Supplementary Information

Cost-Aware Evaluation of Time-Series Foundation Models for Urban Air-Quality and Temperature Forecasting

This document contains the supplementary tables and figures referenced from the main text, together with extended methods that support but do not carry the main-text argument. Table S1 lists every abbreviation used in the paper.

Contents

Abbreviations

Table S1: Abbreviations and full forms. Short forms used in the main text, the tables, and the figure captions.

Abbreviation	Full form
BH	Benjamini–Hochberg (false-discovery-rate procedure)
CI	Confidence interval
CPU	Central processing unit
DM	Diebold–Mariano test of equal predictive accuracy
ERA5	Fifth-generation ECMWF atmospheric reanalysis
FDR	False discovery rate
FM	Foundation model
GPU	Graphics processing unit
GRU	Gated recurrent unit
HLN	Harvey–Leybourne–Newbold small-sample correction to the DM test
Holm	Holm step-down multiple-comparison correction
kJ	Kilojoule
LightGBM	Light Gradient Boosting Machine (gradient-boosted tree ensemble)
MAE	Mean absolute error
MASE	Mean absolute scaled error
NAS	Neural architecture search
NAS-GRU	The GRU architecture discovered by NAS and used as the *search* tier
NSGA-II	Non-dominated Sorting Genetic Algorithm II (multi-objective search)
NVML	NVIDIA Management Library (GPU power telemetry)
NWP	Numerical weather prediction
OpenAQ	Open Air Quality, the open air-quality data platform
$PM_{2.5}$	Particulate matter with aerodynamic diameter ≤ 2.5 μm
PUE	Power usage effectiveness (data-centre overhead multiplier)
RMSE	Root mean squared error
sd	Standard deviation
TOST	Two one-sided tests (equivalence testing procedure)
TSFM	Time-series foundation model
USD/1k	United States dollars per 1,000 forecasts
UTC	Coordinated Universal Time
h	Forecast horizon, in hours
α	Significance level (0.05 throughout)
δ	Equivalence margin for TOST, in MASE units (0.05 in main-text claims)
λ	Cost-penalty coefficient: MASE units traded per USD per 1,000 forecasts
m	Seasonal period of the naive baseline, in hours

Supplementary methods

Extended data-quality gating

The 985 µg m^{-3} ceiling. This rule targets a repeated sensor/API *sentinel* value rather than high pollution as such. The value 985 recurs as an exact, isolated spike (identical to the digit in the raw feed, with no neighbouring elevated readings) across 12 of the 29 raw city files, which is the signature of a saturation or error code rather than a genuine extreme episode. The case for the rule therefore rests on that sentinel pattern, not on 985 being implausibly high as a concentration. Both the ≤ 0 and ≥ 985 rules were fixed before the campaign, and panel membership is insensitive to the exact ceiling.

Excluded cities. Three candidate cities failed the pre-specified gates. London's sensor carries only 247 real hourly readings over its nominal 2016–2026 span; that sparsity is a property of the source record rather than of our retrieval. Kathmandu and Kolkata both fail the minimum-hours gate because of internal gap structure, confirmed by direct inspection.

Retrieval. OpenAQ's deep pagination is unstable for large requests, so fetches were month-windowed, resumable, and checkpointed, with per-request verification that the reported record count matched the records received.

Non-contemporaneous windows and pandemic-period exposure

Each city contributes whichever usable window its source record supports, so the panel is not contemporaneous: windows span August 2016 to July 2026 (Table S2). Ten of the 29 cities have windows overlapping 2020–21, accounting for 85,298 of the 249,070 usable hours in the panel (34%). Three lie wholly inside that window: Mumbai (2020-08-31 to 2021-01-26), Accra (2020-05-01 to 2021-04-24) and Bogotá (2020-02-10 to 2020-09-22). Seven lie partly inside it: Sydney (97% of its window), Hanoi (81%), Karachi (81%), Nairobi (76%), Jakarta (67%), Santiago (49%) and Toronto (36%). Sydney's window also spans the 2019–20 Australian bushfire smoke episode, an extreme PM2.5 anomaly in its own right. We did not exclude this period and did not fit any regime term for it; the figures above are produced by `analysis/pandemic_exposure.py`.

Two features of the design limit what that exposure can do to the reported comparisons. First, every comparison here is within-city and between-tier. MASE is computed against a per-series seasonal-naïve denominator fixed on that city's own pre-test history, and all tiers are evaluated on identical test folds, so a shift in the local pollution regime rescales the difficulty of the forecasting problem for every tier at once rather than favouring one of them. Second, the post-cutoff check in Table S14 evaluates only strictly post-2024-10 windows, which lie entirely outside the pandemic period, and reproduces the panel conclusion: mean MASE 0.415 for zero-shot Chronos-Bolt against 0.395 for the LightGBM specialist, with the foundation model better in 6 of the 10 cities that have sufficient recent coverage.

One caveat we cannot dismiss. The exposure is not balanced across the design: seven of the ten affected cities are scarce-tier and three are rich-tier, so it is correlated with the tier contrast the paper reports. A reader who believes pandemic-era emissions made urban PM2.5 systematically easier or harder to forecast in a tier-specific way should read the rich/scarce comparison with that in mind. The post-cutoff check is the closest available evidence that the conclusion does not depend on the pandemic-era cities, but it covers 10 cities rather than the full panel, so it bounds the concern rather than eliminating it.

Time standardization

All timestamps are handled in Coordinated Universal Time (UTC) end to end: OpenAQ values are taken from each record's UTC period start, the Open-Meteo archive is requested with `timezone=UTC`, and every series is parsed as UTC and made timezone-naive on a common UTC hourly grid before alignment. Calendar features (hour-of-day, day-of-week) are therefore computed on the UTC clock rather than each city's local time. Because the same UTC calendar is used identically for every tier and both domains, and because the covariate-timing contrast (main-text Fig. 3) compares a tier against itself under two covariate regimes, a fixed UTC offset cannot bias the covariate-timing message; it only means the learned diurnal phase is expressed in UTC. No daylight-saving adjustment is applied, since UTC has none.

E4 budget floors and normalization

Two features of the smallest fine-tune budgets are worth stating, and both cut against the transfer model's favour rather than for it.

