STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, July 2026, pp 732–748.
Published online in International Academic Press (www.IAPress.org)

Multivariate Hourly Air Quality Forecasting with MES-LSTM as the Core Framework

Lilis Anggraini [1,2*], Edi Noersasongko [1], Aris Marjuni [1], Purwanto [1]

[1] *Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia*
[2] *Faculty of Information Technology, Islamic University of Kalimantan Muhammad Arsyad Al Banjari, Banjarmasin, Indonesia*

Abstract Accurate air quality forecasting plays an important role in supporting environmental management, protecting public health, and enabling timely early-warning interventions. However, many existing forecasting approaches rely on univariate modeling and may not adequately capture interactions among multiple pollutants. This study therefore explores multivariate hourly air quality forecasting using Multivariate Exponential Smoothing–Long Short-Term Memory (MES-LSTM) as the primary forecasting framework. Experiments were conducted using the Beijing Air Quality dataset to predict carbon monoxide (CO), nitric oxide (NO), nitrogen dioxide (NO_2), ozone (O_3), and sulfur dioxide (SO_2). A standardized evaluation protocol was adopted for all models, including identical preprocessing procedures, chronological data partitioning, a 24-hour lookback window, and a one-hour forecasting horizon. The proposed MES-LSTM model was evaluated against Long Short-Term Memory (LSTM), Temporal Fusion Transformer (TFT), and a hybrid MES-LSTM–TFT architecture. The results show that MES-LSTM achieved the lowest forecasting errors for several pollutants, particularly CO, NO_2, and O_3, while remaining competitive for NO and SO_2. In contrast, the hybrid MES-LSTM–TFT model did not improve forecasting performance and generally produced larger prediction errors than the corresponding single-stage models. Additional robustness analyses, including five-seed experiments, persistence baselines, and multivariate coupling ablations, indicate that the effectiveness of pollutant interaction modeling depends on the target pollutant being predicted. Overall, the findings suggest that combining exponential smoothing with residual learning provides an effective and robust solution for short-term multivariate air quality forecasting. The study further demonstrates that MES-LSTM provides a strong and competitive forecasting framework under a unified evaluation protocol, while increased architectural complexity does not necessarily lead to improved predictive performance. Additional robustness analyses, including five-seed experiments, persistence baselines, and coupling ablations, were conducted to assess stability and reproducibility. Computational cost was not formally benchmarked and remains an avenue for future investigation.

Keywords air quality forecasting, multivariate time series, MES-LSTM, exponential smoothing, LSTM baseline, temporal fusion transformer baseline, hybrid modeling, Beijing air quality

AMS 2010 subject classifications 62M10, 68T07

DOI: 10.19139/soic-2310-5070-3766

1. Introduction

Air pollution remains a major global public health challenge, with substantial evidence linking pollutant exposure to respiratory and cardiovascular diseases [41, 19, 21, 36, 7]. Accurate hourly forecasting is particularly important because intervention opportunities are often limited and environmental management decisions frequently depend on near-real-time information. Consequently, forecasting systems must deliver not only accurate predictions but also stable performance under operational conditions.

*Correspondence to: Lilis Anggraini (Email: p41202300075@mhs.dinus.ac.id). Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Our previous study investigated univariate NO_2 forecasting in Beijing using ARIMA, LSTM, and XGBoost models [2]. The findings revealed notable differences in predictive performance, with XGBoost achieving the strongest results among the evaluated approaches. At the same time, the study highlighted an important limitation of univariate forecasting. Treating each pollutant as an independent process overlooks the fact that urban air quality is shaped by interconnected emission sources, atmospheric chemistry, meteorological influences, and shared temporal patterns [48, 20, 34, 33, 37]. As a result, improvements achieved within a single-pollutant setting do not necessarily translate into a broader understanding of air-quality dynamics. This limitation motivates the present transition from univariate to multivariate forecasting.

Interest in multivariate forecasting has grown rapidly in recent years because pollutant concentrations rarely evolve independently. Instead, they are influenced by complex interactions among multiple pollutants and environmental processes. Despite this progress, the practical benefits of incorporating inter-pollutant information remain difficult to assess. Reported improvements often arise under different preprocessing strategies, forecasting horizons, evaluation metrics, and experimental protocols, making direct comparisons challenging. Consequently, there remains a need for controlled experimental settings that can distinguish genuine architectural improvements from gains introduced by differences in evaluation design [26, 42, 47, 24, 18, 46, 22, 8].

The shift from univariate to multivariate forecasting also changes the nature of the problem being addressed. Rather than modeling a single pollutant trajectory in isolation, the objective becomes understanding coupled temporal dynamics characterized by nonlinear dependencies, heterogeneous noise, and cross-pollutant interactions [26, 42, 47, 24]. Recent advances in recurrent neural networks and attention-based architectures have demonstrated considerable potential for modeling such behavior. However, existing evidence remains inconclusive regarding whether increasingly sophisticated architectures consistently deliver superior forecasting performance in real-world air-quality applications [46, 22, 8].

Within this context, this study focuses on Multivariate Exponential Smoothing–Long Short-Term Memory (MES-LSTM) as the primary forecasting framework. MES-LSTM combines exponential smoothing for extracting stable level components with LSTM-based residual learning for capturing nonlinear short-term variations. The underlying intuition is straightforward: smoothing can represent lower-frequency temporal structures, while the neural component concentrates on residual patterns that remain after decomposition [23, 16, 35, 17, 29]. Such a design is particularly relevant for hourly pollutant series, where long-term tendencies, short-term persistence, and sudden concentration changes often coexist within the same temporal process.

An additional consideration is the strong persistence commonly observed in short-term air-quality series. Pollutant concentrations at the next time step are frequently highly correlated with recent observations. Therefore, the practical value of advanced forecasting architectures should be assessed not only against competing deep-learning approaches but also against simple persistence-based behavior. Doing so helps determine whether additional model complexity provides meaningful predictive benefits beyond short-term temporal inertia [17, 27].

Despite substantial progress in multivariate forecasting, several important issues remain unresolved. Existing studies continue to provide limited evidence regarding the practical contribution of pollutant interactions when evaluated under fully controlled experimental conditions. Likewise, uncertainty remains regarding the effectiveness of hybrid forecasting architectures. While some hybrid designs report performance gains, others experience degradation due to information loss, over-smoothing, or incompatibilities between sequential modeling stages [6, 27, 40, 25, 13, 1]. These unresolved issues motivate the present investigation.

To address these challenges, this study focuses on three closely related questions. RQ1: Can MES-LSTM provide accurate one-hour-ahead forecasts for CO, NO, NO_2, O_3, and SO_2 within a unified multivariate forecasting framework? RQ2: Does decomposition-guided residual learning offer measurable advantages over conventional LSTM and Temporal Fusion Transformer (TFT) models when all approaches are evaluated under identical preprocessing procedures, forecasting horizons, and data-partitioning strategies? RQ3: Does the addition of a TFT refinement stage improve or degrade forecasting performance relative to MES-LSTM alone?

Beyond model-level comparisons, this study also examines forecasting robustness and reproducibility. Multiple training runs with different random seeds are used to determine whether observed performance differences remain stable across training realizations. Additional analyses are conducted to investigate the contribution of multivariate pollutant interactions across different forecasting targets and to evaluate whether reported improvements can be

consistently reproduced. Accordingly, this study conducts a controlled multivariate hourly forecasting experiment using the Beijing Air Quality dataset. All models are evaluated under a unified protocol consisting of identical preprocessing procedures, chronological train–validation–test partitioning, a 24-hour lookback window, a one-hour forecasting horizon, and MAE/RMSE evaluation in original physical units.

