

Supplementary Information for:

Physics Guided AI for Urban Heat Mitigation Planning

Abdul Ghafoor Nizamani^{a,b,*}, Dung Thi Vu^{c,d}, Rafeeqe Ahmed Nizamani^e,

Mostafa Aliyari^a, Cunrong Li^e

^aFaculty of Computer Science, Engineering and Economics, Østfold University of Applied Sciences, BRA Veien 4, Halden, 1757, NO, Norway

^bDepartment of Civil Engineering, Delft University of Technology, Delft, 2628 CD, Netherlands

^cDepartment of Mathematics, Faculty of Fundamental Sciences, Phenikaa University, Nguyen Trac Street, Duong Noi Ward, Hanoi, 12116, Vietnam

^dIstituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna, Viale Carlo Berti Pichat, 6/2, Bologna, 40127 BO, Italy

^eSchool of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan, Hubei, P.R. China

Supplementary Appendix A: Pairwise relationships among LST, NDVI, and emissivity

On 19 May, the LST distribution was concentrated around 31–32 °C, while NDVI values were concentrated mainly between 0.1 and 0.2 (Fig. A1). The reported Pearson correlations were $r = -0.71$ for LST–NDVI, $r = -0.69$ for LST–emissivity, and $r = 0.64$ for NDVI–emissivity.

On 4 June, the LST distribution shifted to approximately 20–21 °C and the NDVI distribution contained more values around 0.4–0.5 (Fig. A2). The reported correlations were $r = -0.70$ for LST–NDVI, $r = -0.73$ for LST–emissivity, and $r = 0.84$ for NDVI–emissivity.

On 20 June, the LST distribution was again concentrated around 30–31 °C, while the NDVI distribution was weighted toward lower values (Fig. A3). The reported correlations were $r = -0.75$ for LST–NDVI, $r = -0.72$ for LST–emissivity, and $r = 0.78$ for NDVI–emissivity.

Across the three acquisitions, LST was negatively correlated with NDVI and emissivity, whereas NDVI and emissivity were positively correlated. These relationships are descriptive associations. Because emissivity was derived partly from NDVI during preprocessing, the NDVI–emissivity correlation is not independent evidence of a causal physical relationship. The three acquisitions represent 19 May, 4 June, and 20 June 2025 only and are not used here to infer a seasonal or phenological trajectory.

Pairwise Relationships - 19 May

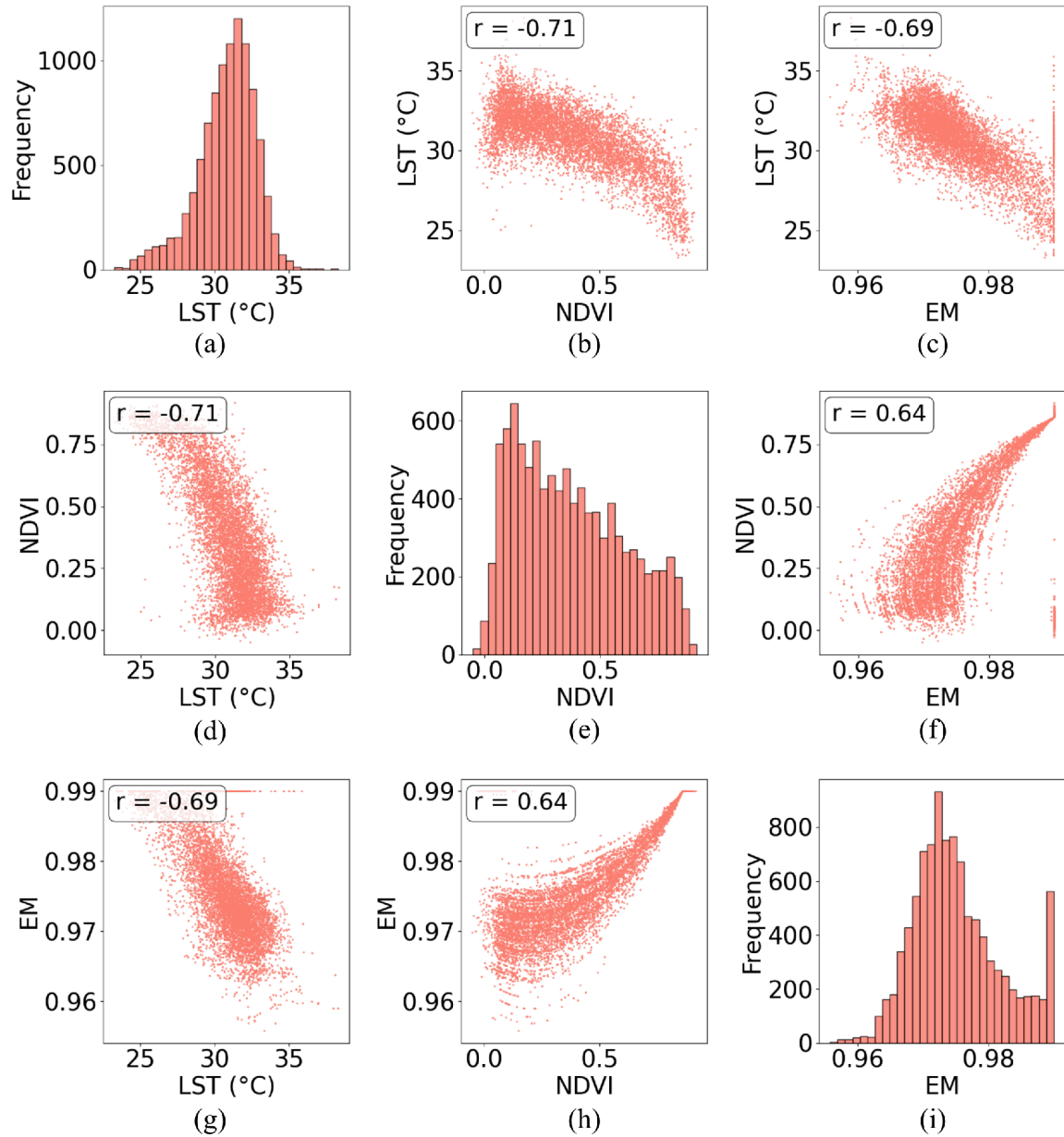


Fig. A1: Pairwise relationships among land surface temperature (LST), Normalised Difference Vegetation Index (NDVI), and emissivity (EM) on 19 May. Panels (a), (e), and (i) present frequency histograms for LST, NDVI, and EM. Panels (b) and (d) compare LST with NDVI; panels (c) and (g) compare LST with EM; panels (f) and (h) compare NDVI with EM. Pearson correlation coefficients (r) are shown in the scatter plots.

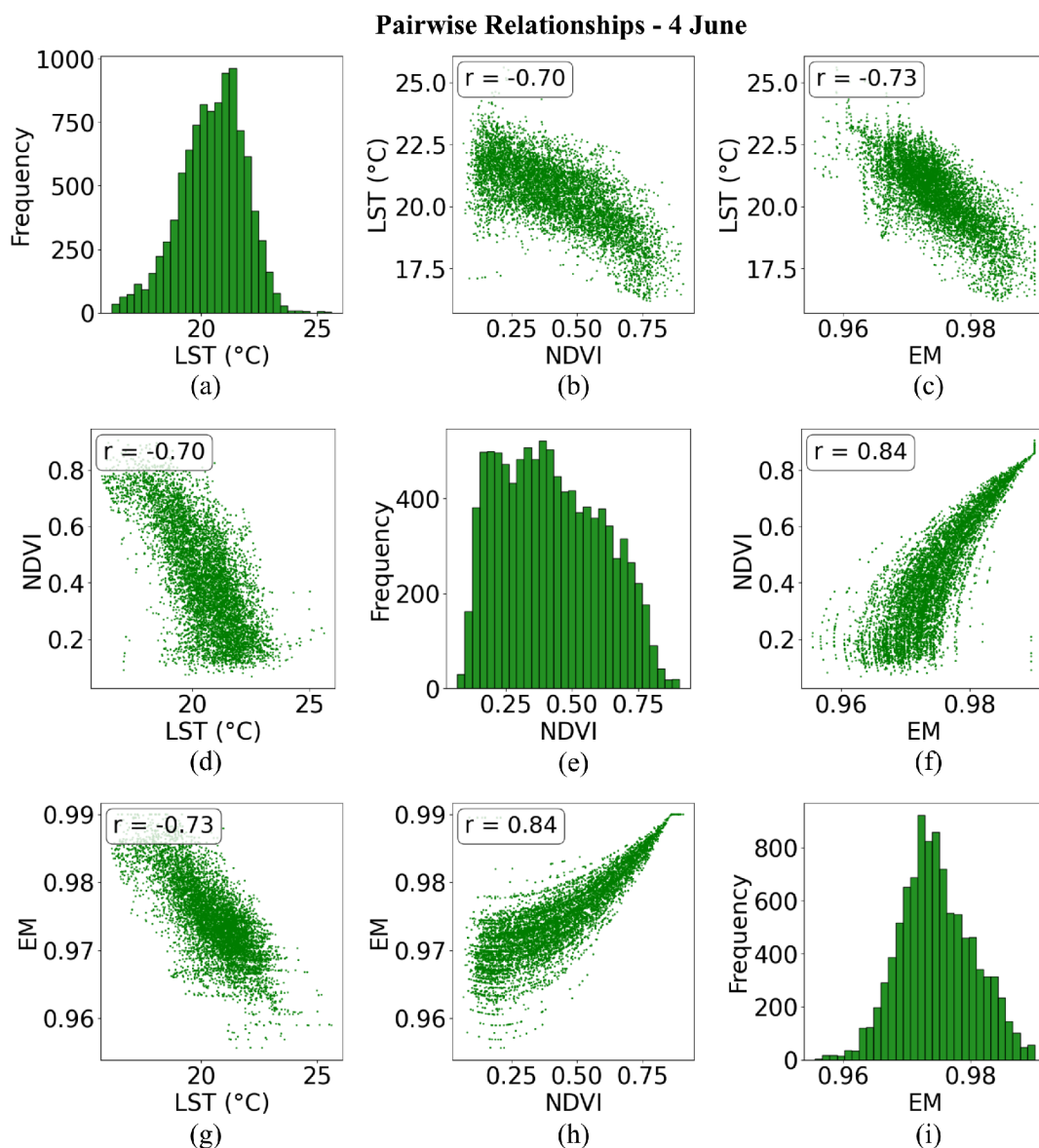


Fig. A2: Pairwise relationships of LST, NDVI, and EM on 4 June. Panel arrangement and annotations follow Fig. A1.

