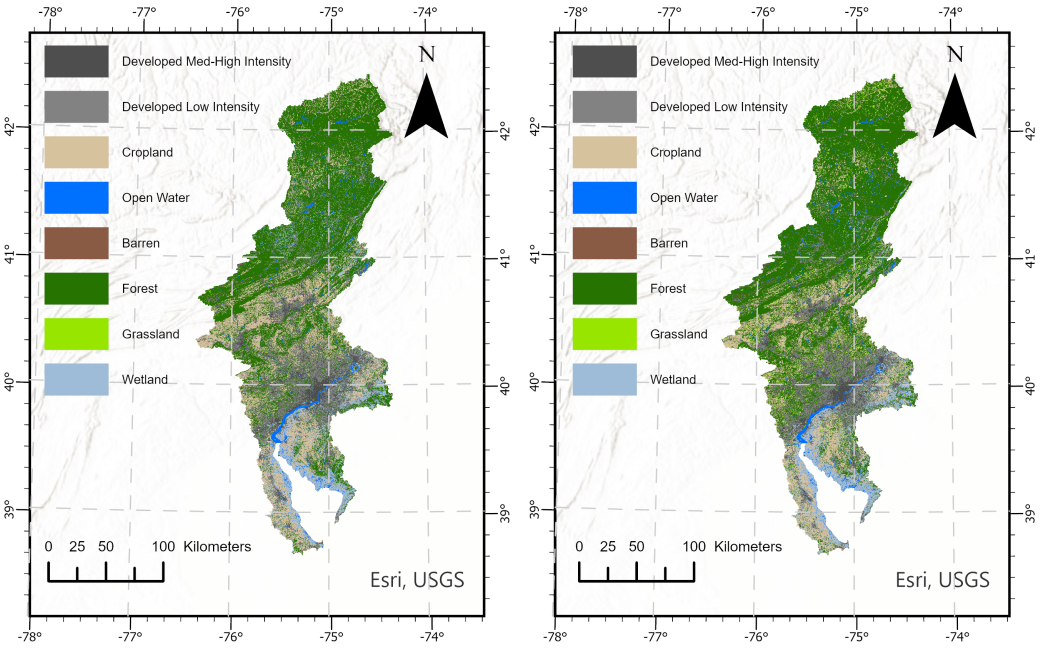

Fig. A1 Landuse in 2021 (Right) and predicted landuse in 2050 in Delaware River Basin under SSP5 (Left)

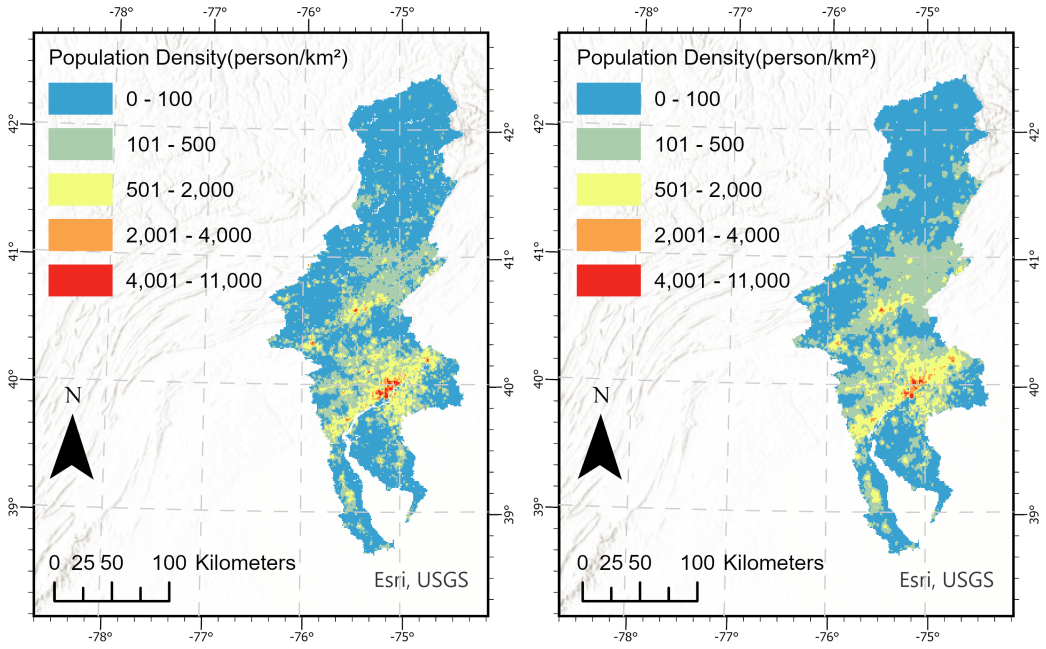

Fig. A2 Polulation density in 2021 (Right) and predicted in 2050 in Delaware River Basin under SSP5 (Left)