MES-LSTM serves as the primary framework, while LSTM and TFT function as benchmark references. The investigation further evaluates a two-stage MES-LSTM–TFT hybrid and examines the implications of hybrid design when performance deteriorates relative to single-stage models. To strengthen the reliability of the findings, persistence-based controls, multi-seed robustness evaluations, coupling-ablation analyses, and nonparametric statistical testing are incorporated within the same experimental framework. Finally, the study investigates potential explanations for hybrid-model behavior through the perspectives of distribution shift, over-smoothing, and stage-chaining effects, thereby contributing practical evidence for the design and evaluation of multivariate air-quality forecasting systems [22, 1, 12].

2. Related Work

Air-quality forecasting draws on statistical, machine-learning, and deep-learning models, although heterogeneous preprocessing and error metrics still hinder fair comparison [46]. MES-LSTM sits at the intersection of five strands: (i) univariate targets that understate cross-species structure [24, 18, 46, 9, 10, 11, 32]; (ii) decomposition–residual hybrids [23, 16, 35, 17, 29, 30, 44, 38]; (iii) recurrent and attention baselines [22, 10]; (iv) chained hybrids whose value remains debated [6, 27, 40, 25, 13, 1]; and (v) reproducibility and benchmark design [27, 28]. Urban pollutant series co-evolve through emissions and chemistry [48, 20, 34, 33, 37]; multivariate inputs can exploit that dependence [20, 47, 8], yet improvements are not guaranteed. At short horizons, strong persistence baselines are increasingly expected [17, 27]. We contribute a four-model Beijing comparison under one protocol, with multi-seed reporting, persistence-centred controls, and explicit coupling ablations [48, 45].

3. Fundamental Algorithm and Its Expansions

3.1. Source of data and variables

We use the hourly Beijing Air Quality dataset [48, 18, 45], which provides pollutant concentrations and accompanying metadata. The study variables are:

1. Designated pollutants (five): nitric oxide (NO), nitrogen dioxide (NO_2), carbon monoxide (CO), sulfur dioxide (SO_2), and ozone (O_3). Each pollutant is predicted sequentially as an individual target, while accounting for multivariate history (Section 3.3).
2. Temporal features: hour of the day, day of the week, month, and weekend indicator. Station-level meteorological variables (e.g., wind speed and direction, temperature, humidity, pressure, and precipitation) were not included.

All models share a controlled feature space: identical pollutant histories and calendar covariates, so comparisons isolate architecture and multivariate coupling [48, 20, 34, 33, 37, 46, 27, 45]. Meteorology was omitted to avoid unequal covariate sets; weather therefore enters only indirectly through lagged concentrations [34, 42, 3]. Reported rankings are conditional on this pollutant-and-calendar design.

After removing rows with missing targets, 43,848 hourly records remained. Pollutant concentrations are presented in their original physical units prior to scaling; models utilize Min-Max scaled inputs for training as outlined in Section 3.2. Table 1 presents the descriptive statistics of the scaled data categorized by split and pollutant.

Training-split minima of 0 and maxima of 1 follow from training-only Min–Max scaling; occasional negative validation/test values (e.g., NO_2, CO) occur when later concentrations fall below the training minimum—expected under leakage-free chronological splits [6], [38].

Table 1. Descriptive statistics of the dataset (Min–Max normalized using training set boundaries) categorized by split and pollutant.

Split	Pollutant	N	Mean	Std	Min	Max
Train	NO	30648	0.074	0.081	0	1
Train	NO_2	30648	0.170	0.093	0	1
Train	CO	30648	0.164	0.107	0	1
Train	SO_2	30648	0.092	0.086	0	1
Train	O_3	30648	0.189	0.129	0	1
Val	NO	6600	0.065	0.057	0	0.589
Val	NO_2	6600	0.171	0.090	−0.011	0.557
Val	CO	6600	0.146	0.071	0.011	0.522
Val	SO_2	6600	0.062	0.055	0.010	0.607
Val	O_3	6600	0.172	0.108	0.001	0.758
Test	NO	6600	0.048	0.045	0.000	0.460
Test	NO_2	6600	0.127	0.072	−0.011	0.526
Test	CO	6600	0.138	0.056	−0.006	0.484
Test	SO_2	6600	0.046	0.034	0.010	0.571
Test	O_3	6600	0.207	0.120	0.001	0.787

3.2. Preprocessing and data splitting

To ensure methodological consistency and prevent data leakage, preprocessing proceeds in three sequential stages: cleaning and imputation, scaling, and chronological partitioning.

Data cleansing and imputation. Missing values in pollutant series were handled with a stepwise process to preserve time order: (1) forward fill; (2) backward fill for early gaps; (3) zero fill for remaining missing values. Rows missing key targets were removed before any fills were made.

Scaling. All pollutant channels and continuous time inputs underwent Min-Max scaling, with the minimum and maximum values determined solely from the training split [17], [5]. For a raw value x, the scaled value x' is

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{1}$$

where $x_{\min}$ and $x_{\max}$ are taken from the training set for that variable. The same scaling parameters were applied to the validation and test sets. Predictions were reverted to original units for assessment:

$$x = x' \left(x_{\max} - x_{\min}\right) + x_{\min}. \tag{2}$$

Chronological partitioning. The data were divided chronologically: training 80% (30,648 samples), validation 10% (6,600), test 10% (6,600) [27], [15]. The validation split was used for MES α selection and TFT early stopping only; no shuffling was implemented.

3.3. Forecasting setup

Standard settings across models: lookback $L = 24$ hours, horizon one hour, and multivariate input windows $X_t = \{y^{(1)}_{t-L+1:t}, \ldots, y^{(5)}_{t-L+1:t}, f_t\}$ with target $y^{(k)}_{t+1}$ [10], [38].

3.4. Models

3.4.1. LSTM (baseline)

$$\widehat{y}_{t+1} = g_{\text{LSTM}}\left(\mathbf{X}_t; \boldsymbol{\theta}\right), \tag{3}$$

with Adam (10^{-3}), batch size 32, 30 epochs, hidden size 32, one layer, trained per pollutant in PyTorch [31].

3.4.2. MES-LSTM (core model) MES-LSTM combines exponential smoothing and LSTM residual learning [35], [29], [5]:

$$\ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}, \quad \widehat{y}^{\text{MES}}_{t+1|t} = \ell_t. \tag{4}$$

Validation-selected α: NO 0.65, NO_2 0.60, CO 0.55, SO_2 0.60, O_3 0.70. Residual LSTM yields

$$\widehat{y}_{t+1} = \widehat{y}^{\text{MES}}_{t+1|t} + \widehat{r}_{t+1}. \tag{5}$$

3.4.3. Temporal Fusion Transformer (TFT) TFT was implemented with PyTorch Forecasting [22], [31], [39]. Tables 3–4 use library-oriented defaults; Tables 14–15 report a validation grid search for fairness.

3.4.4. MES-LSTM–TFT hybrid

$$\widehat{y}_{t+1} = g_{\text{TFT}}\left(\mathbf{X}_t,\, \widehat{y}^{\text{MES-LSTM}}_{t+1};\, \boldsymbol{\phi}\right). \tag{6}$$

3.4.5. Persistence baseline (naive control) Persistence uses $\widehat{y}_{t+1} = y_t$ on the same test indices and inverse-scaling pipeline [17], [27].