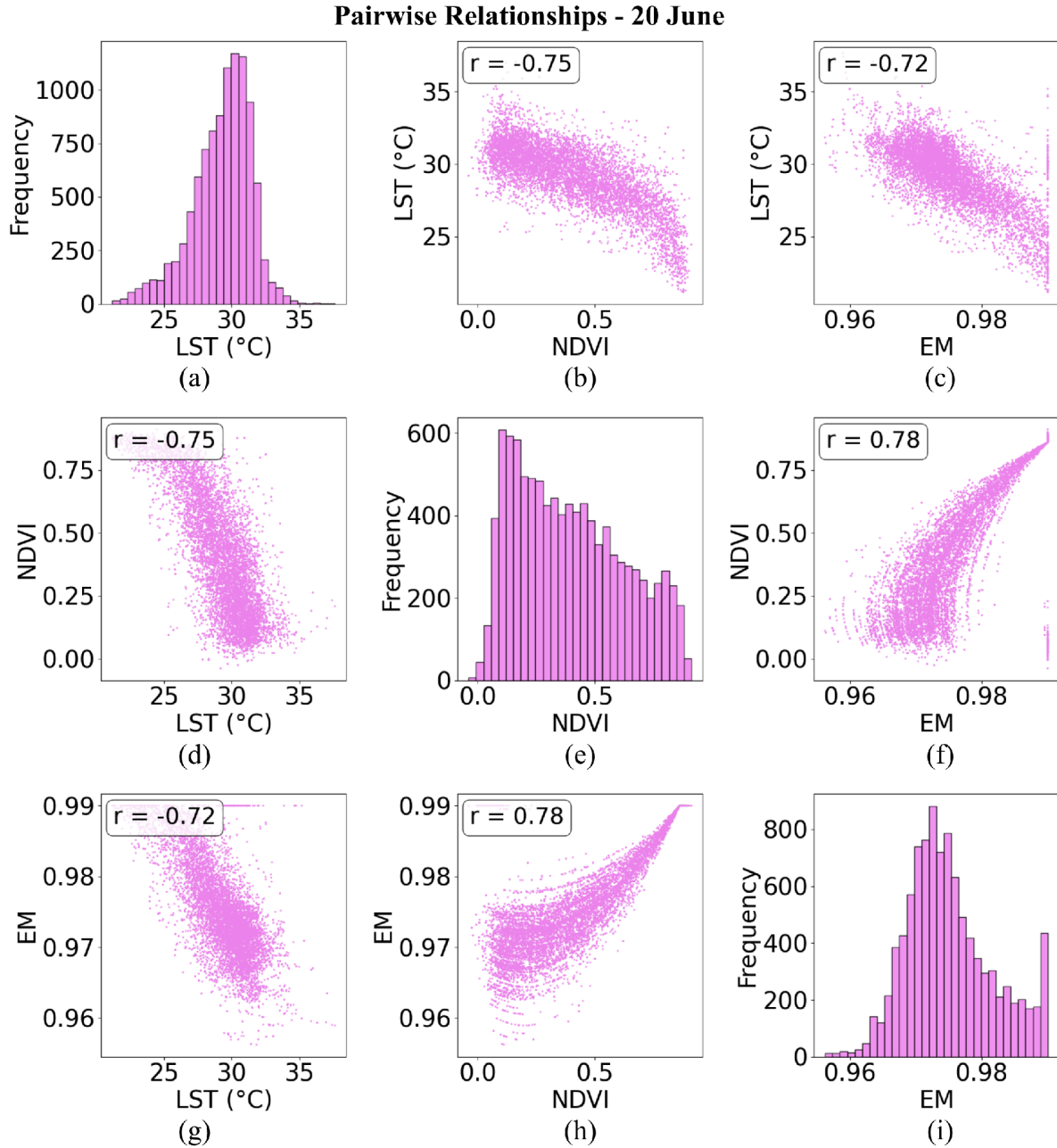


Fig. A3: Pairwise relationships of LST, NDVI, and EM on 20 June. Panel arrangement and annotations follow Fig. A1.

Supplementary Appendix B: Sensitivity of the OSM-based source-sink index to weighting scenarios

Table B1 records four weighting cases for the OSM-based source-sink index used in the model formulation. Among these recorded cases, Case 1 has the lowest fitted-scene RMSE (0.2549–0.3959 °C) and R² values of 0.9638–0.9676. The sign and range columns describe an

uncalibrated spatial index rather than measured heat flux. As noted in the main Methods, the archived training implementation contains both an OSM-based branch and a fallback branch, while the saved checkpoint does not record which branch was executed. Appendix B therefore documents sensitivity of the OSM-based formulation and should not be interpreted as evidence of the exact branch used by the saved checkpoint.

Figure B1 displays the source and sink inputs and the resulting source-sink index for Case 1. Under the recorded weighting, 99.2% of pixels have a positive index and 0.8% have a negative index. The maps and histogram describe the spatial distribution of this constructed index; they do not represent calibrated surface-energy fluxes.

Figure B2 shows the corresponding index components and distributions for the three alternative weighting cases. Case 2 increases the sink weights and produces 4.5% negative-index area, Case 3 produces an entirely positive index, and Case 4 produces an intermediate distribution. The differences in fitted-scene metrics show that the weighting choices affect the fitted temperature field, but they do not establish the physical correctness of any weighting case.

Table B1: Sensitivity of the OSM-based source-sink index weights and fitted-scene metrics for four recorded weighting cases.

Cases	Source factors	Sink factors	Q* index range	Positive/negative index area (%)	Period / Landsat-derived LST (°C)	Fitted R2	Fitted RMSE (°C)	PIN N LST prediction (°C)	Maximum absolute error (°C)
1	BF=0.3; RN=0.25; PMR=0.2; UC=0.15	TC=0.1; GS=0.15; OW=0.1; CF=0.1	(-0.082, 0.766)	99.2/0.8	19 May / (23.3 - 38.3)	0.9638	0.3758	23.8 - 37.7	2.06
					4 June / (16.2 - 25.6)	0.9663	0.2549	16.2 - 25.6	1.36
					20 June / (21.2 - 37.7)	0.9676	0.3959	21.4 - 37.5	2.06
2	BF=0.25; RN=0.2; PMR=0.15; UC=0.1	TC=0.15; GS=0.2; OW=0.15; CF=0.1	(-0.223, 0.593)	95.5/4.5%	19 May	0.9488	0.4471	23.8 - 37.4	2.62
					4 June	0.9530	0.3011	15.9 - 25.3	1.49
					20 June	0.9573	0.4542	21.3 - 37.0	2.78
3	BF=0.25; RN=0.2; PMR=0.15; UC=0.1	TC=0.15; GS=0.2; OW=0.15; CF=0.1	(0.124, 1.032)	100/0%	19 May	0.9416	0.4776	24.1 - 37.3	2.42
					4 June	0.9385	0.3443	16.2 - 25.3	1.76
					20 June	0.9473	0.5046	21.5 - 37.4	3.06
4	BF=0.25; RN=0.2;	TC=0.15; GS=0.2;	(-0.162, 0.745)	96.8/3.2%	19 May	0.9520	0.4330	23.9 - 38.0	2.57

Cases	Source factors	Sink factors	Q* index range	Positive/negative index area (%)	Period / Landsat-derived LST (°C)	Fitted R2	Fitted RMSE (°C)	PIN LST prediction (°C)	Maximum absolute error (°C)
	PMR=0.15 ; UC=0.1	OW=0.15; CF=0.1			4 June	0.9504	0.3093	16.3 - 25.5	1.42
					20 June	0.9521	0.4815	21.4 - 37.6	2.92

Notation: BF = building factor; RN = road network density; PMR = proximity to major roads; UC = urban canyon index; TC = tree canopy cover; GS = green-space/park cover; OW = open-water cover; CF = cooling-feature proximity. Q* is an uncalibrated spatial source-sink index; percentages denote positive/negative index area.

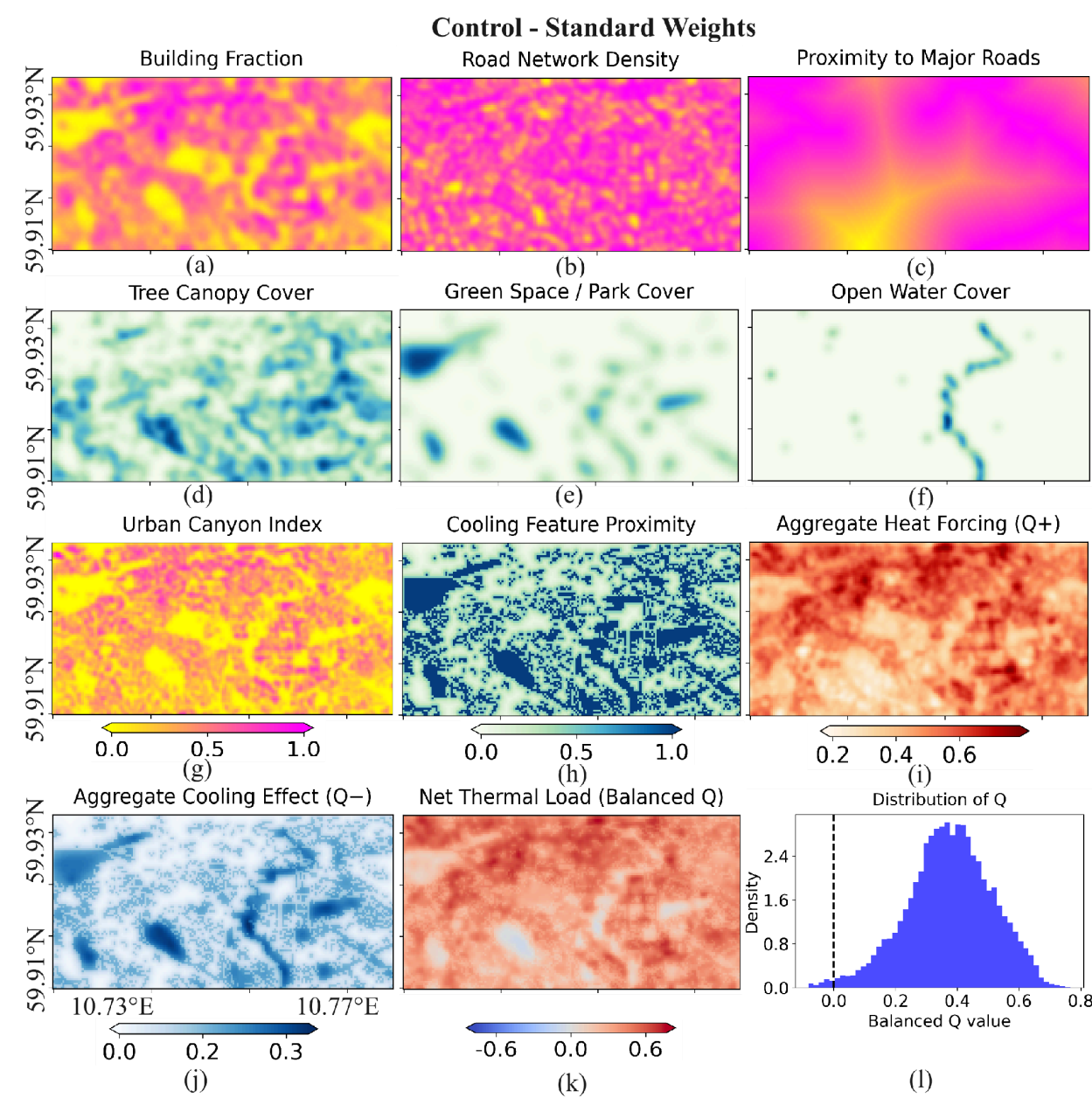


Fig. B1: Spatial distribution of source and sink factors and the resulting OSM-based source-sink index for Case 1. Panels (a)–(h) show source and sink inputs, (i) the aggregate source component, (j) the aggregate sink component, (k) the net source-sink index, and (l) its distribution.

