KDD Ablation Study - Test Evaluation Report

WandB Project: AAAI-Project-channel-ablation-test

Report Date: February 07, 2026

Validation: 2022-2023 | Test: 2024-2025 | Train: 2013-2021 | Input: 12 months | Output: 1 month | Max NODATA: 50%

1. Channel Ablation (cluster = all)

Comparison of feature subsets using two architectures. All models trained on cluster=all (124 US cities). The Earthformer is a CuboidTransformer; the CNN-LSTM is a DMVSTNet.

1.1 Earthformer (CuboidTransformer)

WandB Run (Val / Test)	Feature Set	Ch	Val RMSE (2022-23)	Test RMSE (2024-25)	Val MAE (2022-23)	Test MAE (2024-25)
firm-mountain-8 / polished-firebrand-9	All Features	9	12.83	13.93	9.10	9.53
avid-moon-18 / fresh-snowball-19	Spectral (no LST)	8	13.42	13.89	9.47	9.43
young-smoke-20 / sandy-dragon-21	LST Only	1	13.69	14.67	9.15	9.68
worthy-dream-22 / sandy-eon-23	RGB Only	3	14.60	15.62	10.13	10.13

Checkpoint provenance: call-lucky-water-86 (All), cerulean-deluge-44 (Spectral), laced-deluge-43 (LST Only), swift-waterfall-4 (RGB Only). All from AAAI-Project-channel-ablation / quick-test.

1.2 CNN-LSTM (DMVSTNet)

WandB Run (Val / Test)	Feature Set	Ch	Val RMSE (2022-23)	Test RMSE (2024-25)	Val MAE (2022-23)	Test MAE (2024-25)
northern-rain-32 / fresh-bird-33	All Features	9	20.39	18.75	15.54	13.39
glowing-resonance-2 / jolly-fire-3	Spectral (no LST)	8	22.80	22.44	17.64	16.69
floral-deluge-4 / summer-haze-5	LST Only	1	20.78	19.96	15.94	14.52
skilled-dawn-6 / true-wave-7	RGB Only	3	27.18	26.68	21.93	20.40

Checkpoint provenance: lstm-genial-sky-7/all.ckpt (All Features, md5: 91bffc32), valiant-river-73 (Spectral), dark-sunset-72 (LST Only), mild-music-71 (RGB Only). All.ckpt verified byte-for-byte against WandB artifact legendary-resonance-2/epoch=13.ckpt (3,249,949,000 bytes).

2. Cluster-Specific Evaluation (9 channels, All Features)

Each model evaluated on data filtered to a specific land-use cluster. Earthformer models were trained per-cluster; CNN-LSTM uses a single model (all.ckpt) evaluated on each cluster subset.

2.1 Earthformer per-cluster

WandB Run (Val / Test)	Cluster	Val RMSE (2022-23)	Test RMSE (2024-25)	Val MAE (2022-23)	Test MAE (2024-25)
firm-mountain-8 / polished-firebrand-9	all	12.83	13.93	9.10	9.53

generous-shadow-24 / rare-hill-25	1	17.54	22.82	12.49	18.11
efficient-universe-26 / ethereal-wood-27	2	12.65	15.14	8.47	9.21
tough-pyramid-28 / lunar-breeze-29	3	16.65	12.59	10.74	8.56
visionary-fire-30 / helpful-fog-31	4	12.80	11.84	9.13	8.87

Checkpoints: call-lucky-water-86 (all), c1-cosmic-waterfall-85 (1), c2-spring-darkness-87 (2), c3-fallen-fire-88 (3), C4-woven-music-89 (4).

2.2 CNN-LSTM per-cluster (single model)

WandB Run (Val / Test)	Cluster	Val RMSE (2022-23)	Test RMSE (2024-25)	Val MAE (2022-23)	Test MAE (2024-25)
northern-rain-32 / fresh-bird-33	all	20.39	18.75	15.54	13.39
atomic-mountain-34 / fanciful-grass-35	1	17.83	15.52	15.36	14.55
swept-frog-36 / brisk-glitter-37	2	20.34	19.48	14.97	13.28
sunny-disco-38 / woven-flower-39	3	20.67	14.42	15.78	10.87
polished-silence-40 / lucky-smoke-41	4	20.75	17.61	16.26	13.37

Checkpoint: lstm-genial-sky-7/all.ckpt for all cluster evaluations (single model, cluster=all training, evaluated on each cluster subset).

3. Head-to-Head Summary: Earthformer vs CNN-LSTM

Direct comparison using test set (2024-2025) RMSE in degrees Fahrenheit. All models use 9 input channels.