First, the 0% condition. Per-city z-scoring statistics are computed from the target city's pre-test training history, so the 0% budget involves no target-city *gradient updates* but does use target-city data for input and output normalization. This advantages the transferred model at the smallest budgets. The direction is conservative here, because the transferred model still does not overtake the zero-shot foundation model.

Second, the two window floors are asymmetric. The fine-tune slice floors at 49 h (lookback + horizon + 1) while the LightGBM refit window floors at 202 h (weekly-lag warm-up + horizon + margin), so at the 1% budget the baseline receives up to about four times more data than the transfer model. This advantages the baseline, which still loses on mean MASE. Actual hours used are recorded per row and reported alongside nominal fractions in Table S10.

Equivalence-margin provenance

The equivalence *analysis* was added after peer review and is logged in the deviations record. The $\delta = 0.05$ MASE margin was fixed before the TOST was run and before its outcome was seen. It is not a pre-campaign registered quantity, and we do not describe it as one.

Because the verdict depends on the margin, Table S15 reports the full sensitivity. At a strict $\delta = 0.02$ no comparison is equivalent. At $\delta = 0.05$ only the perfect-foresight PM2.5 comparison is. At a lenient $\delta = 0.10$ the causal PM2.5 comparison and the 100%-budget E4 comparison also clear it, while causal temperature and the smaller E4 budgets never do. Main-text equivalence claims use $\delta = 0.05$ throughout.

Note that δ and α are distinct quantities that happen to share the value 0.05: δ is an equivalence margin measured in MASE units, whereas α is a significance level. The interval TOST tests against is the $1 - 2\alpha = 90\%$ confidence interval, not the 95% one, which is why every TOST interval reported in this paper is a 90% interval while the accompanying bootstrap CI on the paired difference is a conventional 95% one.

Supplementary tables

Table S2: City panel. The 29 cities passing the pre-specified quality gate, with country, data-tier (rich/scarce), OpenAQ reference sensor, usable window, and usable hours after gap-gating. Windows are city-specific and not contemporaneous; ten of them overlap the 2020–21 pandemic period, which is quantified above.

City	Country	Tier	OpenAQ sensor	Window start	Window end	Usable hours
Amsterdam	NL	rich	4235	2025-02-24	2026-07-09	11228
Bangkok	TH	rich	1304255	2023-03-19	2025-04-25	18164
Berlin	DE	rich	1300115	2024-03-23	2026-07-11	19574
Krakow	PL	rich	20854	2018-11-21	2019-12-31	9344
Los Angeles	US	rich	25551	2024-09-12	2025-08-22	8195
Madrid	ES	rich	10378	2025-04-24	2026-01-27	6431
Mexico City	MX	rich	2034	2017-01-30	2017-06-24	3371
New York	US	rich	673	2024-03-07	2025-09-20	13421
Paris	FR	rich	1582598	2023-09-01	2024-03-11	4457
Santiago	CL	rich	1044	2019-01-02	2020-12-11	16288
Seoul	KR	rich	8533919	2024-03-19	2026-07-13	19827
Sydney	AU	rich	21610	2019-12-23	2020-11-08	7277
Toronto	CA	rich	2914	2018-12-16	2020-07-31	13518
Vienna	AT	rich	30223	2025-10-19	2026-07-12	6241
Accra	GH	scarce	30469	2020-05-01	2021-04-24	8160
Addis Ababa	ET	scarce	30190	2016-12-23	2017-05-30	3610
Bogota	CO	scarce	29734	2020-02-10	2020-09-22	5175
Delhi	IN	scarce	35	2016-08-28	2016-12-22	2518
Dhaka	BD	scarce	24434	2023-04-26	2024-05-25	9395
Hanoi	VN	scarce	21632	2019-11-13	2020-07-31	6062
Jakarta	ID	scarce	25072	2019-09-24	2020-07-24	7024
Kampala	UG	scarce	25086	2018-12-16	2019-04-19	2836
Karachi	PK	scarce	23747	2019-11-19	2020-07-01	5214
Lagos	NG	scarce	2251459	2023-04-27	2023-09-15	3378
Lahore	PK	scarce	25135	2024-07-01	2025-02-18	5563
Lima	PE	scarce	4171	2022-03-24	2022-11-06	5227
Manila	PH	scarce	6909373	2024-11-03	2026-02-05	10992
Mumbai	IN	scarce	23320	2020-08-31	2021-01-26	3004
Nairobi	KE	scarce	16091678	2019-08-20	2021-03-10	13576

Table S3: Model configuration per tier. All settings mirror the released harness (`run_forecast.py`) and were fixed before the canonical campaign; none were tuned per city or per domain. The LightGBM `subsample_freq` correction predates the campaign and is logged in the analysis-plan deviations record.

Tier	Setting	Value
Seasonal-naïve	persistence lag	168 h (same hour last week)
LightGBM (specialist)	models	one direct model per lead time (24), retrained per fold
LightGBM (specialist)	n_estimators / learning_rate / num_leaves	250 / 0.05 / 31
LightGBM (specialist)	subsample / subsample_freq	0.8 / 1 (freq added after pre-campaign audit; bagging inert without it)
LightGBM (specialist)	random_state	42
LightGBM (specialist)	features	lagged target, calendar, meteorological covariates
NAS-GRU	architecture	Green-NAS-A: 2 stacked GRU layers, 128 units, direct multi-horizon head
NAS-GRU	optimizer / loss	Adam, learning rate 10^{-3}, MSE
NAS-GRU	early stopping	patience 10, max 50 epochs, validation fraction 0.1
NAS-GRU	lookback / batch size	24 h / 256
NAS-GRU	seeds	{42, 43, 44, 45, 46}; per-city seed-mean before any test
Chronos-Bolt (zero-shot)	checkpoint	amazon/chronos-bolt-small (47,718,016 parameters), no training or fine-tuning
Chronos-Bolt (zero-shot)	context / point forecast	672 h (four weeks) / predicted mean
Chronos-Bolt + covariates	covariate model	ridge regression ($\alpha = 1.0$) on standardized calendar + weather features
Chronos-Bolt + covariates	scheme	FM forecasts the residual of the ridge fit; ridge horizon prediction added back

Table S4: Split-conformal 95% interval calibration, pooled per data tier. Empirical coverage and mean interval width (target units) pooled across cities within each tier, under the deployable causal-covariate configuration and the perfect-foresight upper bound. Covariate-free tiers (seasonal-naïve, NAS-GRU, zero-shot Chronos-Bolt) are identical in both configurations by construction. **(a)** PM2.5. **(b)** Temperature. Per-city calibration in Tables S6 and S7.