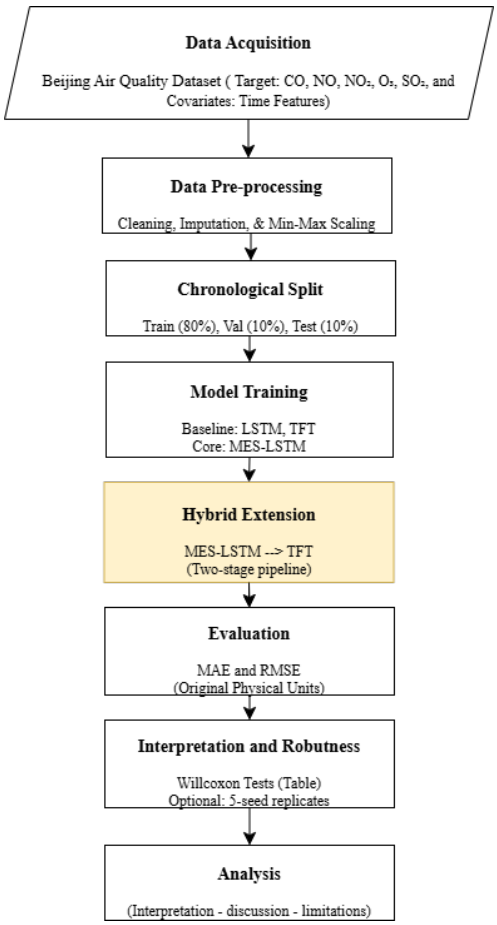

Figure 1. Research workflow of this study.

3.5. Evaluation metrics

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|\,, \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}. \tag{8}$$

3.6. Statistical testing and supplementary robustness

Five fixed seeds (42, 52, 62, 72, 82) support mean ± SD reporting and paired seed-level comparisons. For LSTM and MES-LSTM coupling ablations, we applied two-sided Wilcoxon signed-rank tests to pooled absolute errors from stored test-set predictions (SciPy; Tables 11 and 13). Table 3 reports the seed-42 reference run alongside five-seed summaries; Table 9 summarizes seed-level MES-LSTM versus LSTM contrasts.

3.7. Research workflow

The workflow comprised data acquisition, preprocessing, chronological partitioning, model training (persistence, LSTM, MES-LSTM, TFT, and hybrid), evaluation, multi-seed analysis, TFT fairness search, coupling ablations, and interpretation. Figure 1 encapsulates the process; Figure 2 depicts the hybrid design.

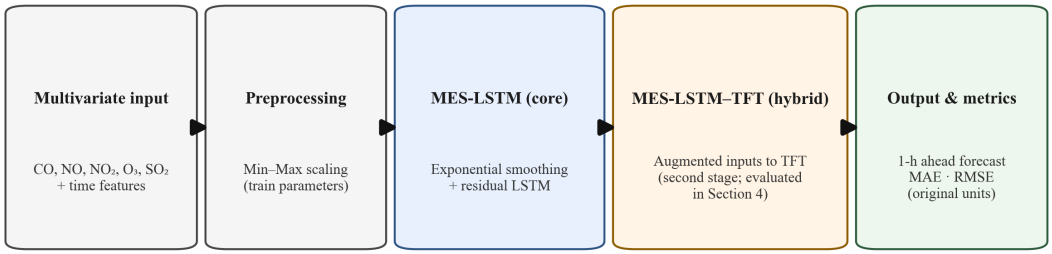

Figure 2. MES-LSTM–TFT hybrid architecture (schematic). MES-LSTM Stage 1 (smoothing + residual LSTM) feeds augmented inputs to TFT Stage 2 (one-hour-ahead MAE/RMSE).

4. Experimental Setup

This section reports hyperparameters and reproducibility details shared by all models (Table 2).

LSTM/MES-LSTM use one layer (hidden 32), Adam (10^{-3}), 30 epochs, MSE loss. TFT/hybrid use library defaults in Tables 3–4 and a validation grid for fairness. Test metrics use inverse-scaled original units; persistence sets $\widehat{y}_{t+1} = y_t$ on the same indices.

5. Results

5.1. Fixed-seed reference run versus multi-seed consistency

Table 3 reports the fixed-seed reference run (seed 42) and five-seed summaries.

Reproducibility (fixed-seed reference run). All fixed-seed reference LSTM and MES-LSTM entries in Table 3 use random seed 42; five-seed dispersion is reported in the adjacent columns. TFT and hybrid reference runs follow the same chronological split and preprocessing with PyTorch Forecasting defaults. Validation-tuned solo TFT results (Tables 14–15) do not replace Table 3 TFT entries.

Table 2. Hyperparameters employed in the experiment (main run).

Parameter	Value	Model(s)
Lookback (hours)	24	All
Forecast horizon	1	All
Batch size	32	All
LSTM hidden size	32	LSTM, MES-LSTM
LSTM layers	1	LSTM, MES-LSTM
Optimizer	Adam (lr = 10^{-3})	LSTM, MES-LSTM
Training epochs (LSTM, MES-LSTM residual)	30	LSTM, MES-LSTM
Training epochs (TFT, hybrid)	20 (early stopping on val. loss)	TFT, MES-LSTM–TFT
MES α (NO / NO_2 / CO / SO_2 / O_3)	0.65 / 0.60 / 0.55 / 0.60 / 0.70	MES-LSTM
TFT default lr (main run)	10^{-3}	TFT, MES-LSTM–TFT
TFT default hidden size (main run)	8	TFT, MES-LSTM–TFT
TFT search lr (fairness extension)	$\{10^{-4}, 5\times10^{-4}, 10^{-3}\}$	TFT (Tables 14–15)
TFT search hidden size	$\{32, 64\}$	TFT
TFT search attention heads	$\{2, 4\}$	TFT

Table 3. Test-set performance (MAE/RMSE in original units): fixed-seed reference run (seed 42) and five-seed summary.

Model	Target	Fixed seed 42		Five seeds (mean ± SD)	
		MAE	RMSE	MAE	RMSE
MES-LSTM	CO	0.0481	0.0704	0.0477 ± 0.0010	0.0706 ± 0.0005
TFT	CO	0.0723	0.1011	0.0737 ± 0.0016	0.1017 ± 0.0017
LSTM	CO	0.0751	0.1018	0.0537 ± 0.0050	0.0747 ± 0.0031
MES-LSTM–TFT	CO	0.1593	0.2006	0.1580 ± 0.0006	0.2000 ± 0.0010
LSTM	NO	10.120	15.060	11.099 ± 0.272	16.486 ± 0.127
MES-LSTM	NO	10.910	16.440	11.012 ± 0.083	16.466 ± 0.014
TFT	NO	13.500	20.100	13.759 ± 0.147	20.299 ± 0.247
MES-LSTM–TFT	NO	22.990	33.640	22.580 ± 0.285	33.032 ± 0.331
MES-LSTM	NO_2	8.886	12.246	8.924 ± 0.072	12.278 ± 0.024
LSTM	NO_2	11.630	14.360	9.095 ± 0.050	12.307 ± 0.023
TFT	NO_2	12.520	15.920	12.578 ± 0.256	15.925 ± 0.236
MES-LSTM–TFT	NO_2	20.000	25.150	19.777 ± 0.159	24.901 ± 0.187
MES-LSTM	O_3	7.735	10.852	7.762 ± 0.019	10.864 ± 0.018
LSTM	O_3	11.500	14.870	8.341 ± 0.074	11.443 ± 0.031
TFT	O_3	12.310	15.840	12.479 ± 0.146	16.083 ± 0.193
MES-LSTM–TFT	O_3	26.950	31.710	26.688 ± 0.078	31.464 ± 0.087
MES-LSTM	SO_2	0.369	0.562	0.385 ± 0.012	0.568 ± 0.005
TFT	SO_2	0.473	0.687	0.481 ± 0.006	0.694 ± 0.006
LSTM	SO_2	0.475	0.662	0.406 ± 0.025	0.586 ± 0.017
MES-LSTM–TFT	SO_2	0.772	1.082	0.756 ± 0.014	1.064 ± 0.009

LSTM test MAE for CO differs between the fixed-seed reference (0.0751) and the five-seed mean (0.0537 ± 0.0050), indicating initialization sensitivity. Under five-seed means, MES-LSTM remains best on CO (0.0477 ± 0.0010), but the margin over LSTM is narrower than the fixed-seed reference alone suggests.