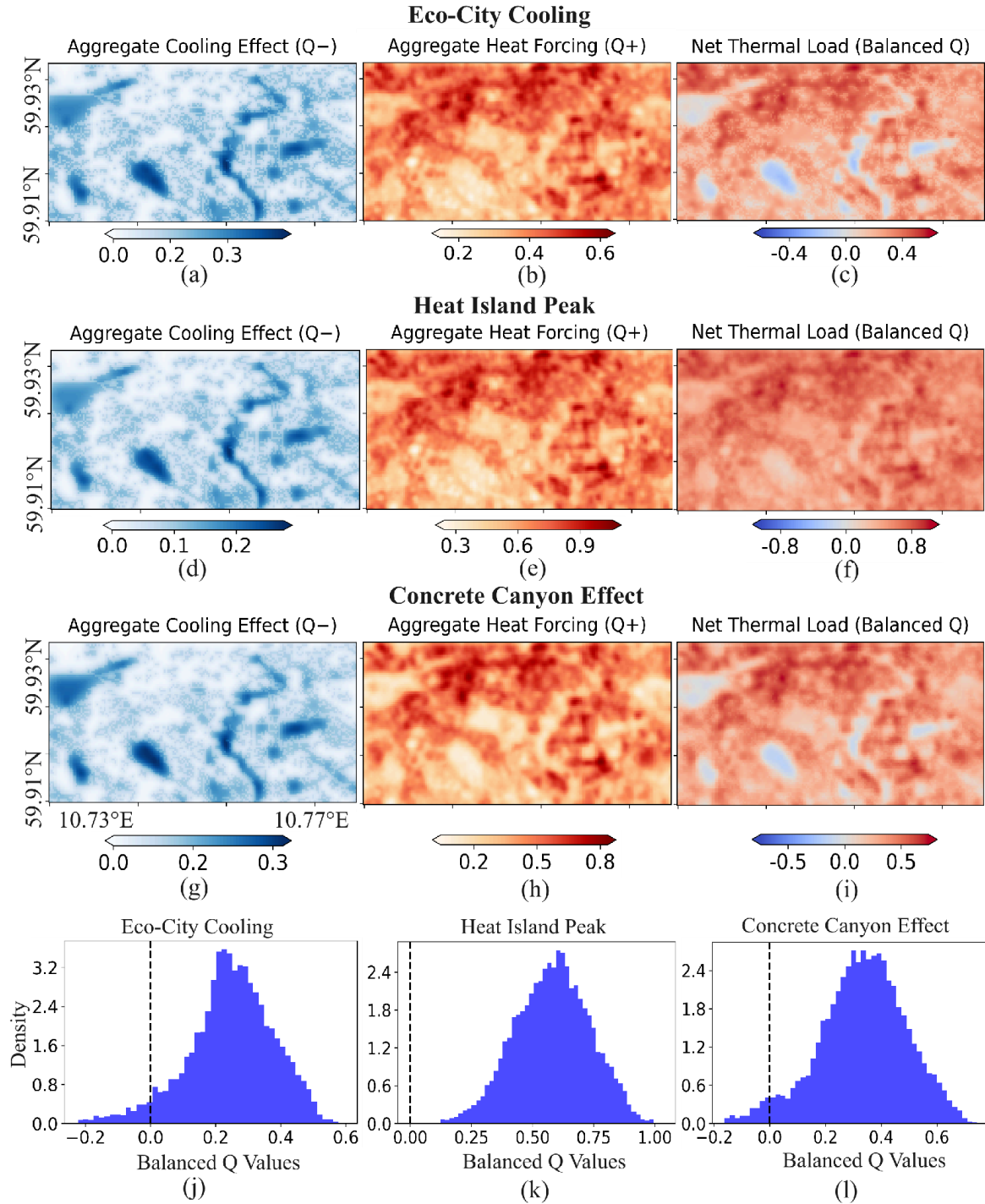


Fig. B2: Source-sink index components and distributions for the three alternative weighting cases. Panels (j)–(l) show the corresponding distributions of the net index.

Supplementary Appendix C: Training and hyperparameter sensitivity record

Table C1 preserves the recorded hyperparameter tests for 19 May, 4 June, and 20 June. The reported R2 and RMSE values compare model predictions with Landsat-derived LST from the same scenes used during fitting and therefore represent fitted-scene agreement. The recorded tests vary training duration, physics- and boundary-loss weights, activation functions, and network size.

For the recorded four-layer sweep configuration, increasing training from 10 to 200 epochs reduced fitted-scene error substantially. Extending the recorded run to 300 epochs did not reduce RMSE relative to the 200-epoch run on all three dates. These comparisons describe the recorded sensitivity runs only.

Within the recorded loss-weight sweep, reducing the physics-loss weight from 1.0 to 0.01 produced lower fitted-scene RMSE on all three dates. The reported main model nevertheless retained a physics-loss weight of 1.0 so that the residual had a larger contribution to the composite training objective; this choice was therefore not selected by minimum fitted-scene RMSE. Increasing the physics-loss weight to 1.5 or the boundary-loss weight to 0.5 increased fitted error relative to the recorded 200-epoch, weight-1.0 run.

The original supplementary record also contains architecture tests based on four-layer configurations, including Tanh-only, three-layer, and narrower four-layer variants. These entries are retained as sensitivity records. They should not be used to define the final network architecture because the surviving main-model implementation described in the main Methods contains seven hidden layers with dimensions 256, 256, 256, 128, 128, 64, and 32.

Accordingly, Appendix C supports statements about the recorded training and loss-weight sensitivity experiments. It does not establish independent predictive performance, physical consistency, or generalisability beyond the three fitted Landsat scenes.

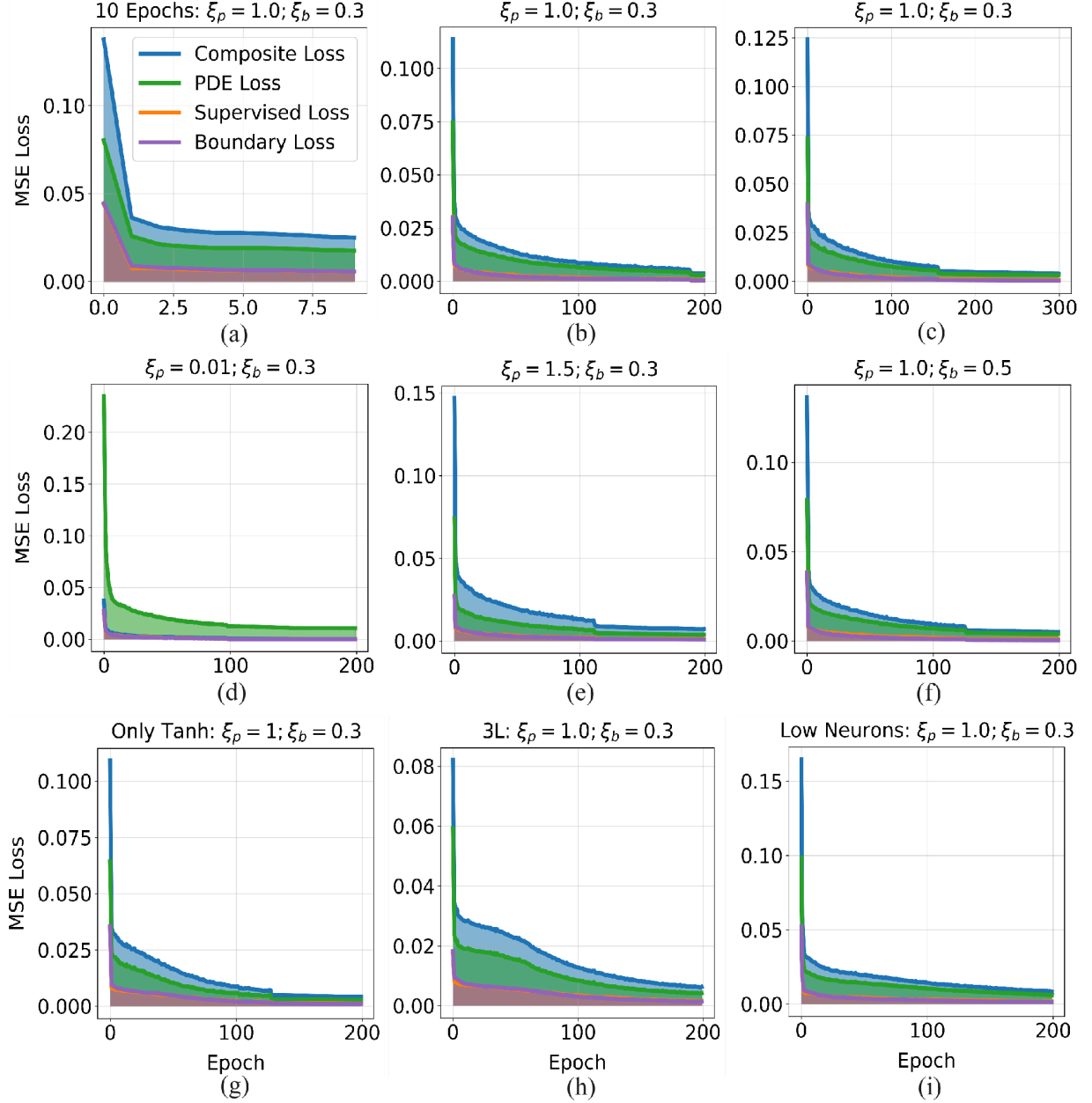


Fig. C1: Recorded hyperparameter sensitivity experiments across nine configurations. Panels (a)–(c) vary training duration; panels (d)–(f) vary physics- and boundary-loss weights; panels (g)–(i) show recorded architecture variants. The physics-loss weight of 0.01 gives lower fitted-scene error than 1.0 in the recorded sweep, whereas the main model retains a weight of 1.0 as described in the main Methods. Architecture labels in this figure describe the recorded sweep configurations and do not define the seven-hidden-layer architecture of the surviving main-model implementation.

Table C1: Recorded hyperparameter sensitivity runs. The table reports training duration, physics- and boundary-loss weights, activation functions, training time, fitted R2 and RMSE against Landsat-derived LST, predicted LST range, and maximum absolute difference. Architecture labels reflect the original supplementary sensitivity record and are not used to define the final main-model architecture.