(a) PM2.5

Data tier	Model	Coverage (causal)	Width (causal)	Coverage (perfect)	Width (perfect)
rich	Chronos-Bolt (zero-shot)	0.929	18.47	0.929	18.47
rich	Chronos-Bolt + covariates	0.910	17.80	0.936	19.48
rich	LightGBM (specialist)	0.935	21.05	0.926	17.69
rich	NAS-GRU	0.914	18.78	0.914	18.78
rich	Seasonal-naïve	0.970	35.07	0.970	35.07
scarce	Chronos-Bolt (zero-shot)	0.931	108.64	0.931	108.64
scarce	Chronos-Bolt + covariates	0.928	123.94	0.931	112.69
scarce	LightGBM (specialist)	0.913	105.83	0.916	102.17
scarce	NAS-GRU	0.931	107.96	0.931	107.96
scarce	Seasonal-naïve	0.949	163.85	0.949	163.85

(b) Temperature

Data tier	Model	Coverage (causal)	Width (causal)	Coverage (perfect)	Width (perfect)
rich	Chronos-Bolt (zero-shot)	0.900	10.72	0.900	10.72
rich	Chronos-Bolt + covariates	0.923	23.91	0.892	9.17
rich	LightGBM (specialist)	0.923	9.56	0.958	7.70
rich	NAS-GRU	0.910	11.75	0.910	11.75
rich	Seasonal-naïve	0.934	18.04	0.934	18.04
scarce	Chronos-Bolt (zero-shot)	0.936	5.90	0.936	5.90
scarce	Chronos-Bolt + covariates	0.939	13.79	0.943	3.79
scarce	LightGBM (specialist)	0.944	4.38	0.949	2.68
scarce	NAS-GRU	0.950	6.02	0.950	6.02
scarce	Seasonal-naïve	0.968	8.96	0.968	8.96

Table S5: Full Diebold–Mariano pairwise win matrices (perfect-foresight panel). Per-pair significant-win counts (raw and FDR-adjusted) across the 29 cities. **(a)** PM2.5. **(b)** Temperature.

(a) PM2.5

Model pair	Cities	Sig. (raw)	A wins (raw)	B wins (raw)	A wins (FDR)	B wins (FDR)	Median *P*
Chronos-Bolt (zero-shot) vs Chronos-Bolt + covariates	29	5	4	1	2	1	0.295
Chronos-Bolt (zero-shot) vs NAS-GRU	29	6	5	1	2	0	0.308
Chronos-Bolt + covariates vs NAS-GRU	29	6	3	3	2	1	0.270
LightGBM (specialist) vs Chronos-Bolt (zero-shot)	29	3	2	1	0	0	0.429
LightGBM (specialist) vs Chronos-Bolt + covariates	29	3	3	0	2	0	0.486
LightGBM (specialist) vs NAS-GRU	29	7	6	1	6	1	0.384
Seasonal-naïve vs Chronos-Bolt (zero-shot)	29	18	0	18	0	13	0.029
Seasonal-naïve vs Chronos-Bolt + covariates	29	12	0	12	0	9	0.080
Seasonal-naïve vs LightGBM (specialist)	29	13	0	13	0	8	0.056
Seasonal-naïve vs NAS-GRU	29	15	0	15	0	11	0.047

(b) Temperature

Model pair	Cities	Sig. (raw)	A wins (raw)	B wins (raw)	A wins (FDR)	B wins (FDR)	Median *P*
Chronos-Bolt (zero-shot) vs Chronos-Bolt + covariates	29	5	2	3	2	1	0.370
Chronos-Bolt (zero-shot) vs NAS-GRU	29	9	8	1	6	0	0.107
Chronos-Bolt + covariates vs NAS-GRU	29	14	12	2	12	2	0.051
LightGBM (specialist) vs Chronos-Bolt (zero-shot)	29	11	11	0	6	0	0.167
LightGBM (specialist) vs Chronos-Bolt + covariates	29	9	9	0	6	0	0.153
LightGBM (specialist) vs NAS-GRU	29	20	20	0	20	0	0.003
Seasonal-naïve vs Chronos-Bolt (zero-shot)	29	20	0	20	0	16	0.017
Seasonal-naïve vs Chronos-Bolt + covariates	29	21	0	21	0	20	0.010
Seasonal-naïve vs LightGBM (specialist)	29	26	0	26	0	26	0.001
Seasonal-naïve vs NAS-GRU	29	16	2	14	2	13	0.022

Table S6: Per-city split-conformal calibration, PM2.5 (causal-covariate configuration). Empirical coverage (nominal 0.95) and mean interval width (μg m^{-3}) per city and tier.