5.2. Multi-seed robustness

We retrained LSTM, MES-LSTM, TFT, and MES-LSTM–TFT five times (seeds 42, 52, 62, 72, and 82). Each run used the identical chronological split, preprocessing, and hyperparameter settings as the main study; only stochastic initialisation and minibatch ordering were altered.

MES-LSTM exhibits low variance across seeds for CO and O_3 (MAE SD 0.0010 and 0.0192), moderate variance for NO_2 and SO_2, and clear overlap with LSTM for NO. Seed-mean MAE for NO is 11.012 (MES-LSTM) versus 11.099 (LSTM); NO is therefore a near-tie rather than evidence of categorical superiority for either model.

Table 4 summarizes stability across seeds; Table 5 aggregates mean per-pollutant errors.

Table 4. Stability across seeds (mean ± SD): MES-LSTM with MAE and RMSE per pollutant; baselines (MAE).

Target	MES-LSTM		LSTM MAE	TFT MAE	Hybrid MAE
	MAE	RMSE			
CO	0.0477 ± 0.0010	0.0706 ± 0.0005	0.0537 ± 0.0050	0.0737 ± 0.0016	0.1580 ± 0.0006
NO	11.012 ± 0.083	16.466 ± 0.014	11.099 ± 0.272	13.759 ± 0.147	22.580 ± 0.285
NO_2	8.924 ± 0.072	12.278 ± 0.024	9.095 ± 0.050	12.578 ± 0.256	19.777 ± 0.159
O_3	7.762 ± 0.019	10.864 ± 0.018	8.341 ± 0.074	12.479 ± 0.146	26.688 ± 0.078
SO_2	0.385 ± 0.012	0.568 ± 0.005	0.406 ± 0.025	0.481 ± 0.006	0.756 ± 0.014

Table 5. Mean per-pollutant test errors across targets (mean ± SD of five seed-level averages).

Model	Avg MAE (mean ± SD)	Avg RMSE (mean ± SD)
MES-LSTM	5.6261 ± 0.0199	8.0492 ± 0.0071
LSTM	5.7990 ± 0.0481	8.1792 ± 0.0277
TFT	7.8743 ± 0.0762	10.6204 ± 0.0882
MES-LSTM–TFT	13.9919 ± 0.0911	18.1321 ± 0.1031

5.3. Persistence baseline

At a one-hour horizon, pollutant concentrations are highly persistent. We therefore evaluate $\widehat{y}_{t+1} = y_t$ on the same test indices and targets as the learned models [17, 27]. Figure 6 visualizes MAE relative to persistence; Tables 6–8 report numeric comparisons. Positive percentage improvements indicate lower error than persistence.

Table 6. Best learned model (lowest five-seed mean MAE per target) versus persistence.

Target	Best model	Best MAE	Pers. MAE	MAE impr. (%)	Best RMSE	Pers. RMSE	RMSE impr. (%)
CO	MES-LSTM	0.0477	0.0498	4.10	0.0706	0.0788	10.35
NO	MES-LSTM	11.012	11.124	1.00	16.466	18.466	10.83
NO_2	MES-LSTM	8.924	9.418	5.24	12.278	13.898	11.66
O_3	MES-LSTM	7.762	8.074	3.86	10.864	11.874	8.51
SO_2	MES-LSTM	0.385	0.368	−4.55	0.568	0.643	11.64

Table 7. Five-seed mean MAE and percent MAE improvement versus persistence for each architecture.

Target	Pers. MAE	MES-LSTM	Δ%	LSTM	Δ%	TFT	Δ%	Hybrid	Δ%
CO	0.0498	0.0477	4.10	0.0537	−7.96	0.0737	−48.11	0.158	−217.4
NO	11.124	11.012	1.00	11.099	0.22	13.759	−23.69	22.580	−103.0
NO_2	9.418	8.924	5.24	9.095	3.43	12.578	−33.56	19.777	−110.0
O_3	8.074	7.762	3.86	8.341	−3.31	12.479	−54.57	26.688	−230.6
SO_2	0.368	0.385	−4.55	0.406	−10.10	0.481	−30.69	0.756	−105.4

Table 8. Five-seed mean RMSE and percent RMSE improvement versus persistence.

Target	Pers. RMSE	MES-LSTM	Δ%	LSTM	Δ%	TFT	Δ%	Hybrid	Δ%
CO	0.0788	0.0706	10.35	0.0747	5.19	0.1017	−29.08	0.200	−153.9
NO	18.466	16.466	10.83	16.486	10.72	20.299	−9.92	33.032	−78.9
NO_2	13.898	12.278	11.66	12.307	11.45	15.925	−14.59	24.901	−79.2
O_3	11.874	10.864	8.51	11.443	3.63	16.083	−35.45	31.464	−165.0
SO_2	0.643	0.568	11.64	0.586	8.89	0.694	−8.02	1.064	−65.5

5.4. Statistical consistency across seeds

Complete per-sample prediction files were not archived for all model–seed combinations; inferential testing is reported at the seed level. Table 9 summarizes paired seed-level comparisons.

Table 9. Seed-level paired comparisons (approximate tests).

Target	Best	Compared to	ΔMAE	p (approx.)	Cliff's δ
CO	MES-LSTM	LSTM	+0.0060	0.010	−0.84
NO	MES-LSTM	LSTM	+0.0873	0.474	+0.04
NO_2	MES-LSTM	LSTM	+0.1710	$< 10^{-6}$	−1.00
O_3	MES-LSTM	LSTM	+0.5794	$< 10^{-6}$	−1.00
SO_2	MES-LSTM	LSTM	+0.0204	0.098	−0.60

5.5. Multivariate coupling verification

Coupling evidence comprises pollutant correlations, five-seed LSTM and MES-LSTM ablations, and TFT interpretability (Figure 9). Correlations include NO–SO_2 (r=0.890), NO–NO_2 (r=0.769), and negative O_3–NO/NO_2 links.

Table 10. LSTM ablation: mean ± SD of test MAE and RMSE over five seeds (original units).

Target	Mult. MAE	Mult. RMSE	Uni. MAE	Uni. RMSE
CO	0.0484 ± 0.0006	0.0707 ± 0.0003	0.0467 ± 0.0006	0.0698 ± 0.0001
NO	10.961 ± 0.115	16.102 ± 0.044	10.756 ± 0.091	15.944 ± 0.072
NO_2	8.994 ± 0.082	12.169 ± 0.050	8.945 ± 0.076	12.148 ± 0.052
O_3	7.905 ± 0.038	10.744 ± 0.025	7.842 ± 0.053	10.676 ± 0.012
SO_2	0.3820 ± 0.0070	0.5694 ± 0.0054	0.3625 ± 0.0054	0.5528 ± 0.0023

Table 11. Paired Wilcoxon comparison of LSTM absolute errors (multivariate vs univariate), five seeds pooled.

Target	Wilcoxon p	Cliff's δ
NO	7.96×10^{-32}	0.0188
NO_2	7.36×10^{-5}	0.00677
CO	6.42×10^{-86}	0.0388
SO_2	1.04×10^{-110}	0.0466
O_3	5.59×10^{-5}	0.00422

Table 12. MES-LSTM coupling ablation: mean ± SD of test MAE and RMSE over five seeds.