Recorded scenario	# Epoch	Physics-loss weight	Boundary-loss weight	Activation function	Training time	Period / Landsat-derived LST (°C)	Fitted R2 & RMSE (°C)	Predicted LST (°C)	Maximum absolute difference (°C)
1 (4L, # neurons =(256, 128, 64, 32))	10	1.0	0.3	Tanh & LeakyReLU	3 min 04 s	19 May/ (23.3 - 38.3)	0.7212 & 1.0432	25.0 - 34.1	4.31
						4 June/ (16.2 - 25.6)	0.6965 & 0.7648	17.1 - 22.8	3.04
						20 June/ (21.2 - 37.7)	0.7163 & 1.1712	22.7 - 33.0	6.68
2 (4L)	200	1.0	0.3	Tanh & LeakyReLU	10 min 12 s	19 May/ (23.3 - 38.3)	0.9638 & 0.3758	23.8 - 37.7	2.06
						4 June/ (16.2 - 25.6)	0.9663 & 0.2549	16.2 - 25.6	1.36
						20 June/ (21.2 - 37.7)	0.9676 & 0.3959	21.4 - 37.5	2.06
3 (4L)	300	1.0	0.3	Tanh & LeakyReLU	10 min 30 s	19 May/ (23.3 - 38.3)	0.9605 & 0.3924	24.0 - 38.1	2.59
						4 June/ (16.2 - 25.6)	0.9635 & 0.2653	16.1 - 25.7	1.70
						20 June/ (21.2 - 37.7)	0.9625 & 0.4260	21.3 - 37.7	2.58
4 (4L)	200	0.01	0.3	Tanh & LeakyReLU	10 min 15 s	19 May/ (23.3 - 38.3)	0.9790 & 0.2863	22.3 - 38.2	2.70
						4 June/ (16.2 - 25.6)	0.9783 & 0.2045	15.5 - 25.8	1.58
						20 June/ (21.2 - 37.7)	0.9795 & 0.3148	20.4 - 38.0	2.29
5 (4L)	200	1.5	0.3	Tanh & LeakyReLU	10 min 16 s	19 May/ (23.3 - 38.3)	0.9443 & 0.4664	24.0 - 36.7	2.24
						4 June/ (16.2 - 25.6)	0.9446 & 0.3266	16.4 - 25.0	1.61
						20 June/ (21.2 - 37.7)	0.9480 & 0.5016	21.9 - 36.9	2.51

Recorded scenario	# Epoch	Physics-loss weight	Boundary-loss weight	Activation function	Training time	Period / Landsat-derived LST (°C)	Fitted R2 & RMSE (°C)	Predicted LST (°C)	Maximum absolute difference (°C)
6 (4L)	200	1.0	0.5	Tanh & LeakyReLU	10 min 14 s	19 May/ (23.3 - 38.3)	0.9550 & 0.4190	23.0 - 38.0	2.33
						4 June/ (16.2 - 25.6)	0.9556 & 0.2926	16.0 - 25.6	1.37
						20 June/ (21.2 - 37.7)	0.9595 & 0.4428	21.5 - 37.8	2.60
7 (4L)	200	1.0	0.3	Only Tanh	10 min 30 s	19 May/ (23.3 - 38.3)	0.9359 & 0.5002	23.7 - 36.5	2.87
						4 June/ (16.2 - 25.6)	0.9368 & 0.3490	16.3 - 24.5	1.85
						20 June/ (21.2 - 37.7)	0.9401 & 0.5380	22.2 - 36.3	3.30
8 (3L, # neurons=(256, 128, 64))	200	1.0	0.3	Only Tanh	9 min 25 s	19 May/ (23.3 - 38.3)	0.9198 & 0.5593	23.9 - 36.6	2.83
						4 June/ (16.2 - 25.6)	0.9173 & 0.3993	16.2 - 25.1	1.76
						20 June/ (21.2 - 37.7)	0.9231 & 0.6098	21.9 - 37.1	3.41
9 (Low neurons=(128, 64, 32, 16), 4L)	200	1.0	0.3	Tanh & LeakyReLU	9 min 20 s	19 May/ (23.3 - 38.3)	0.9127 & 0.5838	23.4 - 36.8	2.88
						4 June/ (16.2 - 25.6)	0.9094 & 0.4178	16.1 - 24.8	2.11
						20 June/ (21.2 - 37.7)	0.9147 & 0.6422	21.2 - 36.6	3.24

Supplementary Appendix D: Recorded comparisons with data-driven models and PINN ablations

This appendix records fitted-scene comparisons among neural-network, tree-based, and reduced-PINN configurations. Figures D1 and D2 compare Landsat-derived LST with model predictions and corresponding absolute residual maps. Tables D1 and D2 record configurations and fitted-scene statistics for the three acquisition dates. Because the same Landsat scenes were used during model fitting and evaluation, these comparisons do not constitute independent out-of-sample validation. Some configurations, including the MLP, ConvLSTM, and separate

PINN ablations, are retained in the supplementary record even though their original scripts are not present in the surviving code archive; this limitation is stated in the main Methods.

Fitted-scene performance varies substantially among the recorded data-driven configurations. For example, the deeper U-Net configuration has $R^2 = 0.9974$ on 19 May but $R^2 = 0.6134$ and $RMSE = 1.3673$ °C on 20 June. The supplementary ConvLSTM record contains configurations with lower fitted error on some dates, but the original ConvLSTM script is not present in the surviving archive. These values therefore document the recorded comparison rather than establish relative generalisation performance.

The recorded PINN comparison has R^2 values of 0.9638–0.9676 and RMSE values of 0.2549–0.3959 °C across the three fitted scenes. Removing advection or removing the air- and dewpoint-temperature inputs increases fitted error relative to this recorded PINN on all three dates. In contrast, the No Heat Source variant has lower RMSE and higher R^2 than the recorded PINN on all three dates. The ablation record therefore does not show an in-sample predictive-accuracy benefit from the explicit source-sink term. Incorporating a differential-equation residual into the loss should not be interpreted as guaranteeing physical consistency or out-of-sample generalisation.

Table D2 records Random Forest and XGBoost fitted-scene results. Several higher-capacity tree configurations reproduce the fitted Landsat scenes with very small residuals, including XGBoost configurations with R^2 close to 1.0. Because no independent test set is used, these values alone do not establish either generalisation or overfitting. Figure D3 reports model-specific feature-importance measures for the tree models; these values indicate how the fitted tree models used the input variables and should not be interpreted as causal thermal drivers or as validation of the PINN feature set.

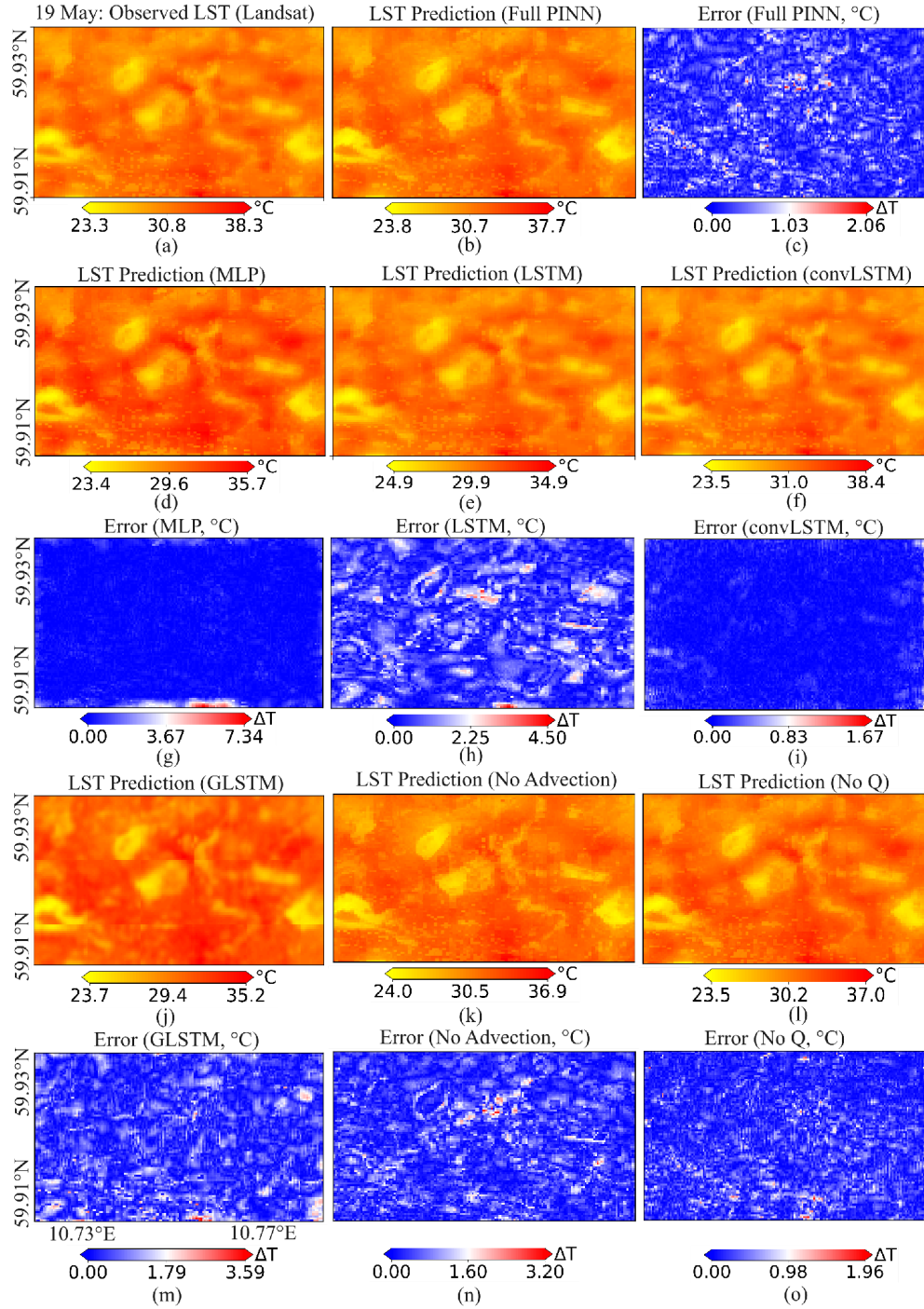


Fig. D1: Landsat-derived LST and recorded model predictions for 19 May, together with corresponding absolute residual maps. The panels include the recorded PINN comparison, MLP, LSTM, ConvLSTM, GLSTM, and reduced PINN variants without advection or the source-sink term. The figure is a fitted-scene spatial comparison and is not an independent validation.

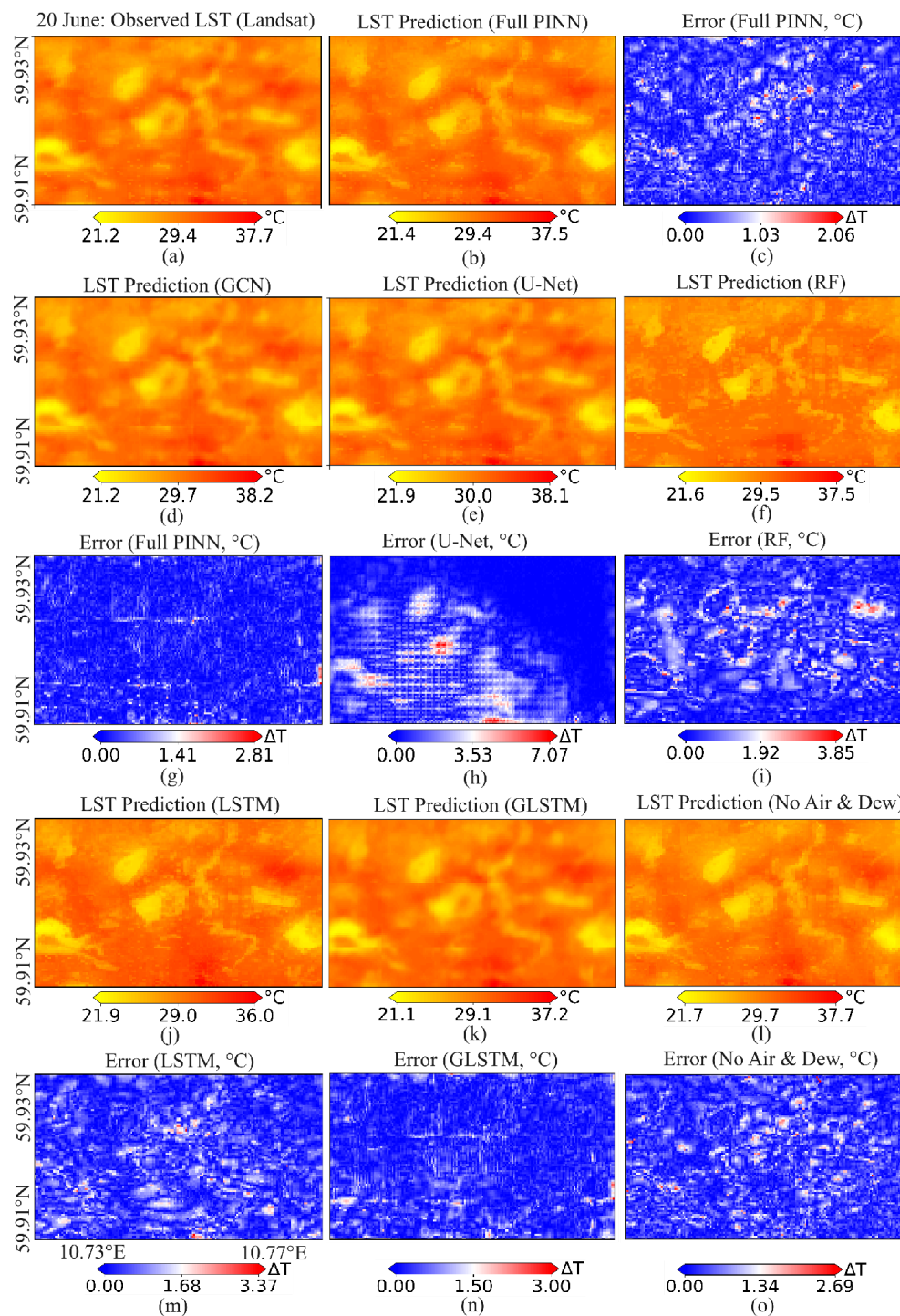


Fig. D2: Landsat-derived LST and recorded model predictions for 20 June, together with corresponding absolute residual maps. The panels include the recorded PINN comparison, GCN, U-Net, Random Forest, LSTM, GLSTM, and the reduced PINN variant without air- and dewpoint-temperature inputs. The figure is a fitted-scene spatial comparison.