City	Tier	Chronos cov.	Chronos width	Chronos+cov cov.	Chronos+cov width	LightGBM cov.	LightGBM width	NAS-GRU cov.	NAS-GRU width	Naïve cov.	Naïve width
Accra	scarce	0.86	24.5	0.89	32.8	0.86	25.4	0.90	29.2	0.93	52.0
Addis Ababa	scarce	0.92	57.7	0.93	68.5	0.81	45.7	0.89	59.2	0.94	84.0
Amsterdam	rich	0.82	5.3	0.92	6.8	0.92	7.9	1.00	12.4	1.00	34.8
Bangkok	rich	0.86	4.6	1.00	18.6	0.97	5.9	0.90	11.1	0.72	17.0
Berlin	rich	1.00	10.4	0.93	8.3	0.93	7.4	0.82	14.6	0.99	14.8
Bogota	scarce	1.00	47.2	0.96	41.6	1.00	43.0	1.00	41.4	1.00	54.0
Delhi	scarce	0.92	339.2	0.92	427.6	0.92	354.2	0.96	454.9	0.89	462.0
Dhaka	scarce	0.99	327.5	0.97	306.1	0.97	284.9	0.97	298.4	0.99	422.0
Hanoi	scarce	0.97	100.5	0.99	81.6	1.00	99.8	0.94	85.2	0.92	116.0
Jakarta	scarce	1.00	85.1	1.00	86.5	0.93	55.8	0.97	55.8	0.90	74.0
Kampala	scarce	0.97	125.1	0.94	106.6	0.90	85.2	0.88	86.4	0.90	158.0
Karachi	scarce	0.89	28.6	0.94	36.0	0.81	26.5	0.93	32.1	0.82	34.0
Krakow	rich	0.60	40.3	0.60	36.5	0.82	37.2	0.89	42.6	0.94	113.2
Lagos	scarce	0.79	22.4	0.83	31.1	0.79	20.7	0.79	20.0	0.97	42.2
Lahore	scarce	0.90	216.5	0.78	346.6	0.94	267.9	0.86	204.1	1.00	576.0
Lima	scarce	0.85	34.7	0.90	37.1	0.86	36.0	0.96	46.0	1.00	70.8
Los Angeles	rich	1.00	23.3	1.00	26.1	1.00	40.4	1.00	24.0	1.00	28.5
Madrid	rich	1.00	25.1	1.00	16.1	0.99	11.2	0.97	10.6	1.00	28.0
Manila	scarce	0.92	47.5	0.89	48.6	0.94	64.0	0.92	59.1	0.97	100.8
Mexico City	rich	0.86	27.7	0.81	24.9	1.00	61.7	1.00	44.9	1.00	64.0
Mumbai	scarce	1.00	143.2	1.00	187.0	0.96	142.3	1.00	116.2	1.00	160.0
Nairobi	scarce	1.00	29.9	0.97	21.5	1.00	36.2	0.99	31.3	1.00	52.0
New York	rich	0.96	12.3	0.99	12.4	0.89	11.9	0.81	11.5	1.00	21.4
Paris	rich	1.00	38.9	0.69	28.0	0.96	37.4	0.88	24.9	1.00	47.2
Santiago	rich	1.00	26.4	0.94	27.2	1.00	28.2	1.00	27.1	1.00	30.0
Seoul	rich	0.96	14.8	0.92	15.3	0.97	18.3	0.83	14.4	1.00	38.0
Sydney	rich	0.97	9.8	1.00	9.7	0.81	12.6	0.96	10.6	0.93	13.8
Toronto	rich	1.00	15.2	0.94	12.8	0.86	9.6	0.81	9.6	1.00	18.0
Vienna	rich	0.97	4.5	1.00	6.5	0.97	5.1	0.93	4.4	1.00	22.4

Table S7: Per-city split-conformal calibration, temperature (causal-covariate configuration). Empirical coverage (nominal 0.95) and mean interval width (°C) per city and tier.

City	Tier	Chronos cov.	Chronos width	Chronos+cov cov.	Chronos+cov width	LightGBM cov.	LightGBM width	NAS-GRU cov.	NAS-GRU width	Naïve cov.	Naïve width
Accra	scarce	0.88	1.8	0.92	10.0	0.88	2.3	0.82	2.7	1.00	4.6
Addis Ababa	scarce	1.00	13.2	1.00	21.2	0.94	4.4	0.96	6.2	0.97	17.2
Amsterdam	rich	1.00	13.7	1.00	30.1	1.00	15.0	1.00	14.0	1.00	16.0
Bangkok	rich	0.99	4.9	0.99	12.9	0.99	5.2	0.99	4.8	0.97	12.8
Berlin	rich	0.78	11.1	0.97	32.3	0.81	13.7	0.89	17.1	0.94	27.0
Bogota	scarce	0.78	3.6	0.81	18.2	0.86	3.5	0.86	5.7	0.89	5.6
Delhi	scarce	0.81	9.7	0.94	18.0	0.86	6.0	0.97	14.8	0.96	27.6
Dhaka	scarce	1.00	10.0	0.92	14.1	1.00	7.0	1.00	7.6	0.99	8.8
Hanoi	scarce	1.00	5.8	1.00	7.9	1.00	7.8	1.00	7.8	0.97	4.8
Jakarta	scarce	1.00	3.0	0.93	13.2	0.99	2.9	1.00	4.0	0.90	2.8
Kampala	scarce	1.00	16.2	1.00	16.4	1.00	7.0	1.00	11.8	1.00	20.8
Karachi	scarce	0.83	1.3	0.97	12.0	0.85	1.6	0.96	3.3	1.00	2.4
Krakow	rich	0.79	4.3	0.78	5.8	0.82	5.3	0.54	3.3	1.00	25.8
Lagos	scarce	0.82	2.1	0.83	5.7	0.83	2.6	0.89	2.7	1.00	6.0
Lahore	scarce	1.00	7.3	0.97	17.5	1.00	7.3	1.00	8.3	1.00	8.4
Lima	scarce	0.99	2.1	0.96	9.7	1.00	2.0	0.90	2.1	1.00	7.6
Los Angeles	rich	1.00	33.2	0.94	29.9	0.97	10.1	1.00	17.7	0.88	26.6
Madrid	rich	0.86	7.3	0.74	12.8	0.79	8.2	0.99	7.9	0.79	10.2
Manila	scarce	1.00	3.5	0.94	11.0	1.00	3.2	0.99	3.6	0.94	4.6
Mexico City	rich	0.58	4.1	0.85	20.5	0.72	4.4	0.58	3.5	0.79	13.2
Mumbai	scarce	0.94	4.7	0.89	12.5	0.97	4.4	0.92	5.9	0.89	7.0
Nairobi	scarce	1.00	4.1	1.00	19.6	0.99	3.8	0.99	3.7	1.00	6.2
New York	rich	0.96	8.4	0.96	31.9	0.94	6.9	0.93	5.1	1.00	18.2
Paris	rich	1.00	9.0	1.00	23.2	1.00	13.6	1.00	15.3	1.00	15.0
Santiago	rich	0.92	12.9	0.92	29.0	0.93	8.7	1.00	16.1	1.00	23.6
Seoul	rich	1.00	10.0	1.00	29.1	1.00	10.9	1.00	10.5	1.00	19.8
Sydney	rich	1.00	14.4	0.88	11.5	1.00	12.9	1.00	20.1	1.00	18.2
Toronto	rich	0.78	10.5	1.00	42.2	1.00	11.1	0.93	13.8	0.96	16.6
Vienna	rich	0.94	6.4	0.90	23.8	0.94	8.0	0.89	15.1	0.74	9.6

Table S8: Measured-energy repeatability. codecarbon-measured joules per 1,000 forecasts, five repetitions on three cities per tier, with the coefficient of variation and the 20% gate flag; flagged cells are reported as ranges in Table 3.