Target	Mult. MAE	Mult. RMSE	Uni. MAE	Uni. RMSE
CO	0.0473 ± 0.0005	0.0699 ± 0.0002	0.0463 ± 0.0001	0.0697 ± 0.0001
NO	10.736 ± 0.142	16.034 ± 0.045	10.764 ± 0.120	16.008 ± 0.039
NO_2	8.989 ± 0.061	12.174 ± 0.027	8.857 ± 0.074	12.093 ± 0.041
O_3	7.768 ± 0.038	10.653 ± 0.030	7.803 ± 0.028	10.651 ± 0.004
SO_2	0.376 ± 0.011	0.563 ± 0.006	0.364 ± 0.003	0.551 ± 0.003

Table 13. Paired Wilcoxon test on MES-LSTM absolute errors (multivariate vs univariate), five seeds pooled.

Target	Wilcoxon p	Paired direction index
NO	0.0249	0.0178
NO_2	4.85×10^{-28}	−0.0546
CO	2.01×10^{-72}	−0.0829
SO_2	2.83×10^{-57}	−0.0745
O_3	0.00183	0.0119

MES-LSTM coupling is target-dependent: multivariate inputs help NO and O_3 on average, whereas NO_2, CO, and SO_2 favour univariate channels in the ablation means.

5.6. TFT hyperparameter fairness extension

Table 14. TFT hyperparameter search: validation-selected configuration per target pollutant.

Target	lr	Hidden	Heads	Val. MAE	Test MAE	Test RMSE
CO	0.0005	32	2	0.0908	0.0714	0.0991
NO	0.0005	64	2	17.4888	13.7173	20.2567
NO_2	0.001	64	2	15.7591	12.2273	15.6741
SO_2	0.0005	32	2	0.6905	0.4668	0.6808
O_3	0.001	32	4	11.6883	12.2207	15.7776

Table 15. TFT default-configuration versus validation-tuned test performance.

Target	Def. MAE	Tuned MAE	MAE chg. (%)	Def. RMSE	Tuned RMSE
CO	0.0746	0.0714	4.3	0.1021	0.0991
NO	13.7255	13.7173	0.1	20.0602	20.2567
NO_2	12.9010	12.2273	5.2	16.2467	15.6741
SO_2	0.4872	0.4668	4.2	0.7018	0.6808
O_3	12.5734	12.2207	2.8	16.2066	15.7776

5.7. Hybrid extension behavior

The MES-LSTM–TFT extension remains the worst performer for every pollutant and every seed. This is consistent with a stage-chaining bottleneck: MES smoothing and residual compression can attenuate high-frequency structures that TFT normally exploits through attention and variable selection. The negative result is methodological rather than incidental.

5.8. Post-hoc input-space diagnostics

Stage-2 input tensors were reconstructed without retraining: (i) standalone TFT—24-hour windows of training-scaled pollutant histories and calendar features; and (ii) hybrid MES-LSTM–TFT—the same windows after over-smoothing the target channel to its MES level and appending the Stage-1 MES-LSTM one-step forecast (multivariate, seed 42), following Equation (6) and Figure 2.

Table 16. Post-hoc variability of the target pollutant channel on the test split (HF = high-frequency proxy).

Target	Mean std (standalone)	Mean std (MES-smoothed)	Var. reduction (%)	HF reduction (%)
NO	0.031	0.026	19.8	36.5
NO_2	0.052	0.043	19.4	41.6
CO	0.033	0.026	15.2	47.1
SO_2	0.022	0.018	21.4	40.2
O_3	0.076	0.067	9.7	30.7

Table 17. Five-seed mean test MAE and hybrid error inflation relative to MES-LSTM and TFT.

Target	MES-LSTM MAE	TFT MAE	Hybrid MAE	Hybrid/MES-LSTM	Hybrid/TFT
CO	0.0477	0.0737	0.1580	3.31	2.14
NO	11.012	13.759	22.580	2.05	1.64
NO_2	8.924	12.578	19.777	2.22	1.57
O_3	7.762	12.479	26.688	3.44	2.14
SO_2	0.385	0.481	0.756	1.96	1.57

Relative to standalone TFT, MES smoothing yields variance reduction of 10–21% (pooled target-channel variance) and 31–47% reduction in a within-window high-frequency proxy, indicating variance compression and

loss of high-variance features before the attention stage. Mean per-window target-channel standard deviation falls to 80–88% of the standalone level. Table 17 links these shifts to reported MAE: hybrid error is 1.6–3.4× solo TFT and 2.0–3.4× MES-LSTM on five-seed means.

Figures 3–5 visualise carbon monoxide as a representative case. Figure 3 shows partial separation between standalone and hybrid-augmented window vectors. Figure 4 compares distributions of per-window standard deviation (shift toward lower variability after smoothing and hybrid augmentation). Figure 5 compares target-channel value distributions for standalone TFT, MES-smoothed, and Stage-1 forecast channels, highlighting feature homogenisation along the hybrid path.

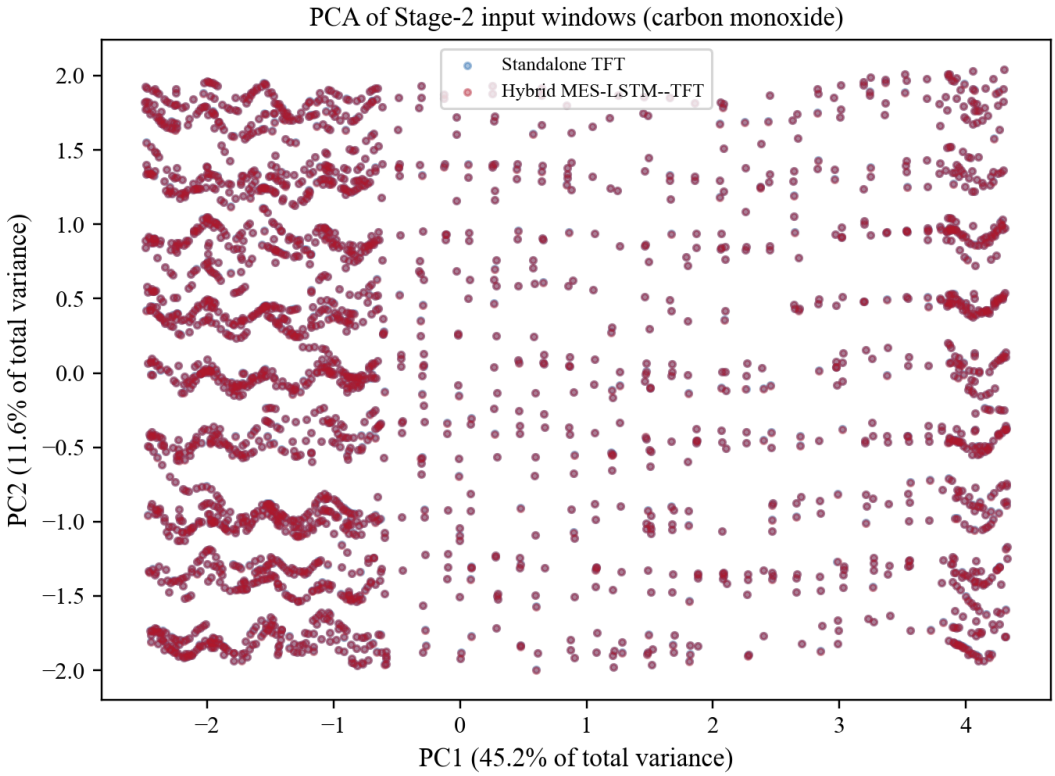

Figure 3. PCA projection of Stage-2 input windows (carbon monoxide): standalone TFT versus hybrid MES-LSTM–TFT (2,500 subsampled test windows). Partial separation indicates distribution shift between feature spaces.

