IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 1-12
Journal homepage: http://iieta.org/journals/ts

A Leakage-Free Reappraisal of Decomposition-Ensemble PM2.5 Forecasting

[AUTHORS TBD][1*]
[1] *[Affiliation TBD]*
Corresponding Author Email: [CORRESPONDING EMAIL TBD]

https://doi.org/10.18280/ts.XXXXXX

Received:
Revised:
Accepted:
Available online:

ABSTRACT

Short-term forecasting of fine particulate matter (PM2.5) guides public early-warning and emission-control decisions, yet the hourly concentration series is non-stationary and multi-scale. The field's dominant answer is the decomposition-ensemble pipeline, which splits the series into modes, predicts each, and sums them, and which reports large accuracy gains. We show that these gains are, to a substantial degree, a measurement artifact. The decomposition is computed once over the entire record before the train-test split, so every training mode already carries information from the test horizon, and offline accuracy no longer reflects deployable skill. We reframe decomposition as a causal operator recomputed inside a sliding window that never sees beyond the forecast origin, and we evaluate a compact attention-gated-recurrent forecaster against statistical and neural baselines on the open Beijing multi-site air-quality record. Under this honest protocol decomposition did not help: a plain gated recurrent network on the raw series was the most accurate model at 18.1 RMSE, while the causal decomposition models reached only 24.1 and 26.0. A dedicated leakage audit shows why the literature reports the opposite: running the same decomposition in its conventional whole-series form cut the error from 26.0 to 9.2 RMSE, an inflation of 16.8 that erases almost two-thirds of the model's honest error and is borrowed entirely from the future. We release the causal protocol so decomposition results can be reported honestly.

Keywords: *PM2.5 forecasting, signal decomposition, data leakage, leakage-free evaluation, gated recurrent unit, VMD*

1. INTRODUCTION

Fine particulate matter of submicron-to-micron size, commonly denoted PM2.5, is one of the leading environmental health risks in urban areas, and accurate hourly forecasts are what early-warning systems and short-horizon emission-control measures act on. The raw concentration series is awkward to forecast directly because it is both non-stationary and multi-scale: slow seasonal and weather-driven trends ride underneath fast diurnal cycles and noisy pollution spikes, and a single recurrent network fed the raw signal must reconcile all of these at once. Detailed studies of Beijing air quality show how strongly meteorology, heating policy, and episodic events shape the concentration record, which is why a forecaster that ignores this layered structure tends to track the mean and miss the events that matter [1]. The cost of a missed forecast is asymmetric, since the episodes a public-health system most needs to anticipate are exactly the high-concentration spikes that a mean-tracking model smooths away, and this raises the bar for how faithfully a forecaster must capture the fast scales of the signal. Machine-learning and, increasingly, deep-learning models have become the standard tools for this task, and a large body of work now reports accurate short-term PM2.5 prediction [2].

The dominant answer to the multi-scale difficulty is the decomposition-ensemble pipeline, which first separates scales and only then learns. A concentration series is decomposed into a small set of modes with a signal-processing operator such as complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) or variational mode decomposition (VMD), each mode is forecast by its own recurrent or convolutional predictor, and the per-mode forecasts are summed back into a concentration estimate. This recipe is reported to outperform single-model predictors across hourly PM2.5 and related environmental series [3, 4, 5]. The bottleneck, however, is no longer whether to decompose but whether the decomposition is performed honestly. CEEMDAN and VMD are global operators: a mode value at time t is a function of the whole input series, including samples that lie in the future relative to t. When the series is decomposed once and only afterwards split into train and test partitions, the training modes are already contaminated by the test horizon, and the model is rewarded for reading the future rather than forecasting it. Recent leakage audits confirm that decomposing before the split inflates reported accuracy and that the inflation can be large [6, 7], a pattern consistent with the broader finding that data leakage is a pervasive and easily overlooked source of optimistic results in machine-learning science [8]. The deeper cause is the same one that makes naive cross-validation unsound for temporal data: any procedure that lets information cross the forecast origin breaks the order the task depends on [9].