Table D1: Recorded deep-learning and PINN comparison configurations and fitted-scene statistics for 19 May, 4 June, and 20 June. The table includes LSTM, ConvLSTM, U-Net, GCN, GLSTM, MLP, the recorded PINN comparison, and reduced PINN variants. Entries labelled as four-layer PINN configurations reproduce the original supplementary record; they should not be used to define the seven-hidden-layer architecture of the surviving main-model implementation described in the main Methods.

Record ed model	# Layers	# Neurons	# Batch size	# Epoch	Training time	Record ed config.	Period / Landsat-derived LST (°C)	Fitted R2 & RMSE (°C)	Predict ed LST (°C)	Maximum absolute error (°C)
LSTM	2	128	512	200	2 min 43 s	Dropout 0.2; Adam.	19 May & (23.3 - 38.3)	0.7655 & 0.9566	24.6 - 34.7	4.21
							4 June & (16.2 - 25.6)	0.8098 & 0.6054	16.2 - 23.6	2.61
							20 June & (21.2 - 37.7)	0.8570 & 0.8317	21.8 - 34.3	3.94
	3	128	512		2 min 52 s		19 May & (23.3 - 38.3)	0.7701 & 0.9472	24.7 - 35.0	4.29
							4 June & (16.2 - 25.6)	0.8431 & 0.5498	16.6 - 24.1	2.43
							20 June & (21.2 - 37.7)	0.8939 & 0.7161	21.4 - 36.6	3.51
	4	128	512		3 min 07 s		19 May & (23.3 - 38.3)	0.7810 & 0.9244	24.8 - 35.5	4.07
							4 June & (16.2 - 25.6)	0.8896 & 0.4614	16.5 - 24.3	2.27
							20 June & (21.2 - 37.7)	0.9290 & 0.5859	21.4 - 36.8	3.10
		64	512		2 min 31 s		19 May & (23.3 - 38.3)	0.7592 & 0.9693	24.9 - 35.1	4.16
							4 June & (16.2 - 25.6)	0.8197 & 0.5895	16.8 - 24.0	2.78
							20 June & (21.2 - 37.7)	0.8728 & 0.7843	22.4 - 36.0	3.64
		512	512		4 min 20 s		19 May & (23.3 - 38.3)	0.9361 & 0.4993	23.8 - 36.7	2.30
							4 June & (16.2 - 25.6)	0.9807 & 0.1930	16.1 - 25.2	1.33
							20 June & (21.2 - 37.7)	0.9830 & 0.2869	20.9 - 37.5	1.66

Record ed model	# Layers	# Neuron s	# Batch size	# Epoch	Trainin g time	Record ed config.	Period / Landsa t-derive d LST (°C)	Fitted R2 & RMSE (°C)	Predict ed LST (°C)	Maxim um absolut e error (°C)	
convLS TM	4	64	8	200	9 min 09 s	32×32 patches; 3×3 kernels.	19 May & (23.3 - 38.3)	0.9934 & 0.1604	23.1 - 38.3	1.67	
							4 June & (16.2 - 25.6)	0.9968 & 0.0790	16.1 - 25.7	1.03	
							20 June & (21.2 - 37.7)	0.9972 & 0.1160	21.1 - 37.6	1.51	
		128			9 min 10 s		19 May & (23.3 - 38.3)	0.9929 & 0.1666	23.3 - 38.5	1.73	
							4 June & (16.2 - 25.6)	0.9958 & 0.0897	16.3 - 25.8	1.28	
							20 June & (21.2 - 37.7)	0.9969 & 0.1220	21.3 - 37.9	2.05	
		256			9 min 15 s		19 May & (23.3 - 38.3)	0.9897 & 0.2006	23.2 - 37.8	1.75	
							4 June & (16.2 - 25.6)	0.9934 & 0.1126	16.0 - 25.4	1.18	
							20 June & (21.2 - 37.7)	0.9957/ 0.1442	20.9 - 37.6	1.62	
U-Net	4	(32, 64, 128, 256)	8	200	6 min 06 s	32×32 patches; enc-dec.	19 May & (23.3 - 38.3)	0.9831 & 0.2565	23.3 - 38.1	2.19	
							4 June & (16.2 - 25.6)	0.9815 & 0.1889	16.1 - 25.7	0.92	
							20 June & (21.2 - 37.7)	0.9848 & 0.2713	21.1 - 37.8	1.26	
		(64, 128, 256, 512)			7 min 04 s		19 May & (23.3 - 38.3)	0.9974 & 0.1006	23.3 - 38.4	1.64	
							4 June & (16.2 - 25.6)	0.8192 & 0.5902	16.2 - 26.0	4.30	
							20 June & (21.2 - 37.7)	0.6134 & 1.3673	21.9 - 38.1	7.07	
GCN	3	128	8	200	5 min 04 s	32×32 patches; 4-nbr graph.	19 May & (23.3 - 38.3)	0.9462 & 0.4582	23.6 - 36.8	2.89	
							4 June & (16.2 - 25.6)	0.9554 & 0.2933	16.1 - 25.1	2.04	
							20 June & (21.2 - 37.7)	0.9604 & 0.4377	21.5 - 37.0	2.86	

Record ed model	# Layers	# Neuron s	# Batch size	# Epoch	Trainin g time	Record ed config.	Period / Landsa t-derive d LST (°C)	Fitted R2 & RMSE (°C)	Predict ed LST (°C)	Maxim um absolut e error (°C)	
	4				5 min 58 s		19 May & (23.3 - 38.3)	0.9624 & 0.3830	23.5 - 38.2	3.04	
							4 June & (16.2 - 25.6)	0.9779 & 0.2064	16.1 - 25.8	1.73	
							20 June & (21.2 - 37.7)	0.9774 & 0.3303	21.2 - 38.2	2.81	
	4	256			9 min 41 s		19 May & (23.3 - 38.3)	0.9678 & 0.3545	23.6 - 38.5	3.01	
							4 June & (16.2 - 25.6)	0.9810 & 0.1911	16.3 - 25.8	1.76	
							20 June & (21.2 - 37.7)	0.9799 & 0.3120	21.3 - 38.3	2.61	
GLSTM	4	128	8 (each containi ng 3 patches)	200	8 min 46 s	32×32 graph patches; recurrent.	19 May & (23.3 - 38.3)	0.8850 & 0.6698	23.7 - 35.2	3.59	
							4 June & (16.2 - 25.6)	0.9463 & 0.3217	16.1 - 25.1	2.10	
							20 June & (21.2 - 37.7)	0.9617 & 0.4305	21.1 - 37.2	3.00	
					8 min 45 s	32×32 graph patches; recurrent model.	19 May & (23.3 - 38.3)	0.8600 & 0.7391	24.3 - 35.1	4.21	
							4 June & (16.2 - 25.6)	0.9323 & 0.3613	16.5 - 24.6	2.49	
							20 June & (21.2 - 37.7)	0.9537 & 0.4732	21.6 - 37.3	3.19	
		256			16 min 40 s	32×32 graph patches; recurrent model.	19 May & (23.3 - 38.3)	0.9323 & 0.5141	23.5 - 36.2	3.28	
							4 June & (16.2 - 25.6)	0.9752 & 0.2186	16.3 - 25.6	1.96	
							20 June & (21.2 - 37.7)	0.9800 & 0.3111	21.5 - 37.6	2.80	
					16 min 45 s	32×32 graph patches; recurrent model.	19 May & (23.3 - 38.3)	0.9008 & 0.5933	23.9 - 35.9	3.53	
							4 June & (16.2 - 25.6)	0.9674 & 0.2507	16.3 - 25.1	2.07	
							20 June & (21.2 - 37.7)	0.9755 & 0.3445	21.4 - 37.5	2.96	

Record ed model	# Layers	# Neurons	# Batch size	# Epoch	Trainin g time	Record ed config.	Period / Landsa t-derive d LST (°C)	Fitted R2 & RMSE (°C)	Predict ed LST (°C)	Maxim um absolut e error (°C)
PINN rec. (4L)	4	(256, 128, 64, 32)	512	200	10 min 12 s	Physics-guided loss.	19 May & (23.3 - 38.3)	0.9638 & 0.3758	23.8 - 37.7	2.06
							4 June & (16.2 - 25.6)	0.9663 & 0.2549	16.2 - 25.6	1.36
							20 June & (21.2 - 37.7)	0.9676 & 0.3959	21.4 - 37.5	2.06
MLP	4	(256, 128, 64, 32)	512	10	3 min 08 s	No physics-guided loss.	19 May & (23.3 - 38.3)	0.8353 & 0.8017	24.7 - 35.0	4.36
				200	8 min 11 s		4 June & (16.2 - 25.6)	0.7866 & 0.6413	16.4 - 23.1	3.03
							20 June & (21.2 - 37.7)	0.8417 & 0.8749	21.6 - 33.7	4.60
							19 May & (23.3 - 38.3)	0.9196 & 0.5600	23.4 - 35.7	7.34
							4 June & (16.2 - 25.6)	0.9002 & 0.4385	16.1 - 24.1	5.00
							20 June & (21.2 - 37.7)	0.9066 & 0.6720	21.0 - 34.6	8.18
No adv.	4	(256, 128, 64, 32)	512	200	12 min 43 s	Advection remove d.	19 May & (23.3 - 38.3)	0.9198 & 0.5595	24.0 - 36.9	3.20
							4 June & (16.2 - 25.6)	0.9053 & 0.4273	16.3 - 24.9	1.79
							20 June & (21.2 - 37.7)	0.9277 & 0.5912	21.7 - 36.9	3.62
No Q	4	(256, 128, 64, 32)	512	200	12 min 27 s	Q* term remove d.	19 May & (23.3 - 38.3)	0.9756 & 0.3086	23.5 - 37.0	1.96
							4 June & (16.2 - 25.6)	0.9766 & 0.2125	15.9 - 25.6	1.54
							20 June & (21.2 - 37.7)	0.9750 & 0.3477	20.9 - 37.3	3.72
No T2/Td2	4	(256, 128, 64, 32)	512	200	13 min 04 s	T2m/Td2m remove d.	19 May & (23.3 - 38.3)	0.9150 & 0.5758	23.9 - 36.5	3.11
							4 June & (16.2 - 25.6)	0.9396 & 0.3412	16.2 - 25.4	1.67
							20 June & (21.2 - 37.7)	0.9451 & 0.5154	21.7 - 37.7	2.69

Table D2: Recorded Random Forest and XGBoost configurations and fitted-scene statistics for the same three Landsat acquisitions. The table reports estimator count, maximum depth, sampling settings, training time, fitted R2 and RMSE, predicted LST range, and maximum absolute error.