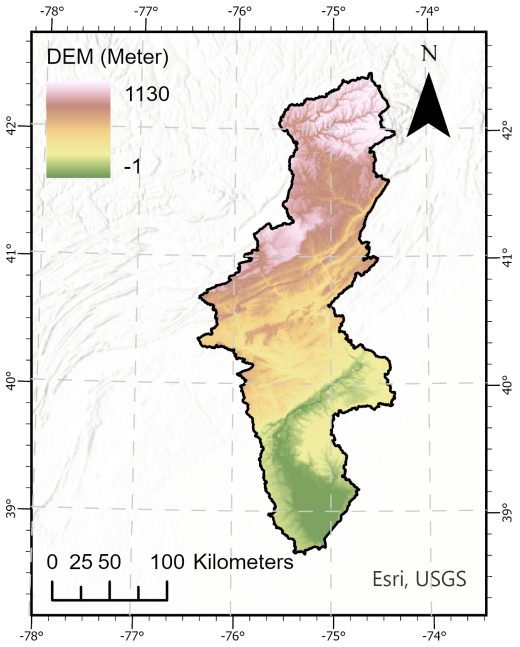

Fig. A3 DEM in Delaware River Basin

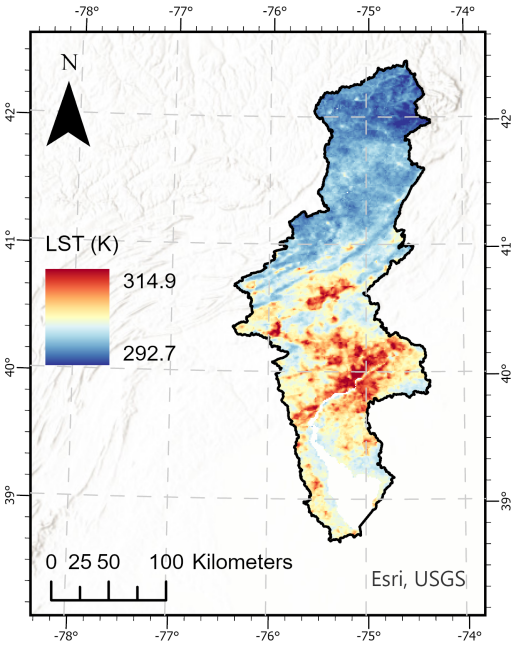

Fig. A4 Land surface temperature in 2020 summer in Delaware River Basin

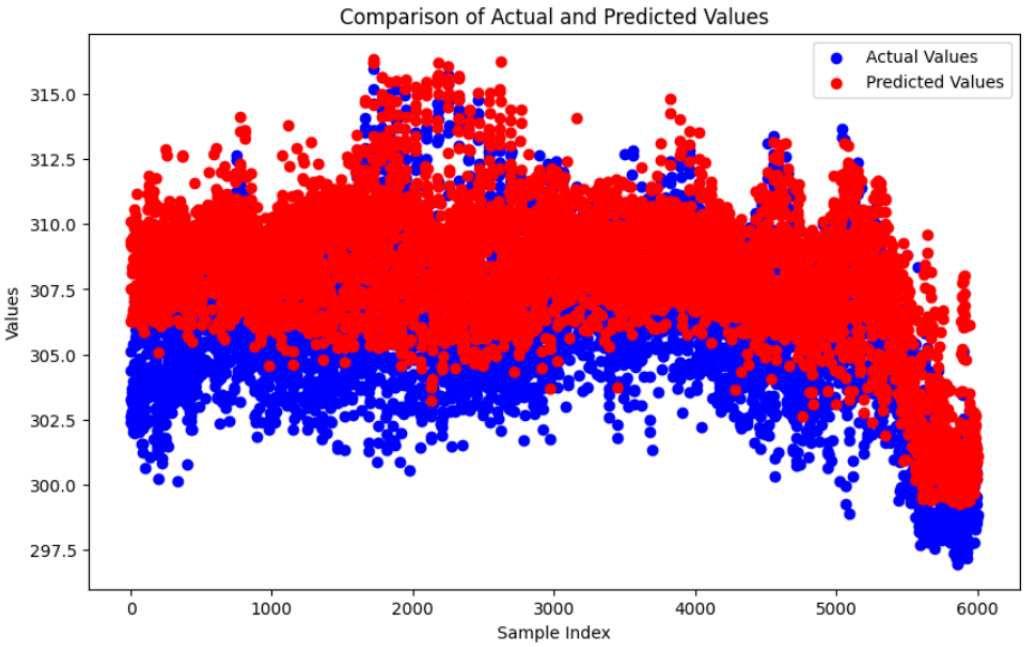

Fig. A5 Comparison of Actual and Predicted Values

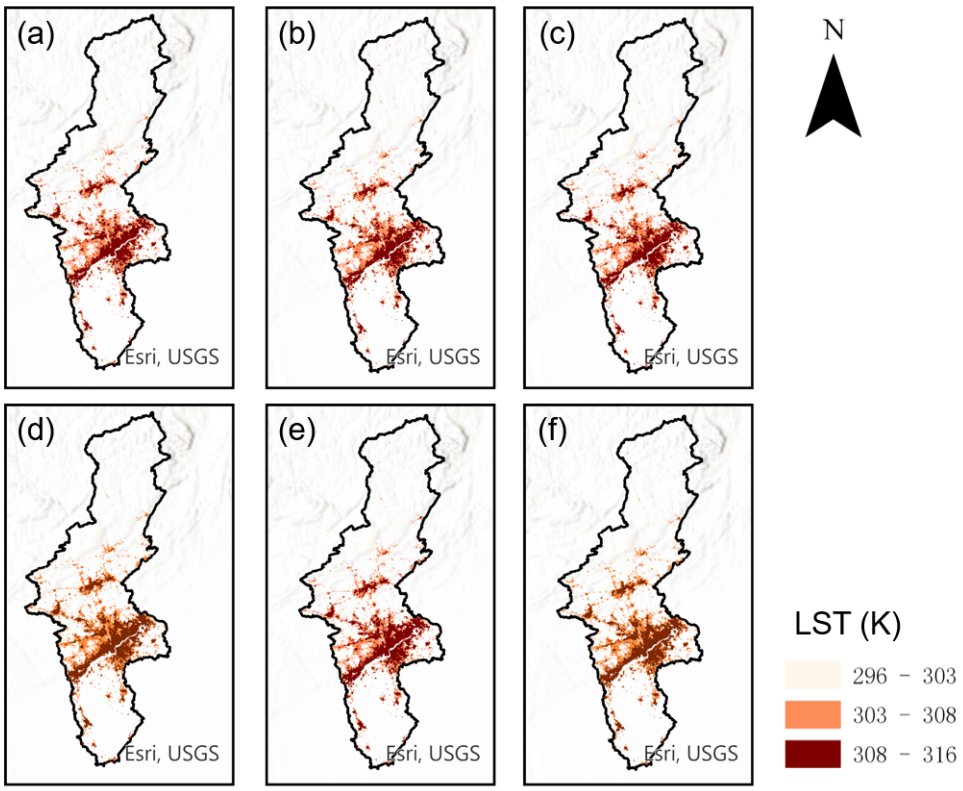

Fig. A6 Comparison of Actual and Predicted Values

Appendix B Individual Roles

Table B2: Individual Roles

Name	Role
Chelsea Liu	Clean data, exploratory data analysis, Visualize the prediction result
Yuewei Shi	Random Forest, model selection
Yajie Zeng	Normalize the data, build an OLS regression model with analysis
Weilai Xu	literature review, Data download and process, Model result, map visualization, conclusion

References

[1] Yao Yuan, Chen Xi, Qian Jing, Research progress on the thermal environment of the urban surfaces. Acta Ecologica Sinica **38**(3) (2018) https://doi.org/10.5846/stxb201611022233 . Accessed 2023-12-13