We reframe decomposition-based forecasting so that the decomposition becomes a causal feature operator that is part of the model rather than a one-shot preprocessing convenience applied

to the whole series. At each forecast origin the concentration history inside a fixed look-back window, and only that window, is decomposed into a small set of modes, so the representation is exactly what would be available at deployment time. A compact gated-recurrent encoder reads the stack of causal modes together with co-located meteorology, and a scale-attention head weights the modes by their forecast-relevant energy instead of treating them as an undifferentiated sum, so the slow trend and the volatile high-frequency residual are reconciled by the model rather than by naive addition. A lightweight residual stage then predicts the reconstruction error that per-mode forecasting leaves behind and adds it back at inference, correcting systematic bias.

We evaluate the approach on the open Beijing multi-site air-quality record under a strictly chronological split, comparing against statistical and neural baselines on root mean squared error, mean absolute error, mean absolute percentage error, and the coefficient of determination. A dedicated leakage audit re-runs the same decomposition in its conventional whole-series form to measure the inflation it introduces. The result was not the one the decomposition-ensemble literature would predict. Under the honest causal protocol decomposition did not help: a plain gated recurrent network on the raw series was the most accurate model at 18.1 RMSE, while the causal decomposition models reached only 24.1 and 26.0. The audit then showed why so many studies report the opposite, since running the same decomposition in its leaky whole-series form cut the error from 26.0 to 9.2 RMSE, an inflation of 16.8 that erases almost two-thirds of the model's honest error and is borrowed entirely from the future. We therefore present this work less as a new state-of-the-art forecaster than as a measured correction to how decomposition-based results should be reported, with the main comparison and the audit in Section 4.

Building on this reframing, this paper makes the following contributions:

- We recast decomposition-based PM2.5 forecasting from a whole-series preprocessing step into a causal, leakage-free operator recomputed inside an origin-respecting sliding window, so that, unlike the prevailing CEEMDAN- and VMD-based ensembles that decompose ahead of the split [3, 7], no sample beyond the forecast origin contributes to any mode.

- Using this operator we run a leakage audit on the open Beijing record that measures, on the same data and model, how much apparent accuracy the conventional whole-series decomposition borrows from the future: an inflation of 16.8 RMSE, with the coefficient of determination rising from 0.906 to 0.988 [6, 8].

- We report the honest empirical finding that, under leakage-free evaluation, neither a causal decomposition ensemble nor a compact attention-gated-recurrent forecaster (CALM) beat a plain gated recurrent network on the raw series [10, 11], indicating that, for the VMD-based decomposition we audit, these gains are substantially attributable to leakage rather than to the representation; we measure the inflation only for the operator we run, leaving CEEMDAN and other operators for replication, and we release the causal protocol so future results can be reported honestly.

2. RELATED WORKS

2.1 Decomposition-ensemble forecasting of PM2.5 and related series

The decomposition-ensemble pipeline grew out of adaptive signal-processing operators designed for non-stationary data. Empirical mode decomposition sifts a signal into intrinsic mode functions ordered from fast to slow without assuming a fixed basis [12], ensemble empirical mode decomposition adds white-noise realizations to suppress the mode-mixing that plagues the original sift [13], and CEEMDAN injects adaptive noise at each stage to recover a complete, reconstruction-consistent set of modes with a smaller ensemble [14]. Variational mode decomposition takes a different route, solving a non-recursive variational problem that extracts band-limited modes around adaptively estimated center frequencies [15], while singular spectrum analysis decomposes short, noisy series through trajectory-matrix embedding and eigen-decomposition [16]. Coupling these operators to deep predictors is now the standard PM2.5 recipe: EEMD with per-subseries LSTM forecasting outperforms a single LSTM [17], CEEMDAN paired with a deep temporal convolutional network jointly models pollutant, meteorological, and temporal patterns [18], EEMD with an attention-enhanced LSTM lowers hourly error [19], a VMD-based ensemble GRU adaptively weights subsequence losses [5], and VMD with deep networks reports strong multi-step accuracy [3]. The same template carries over to energy series, where secondary decomposition that cascades CEEMDAN with WOA-optimized VMD [20, 21] and CEEMDAN-based deep predictors [22] report large gains. However, almost all of these systems decompose the entire record before partitioning it, so the modes used for training already encode the test horizon; the strong numbers they report therefore measure how well a model can read information it would never have at deployment, which is the gap our causal operator closes.