City	Tier	Reps	Mean (J/1k)	SD (J/1k)	SD/mean	>20% gate
Beijing	Chronos-Bolt (zero-shot)	5	1,229	240.9	0.196	False
Beijing	Chronos-Bolt + covariates	5	763	102.3	0.134	False
Beijing	LightGBM (specialist)	5	15,061	2,373.7	0.158	False
Beijing	NAS-GRU	5	13,237	2,690.1	0.203	True
Beijing	Seasonal-naïve	5	2	0.4	0.200	False
Nairobi	Chronos-Bolt (zero-shot)	5	1,946	1,940.3	0.997	True
Nairobi	Chronos-Bolt + covariates	5	842	108.9	0.129	False
Nairobi	LightGBM (specialist)	5	9,461	835.8	0.088	False
Nairobi	NAS-GRU	5	7,298	104.6	0.014	False
Nairobi	Seasonal-naïve	5	16	19.3	1.203	True
Seoul	Chronos-Bolt (zero-shot)	5	1,025	141.3	0.138	False
Seoul	Chronos-Bolt + covariates	5	781	94.2	0.120	False
Seoul	LightGBM (specialist)	5	10,832	406.1	0.037	False
Seoul	NAS-GRU	5	7,496	440.4	0.059	False
Seoul	Seasonal-naïve	5	9	15.2	1.684	True

Table S9: Horizon-48 panel accuracy. MASE (mean ± sd across cities) per tier at the 48-hour secondary horizon, perfect-foresight covariates, split by data tier (rich, $n = 14$ cities; scarce, $n = 15$). **(a)** PM2.5. **(b)** Temperature. The specialist's temperature lead over zero-shot Chronos-Bolt widens with horizon exactly as the perfect-foresight-artifact hypothesis predicts: from 0.259 MASE pooled at $h = 24$ to 0.370 (rich) and 0.425 (scarce) at $h = 48$, since a perfect weather forecast is worth more the further ahead the model must look.

(a) PM2.5

Tier	rich	scarce
Chronos-Bolt (zero-shot)	0.547 ± 0.208	0.892 ± 0.385
Chronos-Bolt + covariates	0.642 ± 0.245	0.902 ± 0.368
LightGBM (specialist)	0.535 ± 0.173	0.885 ± 0.407
NAS-GRU	0.639 ± 0.262	0.988 ± 0.435
Seasonal-naïve	0.843 ± 0.432	1.242 ± 0.545

(b) Temperature

Tier	rich	scarce
Chronos-Bolt (zero-shot)	1.094 ± 0.315	0.958 ± 0.398
Chronos-Bolt + covariates	0.982 ± 0.351	0.710 ± 0.260
LightGBM (specialist)	0.724 ± 0.316	0.533 ± 0.278
NAS-GRU	1.345 ± 0.551	1.238 ± 0.649
Seasonal-naïve	1.926 ± 0.749	1.736 ± 0.972

Table S10: E4 transfer-vs-zero-shot full grid. Per-strategy mean MASE (± sd) at each nominal fine-tune fraction, with the mean actual training hours used (nominal fractions floor differently per strategy, so actual hours are reported alongside).

Strategy	Nominal fraction (%)	MASE (mean ± sd)	Actual training hours (mean)	Cities
Chronos-Bolt (zero-shot)	—	0.843 ± 0.354	0	15
LightGBM (refit on budget)	1	0.941 ± 0.311	202	15
LightGBM (refit on budget)	10	0.944 ± 0.383	599	15
LightGBM (refit on budget)	100	0.858 ± 0.382	5,995	15
NAS-GRU (transfer + fine-tune)	0	0.899 ± 0.340	0	15
NAS-GRU (transfer + fine-tune)	1	0.915 ± 0.331	66	15
NAS-GRU (transfer + fine-tune)	10	0.888 ± 0.340	599	15
NAS-GRU (transfer + fine-tune)	100	0.876 ± 0.374	5,995	15

Table S11: Cost-assumption sensitivity of the decision maps (causal configuration, matching main-text Fig. 5). Winner-cell flip rate for each regime run under an electricity-price × PUE grid (price and PUE act as linear rescalings of the cost-penalty coefficient λ; the central assumption 0.15 USD/kWh × PUE 1.4 has multiplier 1 by definition). No cells flip at the central assumption; at most 40% flip at the most extreme corners of the grid.