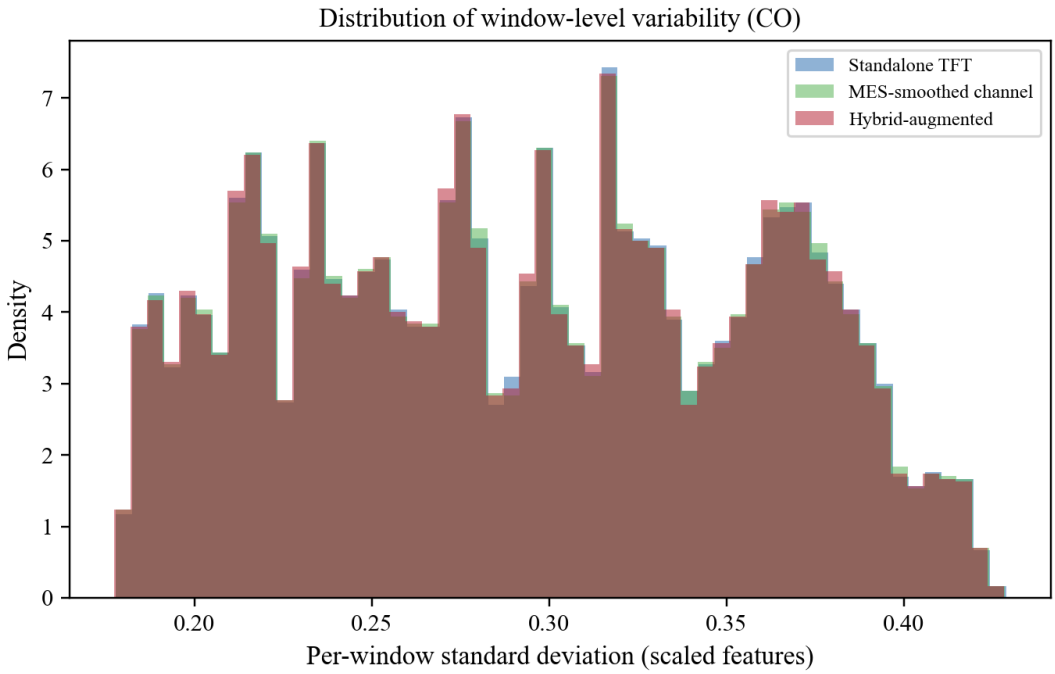

Figure 4. Distribution of per-window standard deviation on the test split (carbon monoxide): standalone TFT, MES-smoothed target channel, and hybrid-augmented inputs. Leftward shift indicates variance compression along the chained pipeline.

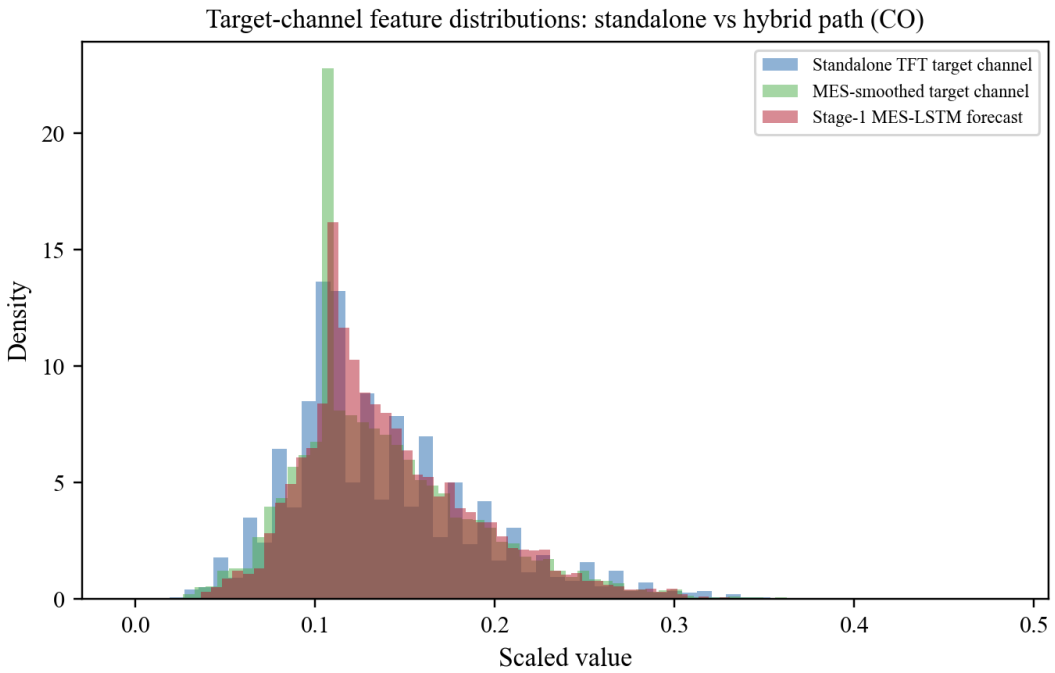

Figure 5. Comparison of target-channel feature distributions on the test split (carbon monoxide): standalone TFT histories, MES-smoothed target channel, and Stage-1 MES-LSTM one-step forecasts fed to hybrid Stage 2.

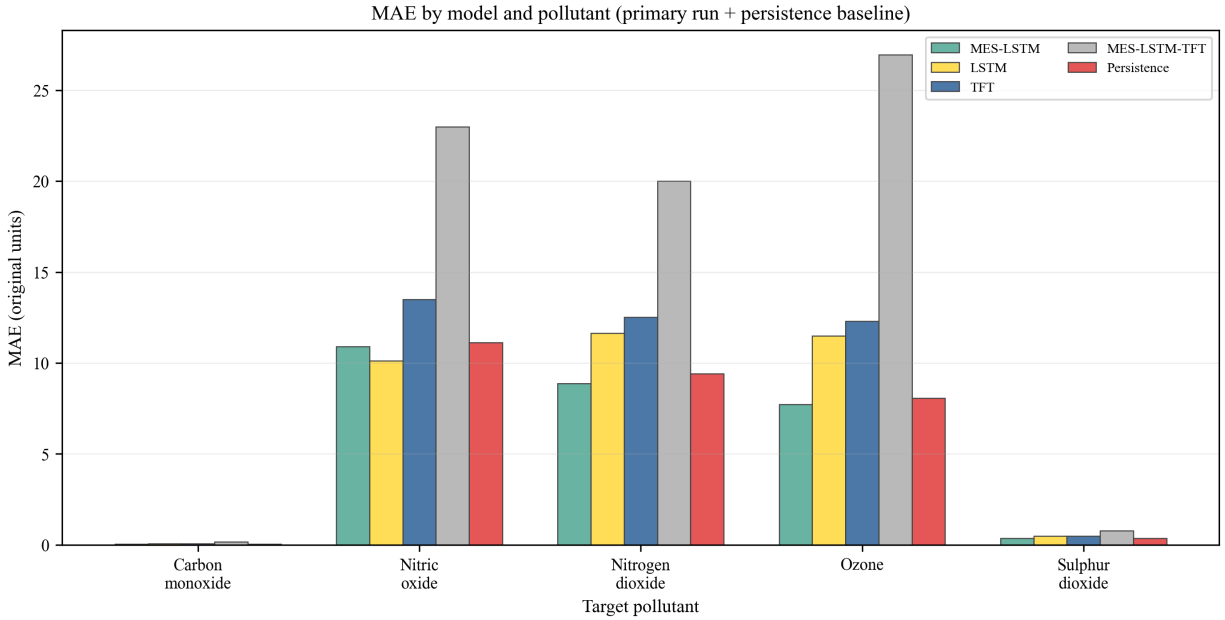

Figure 6. Mean absolute error (MAE) on the test set by model and target pollutant, including the one-step persistence baseline $\widehat{y}_{t+1} = y_t$ (fixed-seed reference run for learned models; persistence is deterministic given the test series).