Recorded tree model	# Estimators	# Max depth	Training time	Recorded config.	Period / Landsat-derived LST (°C)	Fitted R2 & RMSE (°C)	Predicted LST (°C)	Maximum absolute error (°C)
RF	200	10	2 min 05 s	Bootstrap RF; depth varied.	19 May & (23.3 - 38.3)	0.8632 & 0.7307	23.7 - 38.1	3.38
					4 June & (16.2 - 25.6)	0.8399 & 0.5554	16.7 - 25.5	2.97
					20 June & (21.2 - 37.7)	0.8745 & 0.7791	21.6 - 37.5	3.84
		30	2 min 18 s		19 May & (23.3 - 38.3)	0.9935 & 0.1588	23.4 - 38.2	1.57
					4 June & (16.2 - 25.6)	0.9938 & 0.1089	16.2 - 25.6	1.08
					20 June & (21.2 - 37.7)	0.9942 & 0.1681	21.2 - 37.6	1.67
	500	10	2 min 15 s		19 May & (23.3 - 38.3)	0.8635 & 0.7297	23.7 - 38.1	3.38
					4 June & (16.2 - 25.6)	0.8403& 0.5548	16.7 - 25.5	2.96
					20 June & (21.2 - 37.7)	0.8750 & 0.774	21.6 - 37.5	3.82
	1000	10	2 min 40 s		19 May & (23.3 - 38.3)	0.8635 & 0.7298	23.7 - 38.1	3.42
					4 June & (16.2 - 25.6)	0.8402 & 0.5550	16.7 - 25.5	2.96
					20 June & (21.2 - 37.7)	0.8748 & 0.7780	21.6 - 37.5	3.85
XGBoost	200	10	1 min 41 s	sub _{samp} =0.8; col _{samp} =0.8	19 May & (23.3 - 38.3)	0.9975 & 0.0990	23.4 - 38.3	0.79
					4 June & (16.2 - 25.6)	0.9973 & 0.0722	16.2 - 25.6	0.42
					20 June & (21.2 - 37.7)	0.9978 & 0.1042	21.2 - 37.6	0.63

Recorded tree model	# Estimators	# Max depth	Training time	Recorded config.	Period / Landsat-derived LST (°C)	Fitted R2 & RMSE (°C)	Predicted LST (°C)	Maximum absolute error (°C)
			1 min 15 s	sub _{samp} =1.0; col _{samp} =0.8	19 May & (23.3 - 38.3)	0.9972 & 0.1039	23.5 - 38.3	0.77
					4 June & (16.2 - 25.6)	0.9969 & 0.0777	16.3 - 25.6	0.50
					20 June & (21.2 - 37.7)	0.9976 & 0.1069	21.3 - 37.6	0.78
			1 min 17 s	sub _{samp} =0.8; col _{samp} =1.0	19 May & (23.3 - 38.3)	0.9976 & 0.0977	23.4 - 38.2	0.64
					4 June & (16.2 - 25.6)	0.9973 & 0.0723	16.2 - 25.6	0.42
					20 June & (21.2 - 37.7)	0.9977 & 0.1049	21.3 - 37.6	0.75
			1 min 23 s	sub _{samp} =1.0; col _{samp} =1.0	19 May & (23.3 - 38.3)	0.9971 & 0.1068	23.4 - 38.3	0.95
					4 June & (16.2 - 25.6)	0.9967 & 0.0798	16.2 - 25.6	0.43
					20 June & (21.2 - 37.7)	0.9973 & 0.1140	21.3 - 37.6	0.74
	1000	10	1 min 44 s	sub _{samp} =0.8; col _{samp} =0.8	19 May & (23.3 - 38.3)	0.9999 & 0.0162	23.3 - 38.3	0.08
					4 June & (16.2 - 25.6)	0.9999 & 0.0120	16.2 - 25.6	0.06
					20 June & (21.2 - 37.7)	0.9999 & 0.0181	21.2 - 37.7	0.09
			1 min 25 s	sub _{samp} =1.0; col _{samp} =0.8	19 May & (23.3 - 38.3)	0.9997 & 0.0353	23.3 - 38.3	0.23
					4 June & (16.2 - 25.6)	0.9996 & 0.0264	16.2 - 25.6	0.18
					20 June & (21.2 - 37.7)	0.9997 & 0.0383	21.2 - 37.7	0.22
			1 min 27 s	sub _{samp} =0.8; col _{samp} =1.0	19 May & (23.3 - 38.3)	0.9999 & 0.0161	23.3 - 38.3	0.08
					4 June & (16.2 - 25.6)	0.9999 & 0.0119	16.2 - 25.6	0.05
					20 June & (21.2 - 37.7)	0.9999 & 0.0181	21.2 - 37.7	0.09
			1 min 25 s	sub _{samp} =1.0; col _{samp} =1.0	19 May & (23.3 - 38.3)	0.9996 & 0.0374	23.3 - 38.3	0.26

Recorded tree model	# Estimators	# Max depth	Training time	Recorded config.	Period / Landsat-derived LST (°C)	Fitted R2 & RMSE (°C)	Predicted LST (°C)	Maximum absolute error (°C)
					4 June & (16.2 - 25.6)	0.9996 & 0.0276	16.2 - 25.6	0.19
					20 June & (21.2 - 37.7)	0.9997 & 0.0400	21.2 - 37.7	0.29

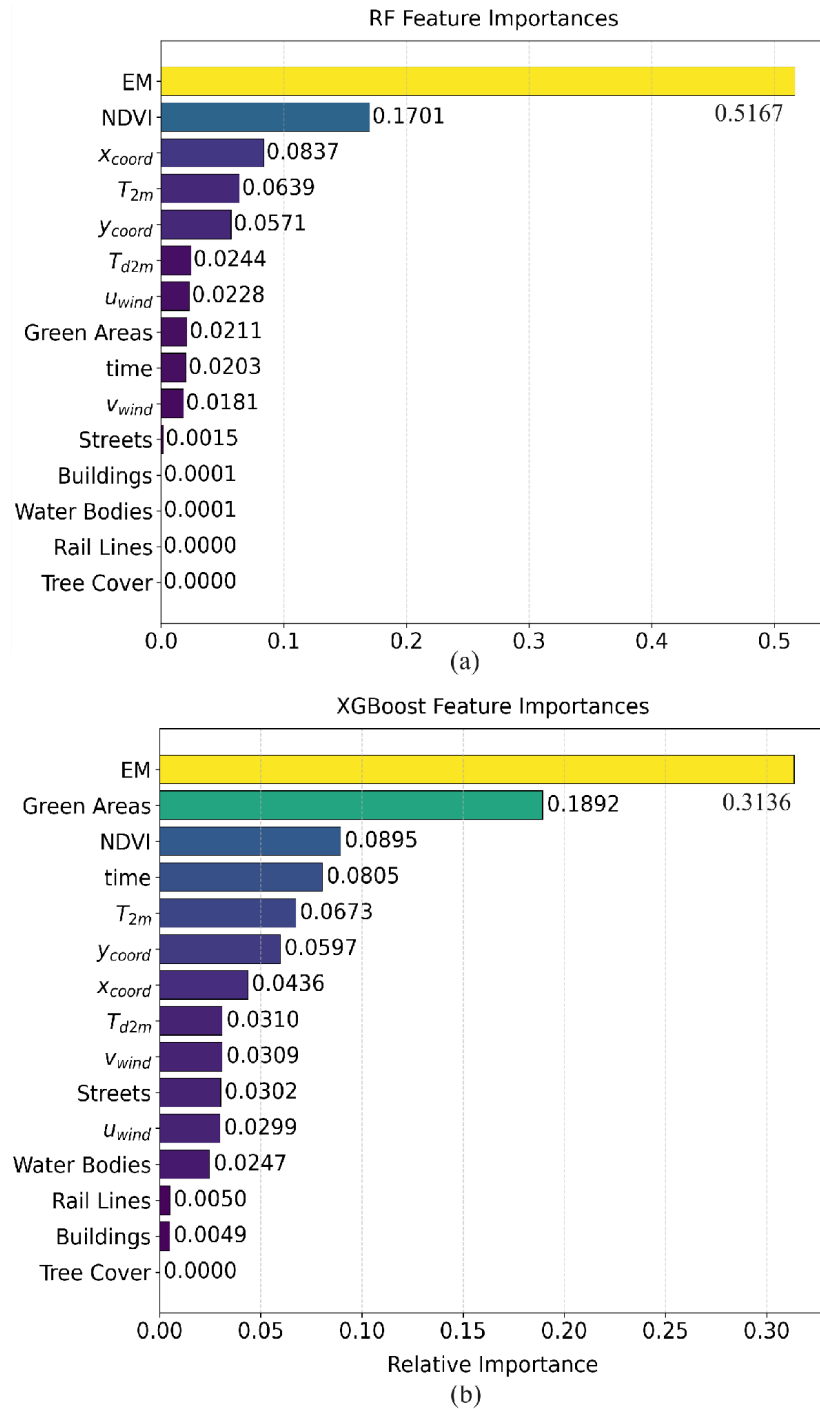


Fig. D3: Feature-importance measures from the recorded Random Forest and XGBoost fits. Panel (a) shows Random Forest feature importance and panel (b) XGBoost feature importance. These are model-specific importance measures for the fitted tree models and are not interpreted as causal effects.

Supplementary Appendix E: NDVI and emissivity distributions for eligible LST-hotspot building pixels

This appendix shows the original and IQR-screened NDVI and emissivity (EM) distributions for eligible building pixels located within the observed Landsat LST hotspot classes. Hotspots were identified from the Landsat-derived LST distributions and then intersected with the building mask. The candidate set was divided into pixels inside and outside the near-rail group defined by the 15-pixel square dilation of the rasterised railway mask.

Figure E1 shows the distributions for eligible hotspot-building pixels outside the near-rail group. The original and IQR-screened distributions document the surface-property values retained before the prescribed scenario values were assigned.

Figure E2 shows the corresponding distributions for the near-rail candidate group. The railway grouping is a spatial classification used in the analysis; differences between the groups are not interpreted here as causal effects of railway infrastructure on local microclimate.

Figures E1 and E2 therefore document the NDVI and emissivity distributions used in candidate screening before the surface-property scenarios. They do not establish roof material, vegetation process, or complete green-roof thermal behaviour.