Regime run	Price ($/kWh)	PUE	Effective λ multiplier	Cells	Flipped	Flip rate
Beijing PM2.5	0.05	1.0	0.238	25	4	0.16
Beijing PM2.5	0.05	1.4	0.333	25	4	0.16
Beijing PM2.5	0.05	2.0	0.476	25	3	0.12
Beijing PM2.5	0.15	1.0	0.714	25	3	0.12
Beijing PM2.5	0.15	1.4	1.000	25	0	0.00
Beijing PM2.5	0.15	2.0	1.429	25	0	0.00
Beijing PM2.5	0.30	1.0	1.429	25	0	0.00
Beijing PM2.5	0.30	1.4	2.000	25	2	0.08
Beijing PM2.5	0.30	2.0	2.857	25	3	0.12
Beijing Temperature	0.05	1.0	0.238	25	4	0.16
Beijing Temperature	0.05	1.4	0.333	25	3	0.12
Beijing Temperature	0.05	2.0	0.476	25	2	0.08
Beijing Temperature	0.15	1.0	0.714	25	1	0.04
Beijing Temperature	0.15	1.4	1.000	25	0	0.00
Beijing Temperature	0.15	2.0	1.429	25	1	0.04
Beijing Temperature	0.30	1.0	1.429	25	1	0.04
Beijing Temperature	0.30	1.4	2.000	25	2	0.08
Beijing Temperature	0.30	2.0	2.857	25	3	0.12
Nairobi PM2.5	0.05	1.0	0.238	20	7	0.35
Nairobi PM2.5	0.05	1.4	0.333	20	5	0.25
Nairobi PM2.5	0.05	2.0	0.476	20	4	0.20
Nairobi PM2.5	0.15	1.0	0.714	20	3	0.15
Nairobi PM2.5	0.15	1.4	1.000	20	0	0.00
Nairobi PM2.5	0.15	2.0	1.429	20	2	0.10
Nairobi PM2.5	0.30	1.0	1.429	20	2	0.10
Nairobi PM2.5	0.30	1.4	2.000	20	8	0.40
Nairobi PM2.5	0.30	2.0	2.857	20	8	0.40
Nairobi Temperature	0.05	1.0	0.238	20	4	0.20
Nairobi Temperature	0.05	1.4	0.333	20	4	0.20
Nairobi Temperature	0.05	2.0	0.476	20	2	0.10
Nairobi Temperature	0.15	1.0	0.714	20	0	0.00
Nairobi Temperature	0.15	1.4	1.000	20	0	0.00
Nairobi Temperature	0.15	2.0	1.429	20	4	0.20
Nairobi Temperature	0.30	1.0	1.429	20	4	0.20
Nairobi Temperature	0.30	1.4	2.000	20	7	0.35
Nairobi Temperature	0.30	2.0	2.857	20	8	0.40
Seoul PM2.5	0.05	1.0	0.238	25	5	0.20
Seoul PM2.5	0.05	1.4	0.333	25	4	0.16
Seoul PM2.5	0.05	2.0	0.476	25	4	0.16
Seoul PM2.5	0.15	1.0	0.714	25	0	0.00
Seoul PM2.5	0.15	1.4	1.000	25	0	0.00
Seoul PM2.5	0.15	2.0	1.429	25	1	0.04
Seoul PM2.5	0.30	1.0	1.429	25	1	0.04
Seoul PM2.5	0.30	1.4	2.000	25	9	0.36
Seoul PM2.5	0.30	2.0	2.857	25	10	0.40
Seoul Temperature	0.05	1.0	0.238	25	0	0.00
Seoul Temperature	0.05	1.4	0.333	25	0	0.00
Seoul Temperature	0.05	2.0	0.476	25	0	0.00
Seoul Temperature	0.15	1.0	0.714	25	0	0.00
Seoul Temperature	0.15	1.4	1.000	25	0	0.00
Seoul Temperature	0.15	2.0	1.429	25	0	0.00
Seoul Temperature	0.30	1.0	1.429	25	0	0.00

continued on next page

Table S11 continued

Regime run	Price ($/kWh)	PUE	Effective λ multiplier	Cells	Flipped	Flip rate
Seoul Temperature	0.30	1.4	2.000	25	0	0.00
Seoul Temperature	0.30	2.0	2.857	25	0	0.00

Table S12: Energy train/inference decomposition and amortization crossover. One-time training energy per fit versus per-forecast inference energy for the specialist and the zero-shot foundation model, on CPU and GPU, and the crossover count (forecasts a once-trained specialist serves before its amortized energy per forecast falls to the foundation model's inference energy). codecarbon measurements at sub-kilojoule scale are noisy (the GPU column is less stable; one negative inference estimate is a measurement artifact), so the CPU rows are the reliable signal.

City	Device	Train energy / fit (J)	LightGBM infer (J/forecast)	Chronos infer (J/forecast)	Crossover (forecasts)
Beijing	CPU	317	0.033	0.112	4,041
Beijing	GPU	550	-0.243	1.942	252
Nairobi	CPU	203	0.065	0.111	4,416
Nairobi	GPU	348	0.075	0.322	1,411
Seoul	CPU	246	0.022	0.110	2,778
Seoul	GPU	423	0.130	0.327	2,147

Table S13: Perfect-foresight panel accuracy (upper bound). Table 1 recomputed with perfect-foresight covariates for every covariate-using tier, giving an upper bound on the specialist's reach that assumes a perfect forecast of its own inputs. The specialist reaches 0.533 on temperature with 6 FDR-significant wins here; under deployable causal covariates (Table 1) it has none.

Tier	Domain	MASE (mean ± sd)	MAE (mean ± sd)	RMSE (mean ± sd)	Avg rank	FDR wins (tier/FM)
Chronos-Bolt (zero-shot)	PM2.5	0.662 ± 0.368	10.553 ± 13.524	15.093 ± 19.273	1.83	—
Chronos-Bolt + covariates	PM2.5	0.730 ± 0.376	11.745 ± 15.266	15.959 ± 20.478	3.14	1 / 2
LightGBM (specialist)	PM2.5	0.662 ± 0.371	10.845 ± 14.060	14.810 ± 18.895	2.17	0 / 0
NAS-GRU	PM2.5	0.734 ± 0.352	11.426 ± 14.931	15.592 ± 20.791	3.07	0 / 2
Seasonal-naïve	PM2.5	1.026 ± 0.530	16.505 ± 21.235	22.274 ± 28.983	4.79	0 / 13
Chronos-Bolt (zero-shot)	Temp.	0.792 ± 0.282	1.303 ± 0.717	1.819 ± 1.040	2.90	—
Chronos-Bolt + covariates	Temp.	0.718 ± 0.278	1.228 ± 0.795	1.559 ± 1.012	2.41	1 / 2
LightGBM (specialist)	Temp.	0.533 ± 0.208	0.888 ± 0.550	1.153 ± 0.730	1.28	6 / 0
NAS-GRU	Temp.	0.999 ± 0.481	1.626 ± 1.093	2.070 ± 1.261	3.59	0 / 6
Seasonal-naïve	Temp.	1.686 ± 1.050	2.716 ± 1.780	3.300 ± 2.085	4.83	0 / 16

Table S14: Post-cutoff (unseen-data) check. Zero-shot Chronos-Bolt versus the LightGBM specialist evaluated only on strictly post-2024-10 PM2.5 windows, which postdate the model's pretraining corpus, for the 10 cities with sufficient recent coverage. The foundation model remains competitive (mean MASE 0.415 vs 0.395; better in 6 of 10 cities), so the PM2.5 result is not a pretraining-exposure artifact. These windows also lie outside the 2020–21 pandemic period, so the same comparison doubles as a check that the result does not depend on pandemic-era cities.