Cluster	LCZ Classes	Description	Earthformer RMSE (F)	CNN-LSTM RMSE (F)	Delta
all	0-17	All Lands	**13.93**	18.75	-4.82
1	1-3	Dense buildings, sparse green	22.82	**15.52**	+7.30
2	4-6	Less dense buildings, more green	**15.14**	19.48	-4.34
3	7-10	Remaining urban classes	**12.59**	14.42	-1.83
4	11-17,0	Natural landscapes	**11.84**	17.61	-5.77

4. Key Observations

1. **Earthformer dominates overall.** On the full dataset (cluster=all), Earthformer achieves 13.93F vs CNN-LSTM 18.75F on the 2024-2025 test set, a 4.82F improvement.

2. **CNN-LSTM wins on Cluster 1 (dense urban).** The CNN-LSTM achieves 15.52F vs Earthformer 22.82F. This is a +7.30F advantage for the simpler architecture on compact high-rise areas (LCZ 1-3). This warrants further investigation.

3. **Earthformer channel ablation shows robustness.** Removing LST (13.89F) barely degrades performance vs all features (13.93F), suggesting spectral features carry strong predictive signal. CNN-LSTM is more sensitive to feature selection (18.75F all vs 22.44F spectral-only).

4. **Generalization gap varies by cluster.** Cluster 1 shows large val-to-test degradation for Earthformer (17.54 to 22.82F), while Clusters 3 and 4 actually improve on the test set. CNN-LSTM shows more consistent val-to-test behavior.

5. **All CNN-LSTM cluster results use a single model** (all.ckpt, legendary-resonance-2), verified byte-for-byte against WandB (md5: 91bffc32). Per-cluster numbers reflect the same model evaluated on different data subsets, not separate cluster-trained models.

5. Complete Run Index

#	Run Name	Model	Cluster	Channels	Years	RMSE (F)	MAE (F)
1	honest-silence-1	lstm	all	8 (no LST)	22-23	FAILED	-
2	glowing-resonance-2	lstm	all	8 (no LST)	22-23	22.80	17.64
3	jolly-fire-3	lstm	all	8 (no LST)	24-25	22.44	16.69
4	floral-deluge-4	lstm	all	1 (LST)	22-23	20.78	15.94
5	summer-haze-5	lstm	all	1 (LST)	24-25	19.96	14.52
6	skilled-dawn-6	lstm	all	3 (RGB)	22-23	27.18	21.93
7	true-wave-7	lstm	all	3 (RGB)	24-25	26.68	20.40
8	firm-mountain-8	earthnet	all	9 (All)	22-23	12.83	9.10
9	polished-firebrand-9	earthnet	all	9 (All)	24-25	13.93	9.53
18	avid-moon-18	earthnet	all	8 (no LST)	22-23	13.42	9.47
19	fresh-snowball-19	earthnet	all	8 (no LST)	24-25	13.89	9.43
20	young-smoke-20	earthnet	all	1 (LST)	22-23	13.69	9.15
21	sandy-dragon-21	earthnet	all	1 (LST)	24-25	14.67	9.68
22	worthy-dream-22	earthnet	all	3 (RGB)	22-23	14.60	10.13
23	sandy-eon-23	earthnet	all	3 (RGB)	24-25	15.62	10.13
24	generous-shadow-24	earthnet	1	9 (All)	22-23	17.54	12.49
25	rare-hill-25	earthnet	1	9 (All)	24-25	22.82	18.11
26	efficient-universe-26	earthnet	2	9 (All)	22-23	12.65	8.47
27	ethereal-wood-27	earthnet	2	9 (All)	24-25	15.14	9.21
28	tough-pyramid-28	earthnet	3	9 (All)	22-23	16.65	10.74
29	lunar-breeze-29	earthnet	3	9 (All)	24-25	12.59	8.56
30	visionary-fire-30	earthnet	4	9 (All)	22-23	12.80	9.13
31	helpful-fog-31	earthnet	4	9 (All)	24-25	11.84	8.87
32	northern-rain-32	lstm	all	9 (All)	22-23	20.39	15.54
33	fresh-bird-33	lstm	all	9 (All)	24-25	18.75	13.39
34	atomic-mountain-34	lstm	1	9 (All)	22-23	17.83	15.36
35	fanciful-grass-35	lstm	1	9 (All)	24-25	15.52	14.55
36	swept-frog-36	lstm	2	9 (All)	22-23	20.34	14.97
37	brisk-glitter-37	lstm	2	9 (All)	24-25	19.48	13.28
38	sunny-disco-38	lstm	3	9 (All)	22-23	20.67	15.78
39	woven-flower-39	lstm	3	9 (All)	24-25	14.42	10.87
40	polished-silence-40	lstm	4	9 (All)	22-23	20.75	16.26
41	lucky-smoke-41	lstm	4	9 (All)	24-25	17.61	13.37