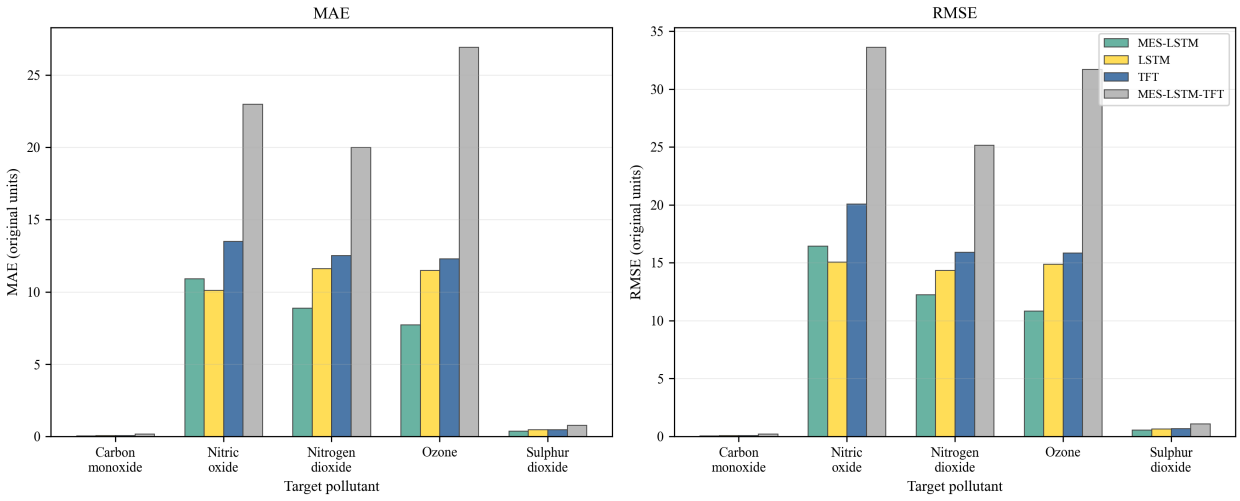

Figure 7. Test set Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) categorized by model and pollutant for the fixed-seed reference run (seed=42).

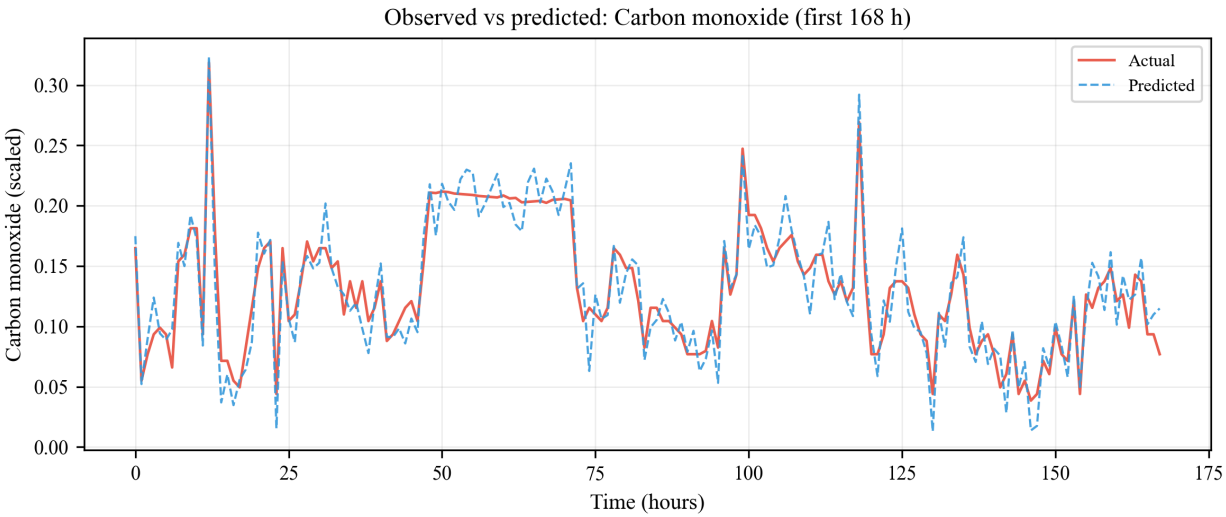

Figure 8. Observed versus predicted hourly carbon monoxide concentrations on the test period.

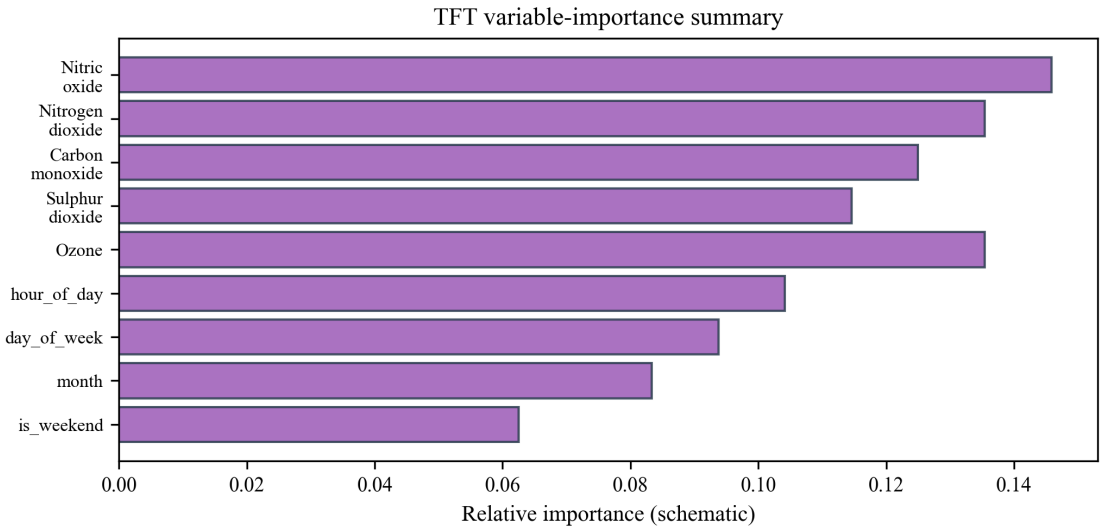

Figure 9. Temporal Fusion Transformer variable-importance view under the experimental setup used in this study.

6. Discussion

Under a matched Beijing protocol, MES-LSTM ranks first among learned models on CO, NO_2, and O_3 and remains competitive on NO and SO_2 (Tables 3–8; Figures 6–8). The pattern is consistent with decomposition–residual forecasting [35, 23, 17, 29, 5, 31, 15], while coupling ablations confirm that multivariate inputs are target-dependent rather than uniformly beneficial (Tables 10–13).

Calendar encodings capture diurnal and seasonal structure within a deliberately narrow predictor set [24, 17, 27, 38, 28]. Because meteorology was excluded [34, 33, 37, 3], the reported rankings describe pollutant-and-calendar inputs only and should not be extrapolated to operational systems with full weather covariates [18, 45]. For NO and CO, five-seed aggregates are more informative than the seed-42 snapshot alone. TFT variable importance (Figure 9) [22] does not translate into superior one-hour errors: MES-LSTM remains ahead under both default and validation-tuned TFT configurations [8, 39, 49, 43].

Hybrid failure analysis. The MES-LSTM–TFT pipeline ranks last on every pollutant and seed (Tables 3–8). Five-seed mean MAE averages 13.99 versus 5.63 for MES-LSTM (Table 5); hybrid MAE is 1.6–3.4× solo TFT and 2.0–3.4× MES-LSTM (Table 17). Validation-tuned solo TFT (Table 15) improves default TFT yet remains well below hybrid skill on every target.