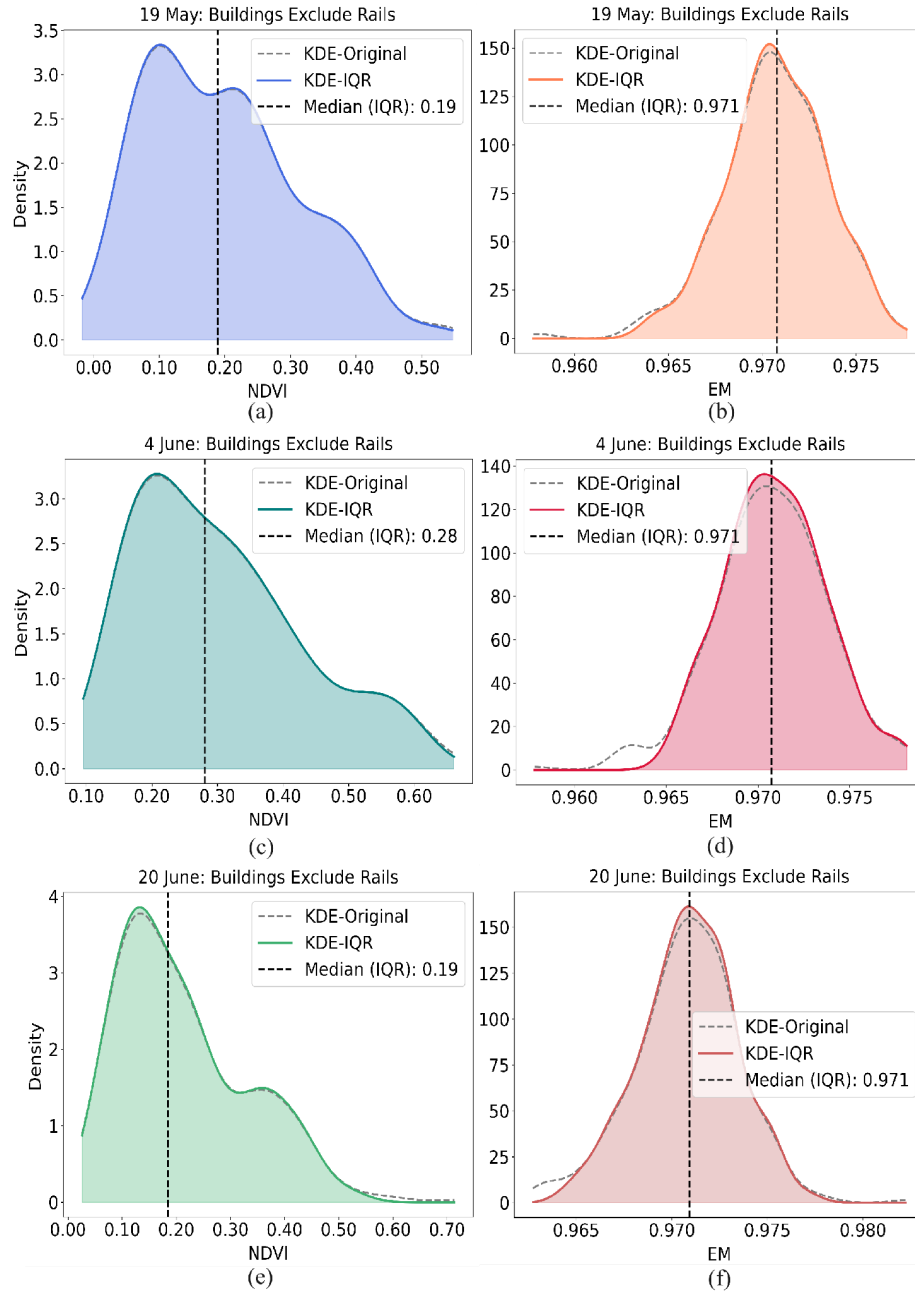


Fig. E1: NDVI and emissivity (EM) distributions for eligible LST-hotspot building pixels outside the near-rail group. (a) 19 May NDVI, median of the IQR-screened distribution 0.19; (b) 19 May EM, median 0.971; (c) 4 June NDVI, median 0.28; (d) 4 June EM, median 0.971; (e) 20 June NDVI, median 0.19; and (f) 20 June EM, median 0.971. KDE-Original and KDE-IQR denote the original and IQR-screened distributions shown in the figure.

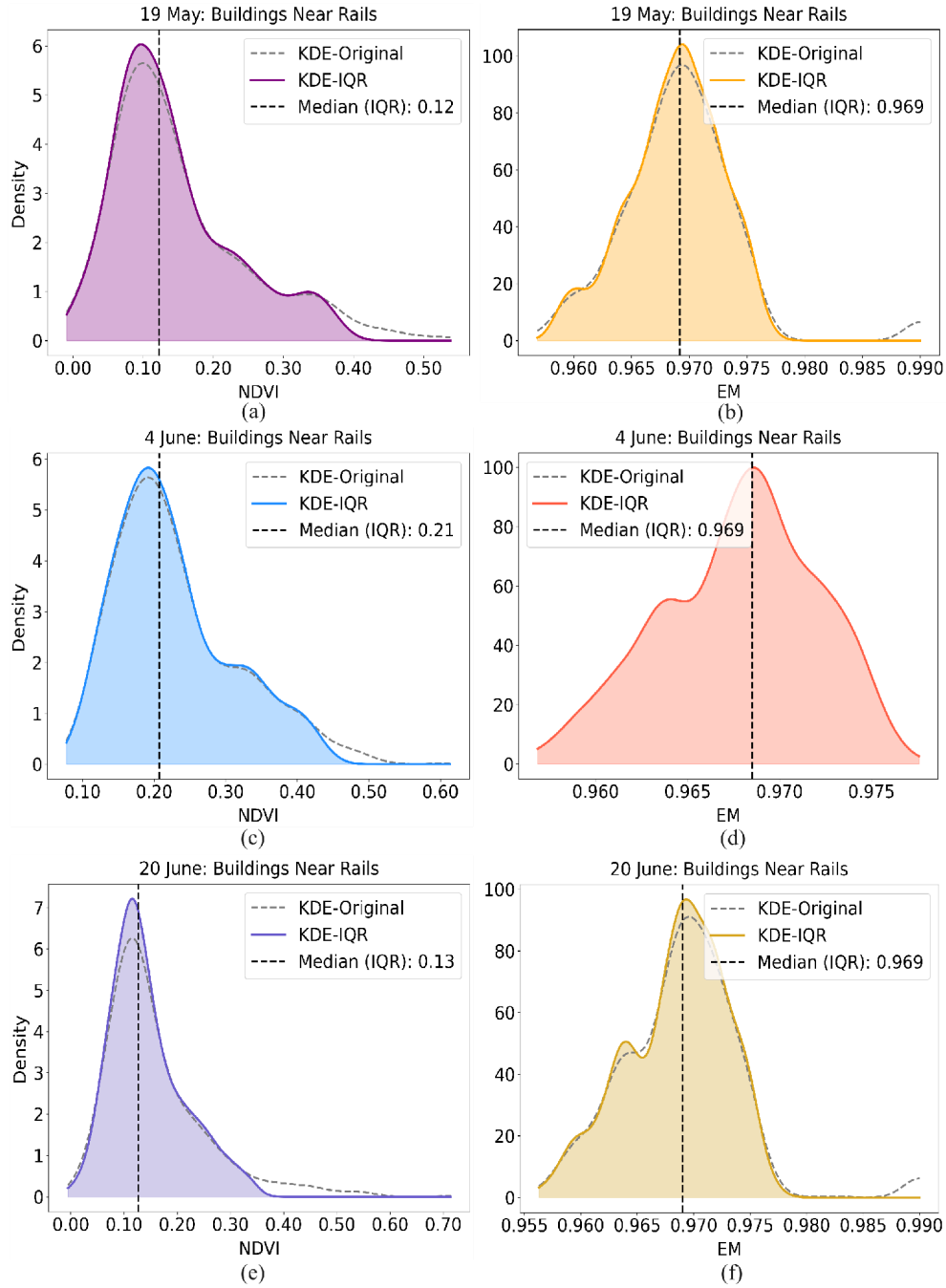


Fig. E2: NDVI and emissivity (EM) distributions for eligible LST-hotspot building pixels in the near-rail group. (a) 19 May NDVI, median of the IQR-screened distribution 0.12; (b) 19 May EM, median 0.969; (c) 4 June NDVI, median 0.21; (d) 4 June EM, median 0.969; (e) 20 June NDVI, median 0.13; and (f) 20 June EM, median 0.969. Line styles follow Fig. E1.

Supplementary Appendix F: NDVI = 0.55 surface-property scenario

This appendix provides additional spatial outputs for the prescribed NDVI = 0.55 surface-property scenario. For each acquisition date, the fitted PINN was first evaluated with the original inputs to obtain a baseline modelled LST field. NDVI and emissivity were then replaced at the eligible hotspot-building pixels included in the scenario, while the remaining model inputs and trained network parameters were unchanged. The resulting differences are responses of the fitted PINN to the prescribed surface-property changes rather than predictions from a complete green-roof process model.

Figure F1 presents the baseline PINN LST, scenario PINN LST, and pixel-wise temperature change for the candidate group outside the near-rail classification on 19 May, 4 June, and 20 June. The corresponding distributions summarise modelled temperature change over the affected pixels. Negative ΔT indicates lower modelled LST after the prescribed surface-property change.

Figure F2 presents the same outputs for the near-rail candidate group. The comparison between the two spatial groups is descriptive; the grouping itself does not establish that railway proximity causes differences in model response.

These figures should be interpreted within the three observed acquisition states only. They do not establish performance across seasons, represent an upper bound for citywide heat mitigation, or quantify complete green-roof cooling.