[2] He, B.-J., Wang, J., Zhu, J., Qi, J.: Beating the urban heat: Situation, background, impacts and the way forward in China. Renewable and Sustainable Energy Reviews **161**, 112350 (2022) https://doi.org/10.1016/j.rser.2022.112350 . Accessed 2023-12-13

[3] Zhou, L., Yuan, B., Hu, F., Wei, C., Dang, X., Sun, D.: Understanding the effects of 2D/3D urban morphology on land surface temperature based on local climate zones. Building and Environment **208**, 108578 (2022) https://doi.org/10.1016/j.buildenv.2021.108578 . Accessed 2023-12-13

[4] Brans, K.I., Engelen, J.M.T., Souffreau, C., De Meester, L.: Urban hot-tubs: Local urbanization has profound effects on average and extreme temperatures in ponds. Landscape and Urban Planning **176**, 22–29 (2018) https://doi.org/10.1016/j.landurbplan.2018.03.013 . Accessed 2023-12-13

[5] Geng, X., Yu, Z., Zhang, D., Li, C., Yuan, Y., Wang, X.: The influence of local background climate on the dominant factors and threshold-size of the cooling effect of urban parks. Science of The Total Environment **823**, 153806 (2022) https://doi.org/10.1016/j.scitotenv.2022.153806 . Accessed 2023-12-13

[6] 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 Sensing **9**(5), 398 (2017) https://doi.org/10.3390/rs9050398 . Number: 5 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2023-12-13

[7] Kashki, A., Karami, M., Zandi, R., Roki, Z.: Evaluation of the effect of geographical parameters on the formation of the land surface temperature by applying OLS and GWR, A case study Shiraz City, Iran. Urban Climate **37**, 100832 (2021) https://doi.org/10.1016/j.uclim.2021.100832 . Accessed 2023-12-13

[8] Chen, H., Zhao, L., Cheng, L., Zhang, Y., Wang, H., Gu, K., Bao, J., Yang, J., Liu, Z., Huang, J., Chen, Y., Gao, X., Xu, Y., Wang, C., Cai, W., Gong, P., Luo, Y., Liang, W., Huang, C.: Projections of heatwave-attributable mortality under climate change and future population scenarios in China. The Lancet Regional Health. Western Pacific **28**, 100582 (2022) https://doi.org/10.1016/j.lanwpc.2022.100582

[9] Chen, M., Chen, L., Zhou, Y., Hu, M., Jiang, Y., Huang, D., Gong, Y., Xian, Y.: Rising vulnerability of compound risk inequality to ageing and extreme heatwave exposure in global cities. npj Urban Sustainability **3**(1), 1–11 (2023) https://

doi.org/10.1038/s42949-023-00118-9 . Number: 1 Publisher: Nature Publishing Group. Accessed 2023-12-13

[10] Lan, T., Peng, J., Liu, Y., Zhao, Y., Dong, J., Jiang, S., Cheng, X., Corcoran, J.: The future of China's urban heat island effects: A machine learning based scenario analysis on climatic-socioeconomic policies. Urban Climate **49**, 101463 (2023) https://doi.org/10.1016/j.uclim.2023.101463 . Accessed 2023-12-13

[11] Hawkins, T.W., Woltemade, C.J.: Impact of projected 21st century climate change on basin hydrology and runoff in the Delaware River Basin, USA. Journal of Water and Climate Change **12**(1), 60–81 (2019) https://doi.org/10.2166/wcc.2019.140 . Accessed 2023-12-13

[12] Dornbierer, J.C.M., Wika, S.C., Robison, C.J., Rouze, G.S., Sohl, T.L.: Long-term database of historical, current, and future land cover for the Delaware River Basin (1680 through 2100). U.S. Geological Survey (2021). https://doi.org/10.5066/P93J4Z2W . https://www.sciencebase.gov/catalog/item/605c987fd34ec5fa65eb6a74 Accessed 2023-12-13

[13] Li, M., Zhou, B.-B., Gao, M., Chen, Y., Hao, M., Hu, G., Li, X.: Spatiotemporal dynamics of global population and heat exposure (2020–2100): based on improved SSP-consistent population projections. Environmental Research Letters **17**(9), 094007 (2022) https://doi.org/10.1088/1748-9326/ac8755 . Accessed 2023-10-22