City	Post-cutoff hours	Chronos MASE	LightGBM MASE	Naive MASE
Amsterdam	11,228	0.243	0.249	0.814
Bangkok	4,823	0.133	0.196	0.428
Berlin	15,276	0.238	0.255	0.493
Lahore	3,369	0.468	0.512	0.734
Los Angeles	7,757	0.532	0.521	0.710
Madrid	6,431	0.690	0.346	0.755
Manila	10,992	0.935	0.911	1.655
New York	8,499	0.425	0.470	0.631
Seoul	15,237	0.321	0.339	0.742
Vienna	6,241	0.167	0.151	0.528

Table S15: Sensitivity of the TOST equivalence verdict to the margin. For each of the seven pre-specified tie comparisons: the paired point difference (positive favours the zero-shot foundation model), the TOST 90% confidence interval, the smallest equivalence margin at which the interval fits inside $\pm\delta$ ("min. margin"), and the verdict (✓ = equivalent) at a strict ($\delta = 0.02$), the main-text ($\delta = 0.05$), and a lenient ($\delta = 0.10$) MASE margin. Here δ is a margin in MASE units, distinct from the significance level $\alpha = 0.05$; the reported interval is 90% because TOST tests against the $1 - 2\alpha$ interval. The main-text $\delta = 0.05$ was fixed before the test was run but is not a registered pre-campaign quantity (Supplementary methods); the table makes the margin dependence explicit. At $\delta = 0.05$ only the perfect-foresight PM2.5 comparison is equivalent; a lenient 0.10 margin additionally admits causal PM2.5 and the 100%-budget E4 comparison.

Comparison	Point diff	TOST 90% CI	Min. margin	δ=0.02	δ=0.05	δ=0.10
PM2.5, perfect-foresight (specialist vs FM)	-0.000	[-0.042, +0.042]	0.042	–	✓	✓
PM2.5, causal (specialist vs FM)	+0.030	[-0.012, +0.071]	0.071	–	–	✓
Temperature, causal (specialist vs FM)	-0.048	[-0.114, +0.018]	0.114	–	–	–
E4 transfer 0% vs zero-shot	+0.056	[+0.005, +0.107]	0.107	–	–	–
E4 transfer 1% vs zero-shot	+0.072	[+0.015, +0.129]	0.129	–	–	–
E4 transfer 10% vs zero-shot	+0.045	[-0.012, +0.102]	0.102	–	–	–
E4 transfer 100% vs zero-shot	+0.033	[-0.009, +0.075]	0.075	–	–	✓

Applying the framework to a new city

The decision rule is not specific to the 29 cities studied here. Any operator can re-derive a winner map for their own city by inserting local electricity prices, hardware, and retraining frequency. Table S16 sets out the procedure in the order the released code runs it.

Table S16: Replication and deployment recipe. The steps needed to reproduce this evaluation for a new city and read a deployment decision from it. Steps 1–5 reproduce the accuracy comparison; steps 6–9 turn it into a cost-aware choice.

Step	Action	What to do
1	Gate the series	Drop physically invalid readings, split at gaps $>$ 48 h, interpolate only gaps $\leq$ 6 h, and keep a window of $\geq$ 2,160 h at $\geq$ 60% real coverage. A city that fails this has too little history to compare strategies at all.
2	Fix the MASE scale	Compute the in-sample seasonal-naive (m = 24) error on the pre-test history once per series and reuse it for every tier and every training-history regime, so scores stay comparable as history varies.
3	Build causal covariates	Enter each meteorological covariate at its last value known at the forecast origin; leave calendar features future-known. Use perfect-foresight covariates only as a labelled ceiling, never as the headline.
4	Assemble the candidates	At minimum: the zero-shot foundation model, one locally trained specialist, and the seasonal-naive floor. The foundation model needs no training; the specialist needs the covariate frame from step 3.
5	Backtest	Rolling-origin, six folds, all evaluated on the same final test window per series, so different training-history regimes meet on identical test data.
6	Measure energy	Wrap each tier's entire run and record joules per 1,000 forecasts. Repeat the workload five times and report any cell with sd/mean $>$ 20% as a range. Record training and inference separately.
7	Set λ	Convert energy to cost with the local electricity price and PUE, then choose λ, the MASE units worth one USD per 1,000 forecasts. Price and PUE rescale λ linearly, so a sweep covers a range of local conditions.
8	Read the winner	Select the tier minimizing MASE + $\lambda \cdot$ USD/1k in the cell matching the available training history. Check the neighbouring cells: a winner that flips under a small change in λ is not a robust choice.
9	Check the retraining cadence	Compare the planned refresh interval against the amortization crossover. Refresh more often than the crossover and the zero-shot model is cheaper; train once and serve for longer and the specialist wins.