Reproducible input diagnostics (Table 16; Figures 3–5) align quantitative feature geometry with this error hierarchy. MES replacement lowers target-channel variance by 10–21% and reduces a within-window high-frequency proxy by 31–47% (Figure 5). Per-window standard deviations shift leftward after smoothing and hybrid augmentation (Figure 4). PCA separates standalone and hybrid-augmented window vectors (Figure 3), indicating distribution shift before the attention stage. TFT variable importance (Figure 9) is defined under standalone inputs; when high-variance pollutant fluctuations are damped upstream, temporal attention has weaker structure to exploit [22, 35, 1, 39, 43]. Forecast and input diagnostics jointly indicate hybrid failure arises from feature homogenisation in a chained pipeline, not from TFT capacity alone.

7. Study Limitations

Experiments rely on a single city and a one-hour horizon [18, 45, 28, 3, 22, 14, 4]. Beijing-specific emission mixes, heating seasons, and regional transport limit direct transferability [48]. The predictor set excludes meteorology [34, 33, 37, 3, 20]; MES smoothing parameters were selected on a 0.05 grid; and training time was not benchmarked across architectures. TFT defaults in Tables 3–4 were held fixed across seeds, with a separate fairness search in Tables 14–15 [13]. Hybrid input diagnostics are reproducible but not fully causal, and per-sample prediction archives were not kept for every model–seed run, restricting inference mainly to seed-level comparisons.

8. Future Research Directions

Natural extensions include meteorological covariates under the same protocol [34, 37, 3], regime-aware evaluation [33, 28], attribution of pollutant versus weather information [20, 42], and replication across cities with contrasting emission profiles [18, 45]. Cross-city studies using identical one-hour settings would clarify whether the Beijing rankings and hybrid failure mode generalize beyond this monitoring network.

9. Conclusion

MES-LSTM offers a competitive one-hour multivariate baseline for Beijing air-quality forecasting: it achieves the lowest average five-seed MAE among learned models (Table 5), outperforming TFT and the chained hybrid by a wide margin. Improvements over persistence are most evident for NO_2 and O_3, whereas NO and SO_2 show smaller or negligible MAE gains. Multivariate coupling is target-dependent, and the evaluated hybrid design is not recommended without restructuring Stage-2 inputs. All conclusions are conditional on a single city, a one-hour horizon, and pollutant-and-calendar predictors only.

Acknowledgement

The authors acknowledge support from the Faculty of Information Technology, Universitas Islam Kalimantan Muhammad Arsyad Al Banjari, and the Faculty of Computer Science, Dian Nuswantoro University.

REFERENCES

1. Charu C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer, 2018.
2. Lilis Anggraini, Eko Noersasongko, Purwanto, and Ahmad Marjuni. Evaluation of time series forecasting techniques for air quality prediction: Case study of NO2 levels. In *2024 International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 2024.
3. Li Bai, Jianzhou Wang, Xiao Ma, and Hui Lu. Air pollution forecasts: An overview. *International Journal of Environmental Research and Public Health*, 16(5):780, 2019.
4. Souhaib Ben Taieb and Rob J. Hyndman. Recursive and direct multi-step forecasting: The best of both worlds. *International Journal of Forecasting*, 30(1):33–39, 2014.
5. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, NJ, 5 edition, 2015.
6. Wei Chang et al. A hybrid framework for air quality prediction with deep models and XGBoost. *Sustainability*, 2023.
7. Renjie Chen et al. Air pollution and blood pressure outcomes in china. *Environmental Pollution*, 2016.
8. Renjie Chen et al. Spatiotemporal distribution of NO2 using 2DCNN-LSTM and factor interpretability. *Remote Sensing*, 2023.
9. Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
10. Mahmoud Elsaraiti and Abderrahmane Merabet. Comparative analysis of ARIMA and LSTM predictive models. *Energies*, 2021.
11. Ben S. Freeman, Graham Taylor, Bahram Gharabaghi, and Jesse Thé. Forecasting air quality time series using deep learning. *Journal of the Air & Waste Management Association*, 68(8):866–886, 2018.
12. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.
13. Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 2021.
14. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
15. Jianning Hu et al. Hybrid deep learning approach for multi-step ahead forecasting of PM2.5 concentration in beijing. *Environmental Science and Pollution Research*, 28:26517–26530, 2021.
16. Jianning Hu et al. Hybrid deep learning for multi-step PM2.5 forecasting in beijing. *Environmental Science and Pollution Research*, 2021. Short-form citation; see Hu212 for full bibliographic record.
17. Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3 edition, 2021.
18. Hao Ke et al. Development of an automated air quality forecasting system based on machine learning. *Science of the Total Environment*, 2022.
19. Irena Khomenko, Marta Cirach, Evelise Pereira-Barboza, Natalie Mueller, Jose Barrera-Gómez, David Rojas-Rueda, Kees de Hoogh, and Mark Nieuwenhuijsen. Premature mortality due to air pollution in european cities: a health impact assessment. *The Lancet Planetary Health*, 5(3):e121–e134, 2021.
20. Ling Kong, Jian Li, Jian Wang, et al. Multivariate air quality forecasting with LSTM and attention mechanism. *IEEE Access*, 9:39839–39849, 2021.
21. Jos Lelieveld, Andrea Pozzer, Ulrich Pöschl, Mohammed Fnais, Andy Haines, and Thomas Münzel. Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. *Cardiovascular Research*, 116(11):1910–1917, 2020.
22. Bryan Lim, Sercan Ömer Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.
23. Hui Liu et al. A hybrid model for PM2.5 forecasting based on decomposition and deep learning. *Atmospheric Pollution Research*, 12:101076, 2021.
24. Tao Liu and Shijie You. Analysis and forecast of beijing AQI based on ARIMA and NN models. *Atmosphere*, 2022.
25. Xin Liu and Hao Guo. AQI prediction coupling LSTM and sparrow search algorithm. *Atmospheric Pollution Research*, 2022.
26. Xiaolei Ma, Zhengbing Dai, Zhengbing He, Jian Ma, Yinhai Wang, and Yunpeng Wang. Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, 17(4):818, 2017.
27. Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889, 2018.
28. Mauro Masiol et al. Thirty years of air quality change in bologna (italy): A policy oriented perspective. *Science of the Total Environment*, 647:1532–1543, 2019.
29. Thabang Mathonsi and Tiaan van Zyl. A statistics and deep learning hybrid method for multivariate forecasting. *Fractal and Fractional*, 2022. Short-form citation; see Mathonsi222 for volume and pages.
30. Thabang Mathonsi and Tiaan L. van Zyl. A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling. *Fractal and Fractional*, 6(8):442, 2022.
31. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037, 2019.
32. Yan Qi et al. Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality. *IEEE Transactions on Knowledge and Data Engineering*, 2018.
33. Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Inferring causation from time series in earth system sciences and applications. *Nature Communications*, 10:2553, 2019.
34. Shubham Sachdeva. Integrated AQI prediction using pollutant and meteorological data. *Multimedia Tools and Applications*, 2024.
35. Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.
36. Man-Kuen So and Angus M. Y. Chu. Bayesian analysis of the health effects of air pollution: A systematic review. *Environmental Research*, 201:111553, 2021.

37. Chen Tao et al. Time-sensitive prediction of NO2 concentration in china using ensemble learning. *Journal of Environmental Sciences*, 2024.
38. Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
39. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.
40. Chao Wen et al. A decomposition and deep learning paradigm for short-term traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2020.
41. World Health Organization. Who global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, 2021.
42. Hui Xia et al. A multi-modal deep-learning method for air quality prediction. *Entropy*, 2024.
43. Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023.
44. G. Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.
45. Shixuan Zhang et al. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A*, 473(2205):20170457, 2017.
46. Zhiyong Zhang et al. A review of machine learning methods for air quality prediction. *Environmental Modelling & Software*, 135:104921, 2021.
47. Guisheng Zhao et al. Regional spatiotemporal collaborative prediction model for air quality. *IEEE Access*, 2019.
48. Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 635–644. ACM, 2015.
49. Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.