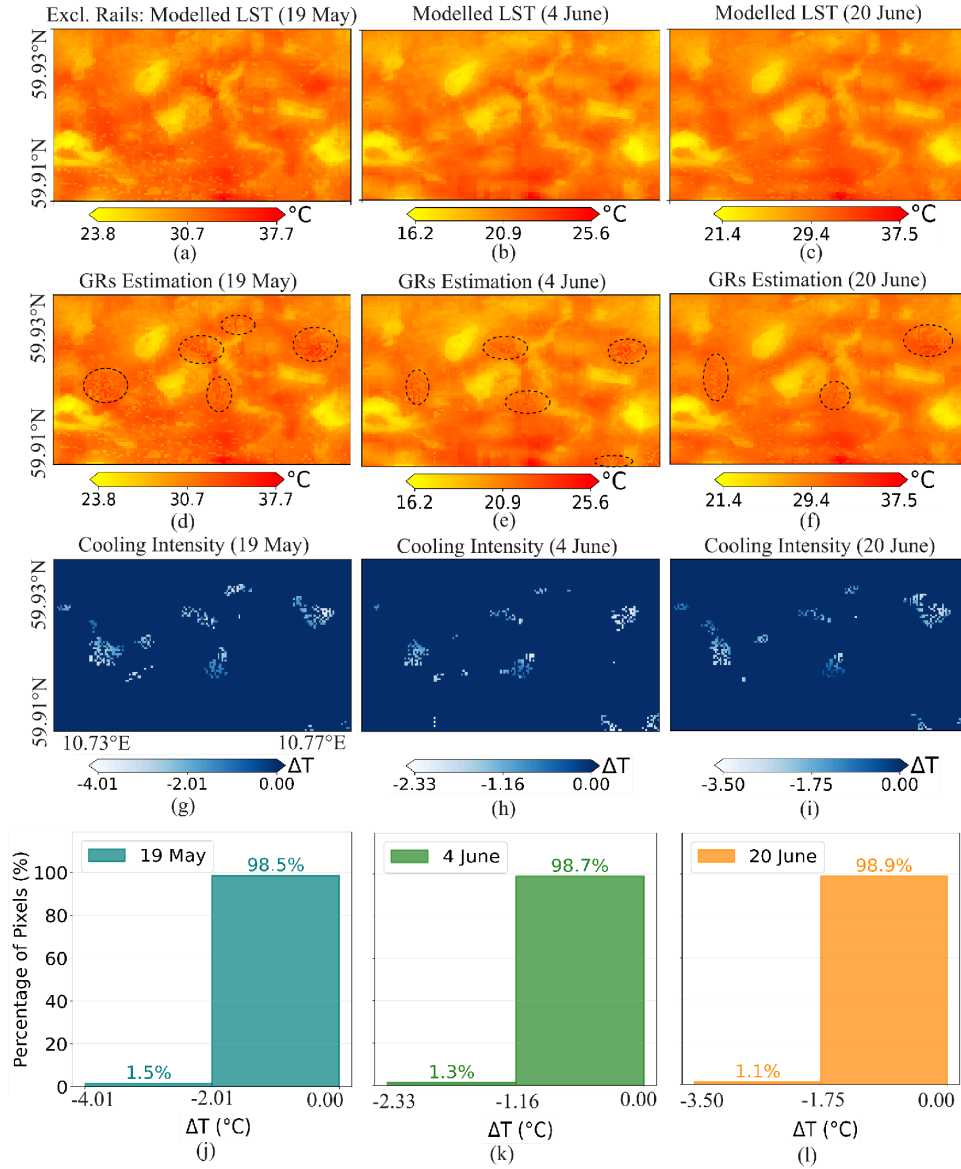


Fig. F1: Baseline PINN LST, scenario PINN LST, and modelled temperature change for the NDVI = 0.55 surface-property scenario for eligible hotspot-building pixels outside the near-rail group. Panels (a)–(c) show baseline PINN LST for 19 May, 4 June, and 20 June; panels (d)–(f) show scenario PINN LST at the prescribed pixels; panels (g)–(i) show ΔT ; and panels (j)–(l) show the corresponding ΔT distributions. Negative ΔT indicates lower modelled LST.

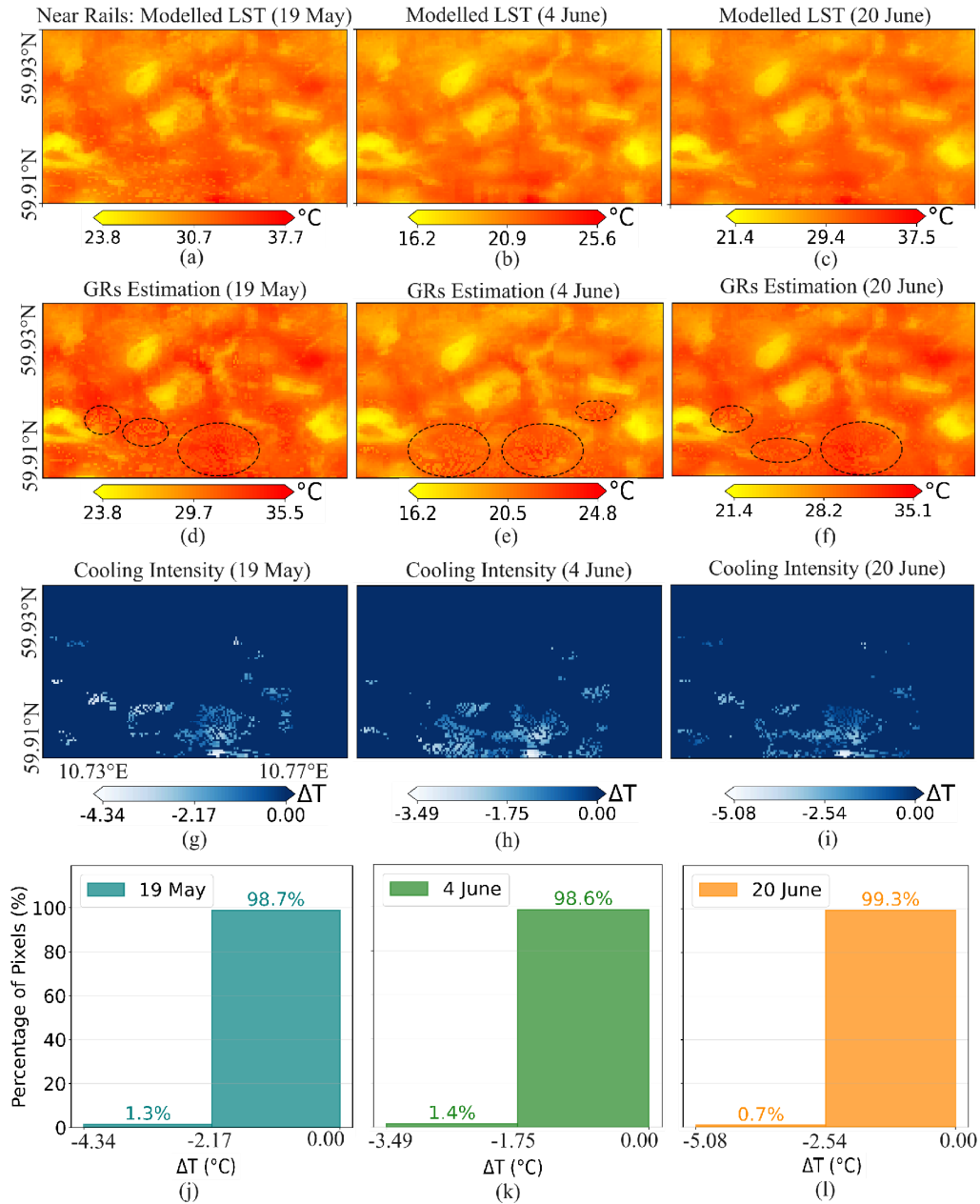


Fig. F2: As Fig. F1, but for eligible hotspot-building pixels in the near-rail group. Panels show baseline PINN LST, scenario PINN LST, modelled ΔT , and the corresponding ΔT distributions for the three acquisition dates. Negative ΔT indicates lower modelled LST.

Supplementary Appendix G: NDVI = 0.75 deployment and spatial prioritisation

This appendix reports coverage-response results for the prescribed NDVI = 0.75 surface-property scenario. Coverage is defined relative to the eligible candidate pixels for each acquisition date and railway-proximity group and ranges from 15% to 100%; it is not the fraction of total roof area or of all buildings in Oslo. Figure G1 compares two proximity-based ranking rules. The

park/water rule orders candidates by Euclidean raster distance to the combined park-and-water mask, and the railway rule orders candidates by Euclidean raster distance to the railway mask.

For the park/water-proximity ranking, mean ΔT varies with acquisition date and candidate group as coverage increases. The curves therefore describe the response of the selected eligible pixels under each fitted thermal state rather than a citywide temperature change.

For the railway-proximity ranking, the near-rail group on 20 June has a mean ΔT of approximately $-2.576\text{ }^{\circ}\text{C}$ at 15% coverage and $-2.084\text{ }^{\circ}\text{C}$ at 100% coverage. The corresponding 15% value is the more negative of the two proximity-based rules for that date and group. This comparison is a ranking result within the eligible candidate set and is not evidence of a causal railway effect.

Across the recorded date-group combinations, neither proximity-based ranking produces the more negative mean ΔT in every case. At 100% coverage, the proximity rankings converge within a given date and candidate group because all eligible pixels in that group have been selected.

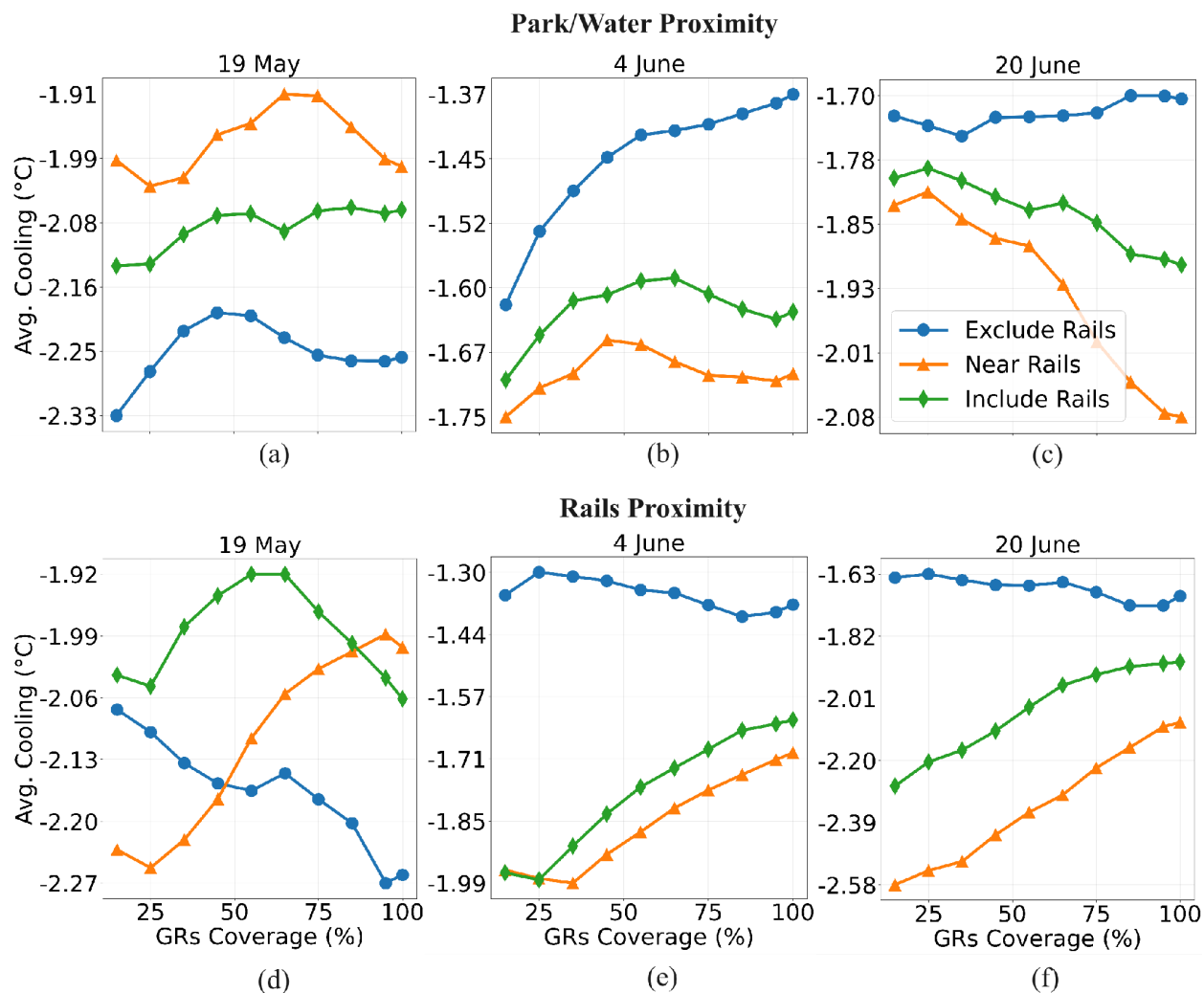


Fig. G1: Mean modelled temperature change (ΔT) over selected eligible hotspot-building pixels under the NDVI = 0.75 surface-property scenario, comparing park/water-proximity and railway-proximity ranking as coverage increases from 15% to 100% of the eligible candidate set. Negative ΔT indicates lower modelled LST. Panels (a)–(c) show park/water-proximity ranking for 19 May, 4 June, and 20 June; panels (d)–(f) show railway-proximity ranking for the same dates. In panel (f), the near-rail group reaches approximately -2.58 °C at 15% coverage.