Supplementary figures

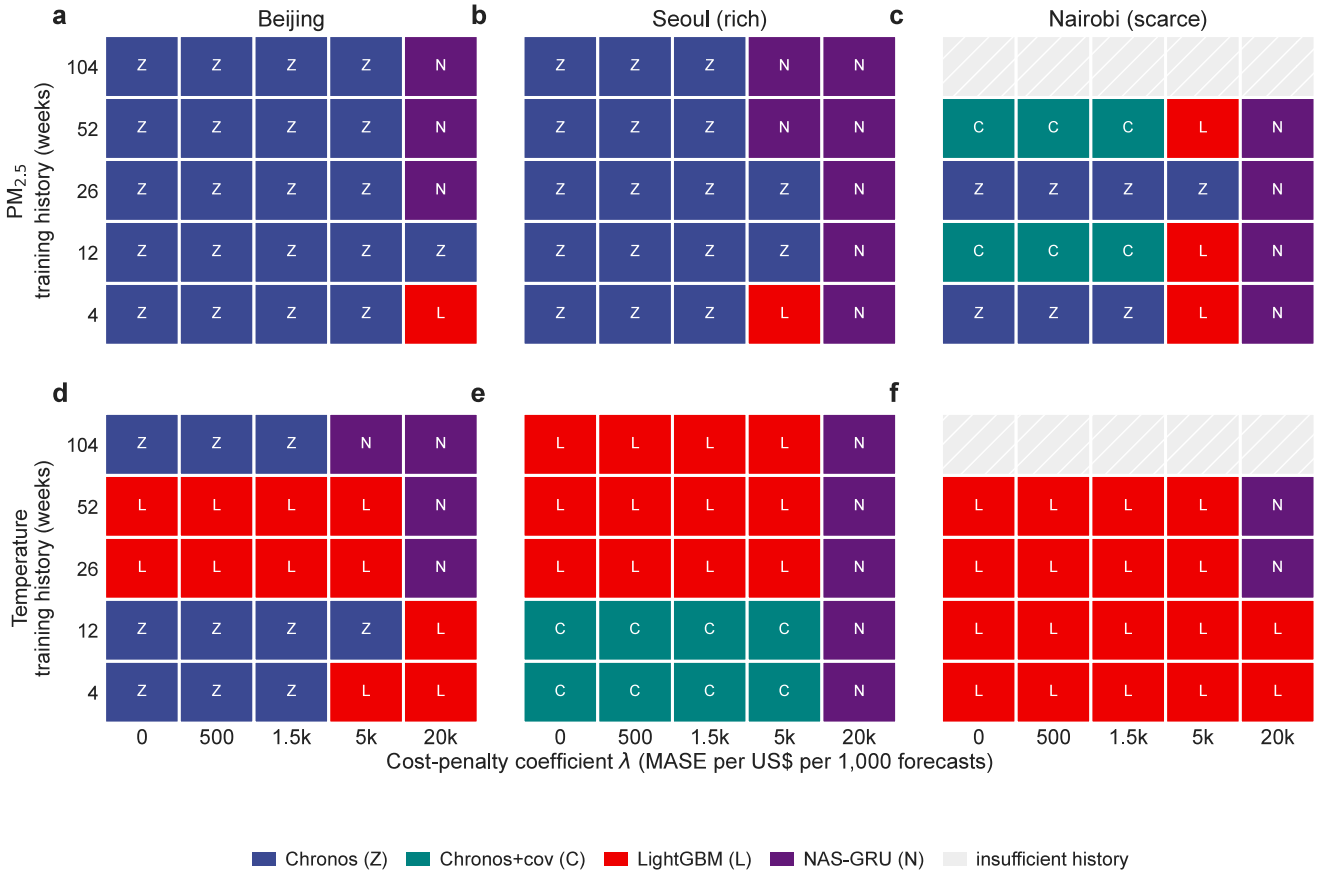

Figure S1: Cost-adjusted deployment winner maps, perfect-foresight covariates (upper bound). As Figure 5 of the main text but with perfect-foresight covariates for every covariate-using tier (panels **a**–**f** as in Fig. 5; cell letters duplicate the colour coding for grayscale readers). The temperature maps favour the specialist here (Beijing 11/25, Seoul 12/25, Nairobi 18/20 cells); under the deployable causal covariates of the main Figure 5 that dominance collapses.

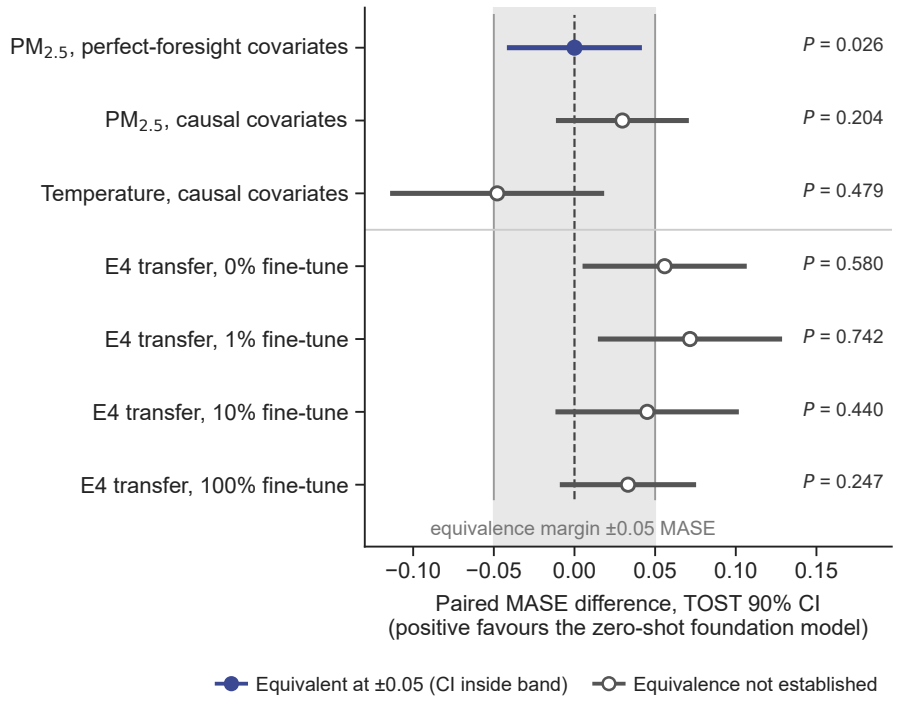

Figure S2: TOST equivalence tests for all seven pre-specified tie comparisons. Paired MASE difference (positive favours the zero-shot foundation model) with the TOST 90% confidence interval against the pre-specified equivalence margin $\delta = 0.05$ MASE (shaded band); P values are the paired TOST P at that margin. The interval is 90% rather than 95% because TOST tests against the $1 - 2\alpha$ interval at $\alpha = 0.05$. Equivalence is established only where the whole interval lies inside the band (filled marker): the perfect-foresight PM2.5 specialist-versus-foundation-model comparison ($n = 29$ cities; E4 rows $n = 15$ scarce cities). All other comparisons (causal PM2.5, causal temperature, and every E4 fine-tune budget) remain "no significant difference" verdicts, not demonstrated equivalence.