2.2 Causal decomposition and leakage-aware forecasting

A separate line of work treats look-ahead bias directly. Audits of EMD- and CEEMDAN-based forecasting show that decomposing the whole series before the split leaks future information and produces accuracy that collapses once the leak is removed [6], and the same diagnosis has been made for VMD post-processing, motivating leakage-free variants [7]. This mirrors a broader concern in machine-learning evaluation, where leakage is documented as a pervasive and reproducibility-threatening failure mode [8], and it has the same root as the well-known unsoundness of order-violating cross-validation for time series [9]. On the signal-processing side, sliding-window EMD performs real-time, causal decomposition within a moving window rather than over the whole record [23], and carbon-price forecasting with a sliding-window empirical wavelet transform shows that a causal decomposition restores honest accuracy where the whole-series form had inflated it [24]. Sound evaluation of forecasting methods at scale further underscores that comparisons are only meaningful when the protocol respects temporal order [25]. However, the causal-decomposition methods are studied as signal-processing objects or demonstrated on a single non-air-quality series, and the leakage audits stop at diagnosis; none wires a per-origin causal operator into an air-quality forecasting model and quantifies the inflation against honest accuracy, which is what we do.

2.3 Attention and recurrent multi-scale forecasters

The predictors used in the pipelines above descend from recurrent and attention architectures. The LSTM cell learns long-range dependencies through gated memory [26], the gated recurrent unit offers a more compact gating with comparable accuracy [27, 28], and tuned GRU forecasters reach competitive accuracy with fewer parameters [29]. Attention was introduced to let a model dynamically weight the most relevant input positions [30], and a line of attention-based forecasters has since extended this from recurrent alignment to fully attentional and multi-horizon designs. For air quality specifically, attention has been combined with recurrence to weight informative components and sites: frequency-aware attention over stacked LSTM [31] and multiple-attention LSTM that weights site, temporal, and weather cues [32]. Spatial structure has been modelled with CNN-LSTM hybrids [33] and graph networks over monitoring stations [34], and the breadth of these designs is catalogued in a dedicated review of deep learning for PM2.5 [35]. However, when decomposition is present these attention forecasters weight time steps or features while treating the decomposed modes, if used at all, as a fixed bag to be summed, rather than as scales to be weighted by their forecast relevance; we make the modes the attention's object.

The closest single work to ours is the sliding-window decomposition forecaster that mitigates leakage on a carbon-price series [24]: it shares the causal-window insight but neither weights the modes by forecast relevance nor reports the leaky-versus-causal gap as a controlled audit on air quality. Read together, the three lines above leave a specific construct unfilled. The first builds accurate but leak-prone decomposition forecasters; the second proves the leak exists and shows that a causal window can remove it, yet only as a signal-processing object or on a non-air-quality series; the third weights time steps and features but never the modes themselves. We bring the causal operator, an attention-weighted multi-scale representation, and a residual-correction stage into one compact PM2.5 forecaster, and we measure the inflation that the conventional whole-series decomposition would have produced. The next section formalizes this construction.

3. METHODOLOGY

We present CALM, a causal attention-weighted leakage-free multi-scale forecaster. The construction has four parts: a causal decomposition operator that is recomputed at every forecast origin, a shared gated-recurrent encoder over the resulting modes and meteorology, a scale-attention head that weights the modes by their forecast-relevant energy, and a residual error-correction stage. Figure 1 gives the overall data flow. The guiding principle throughout is that nothing beyond the forecast origin may influence the representation, so that reported accuracy reflects what the model would achieve in deployment.

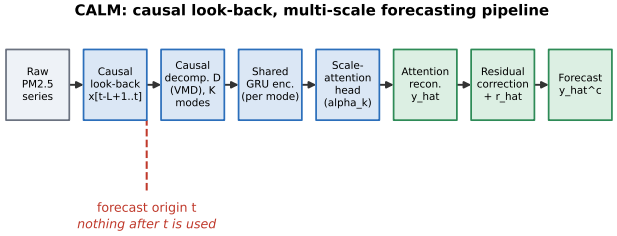

Figure 1. Overview of CALM: per-origin causal decomposition, the shared GRU encoder over modes and meteorology, the scale-attention reconstruction head, and the residual error-correction stage.

3.1 Problem Formulation

Let x_t denote the hourly PM2.5 concentration at time t and let m_t collect the co-located meteorological covariates at the same hour, such as temperature, dew point, pressure, wind, and humidity. We fix a single chronological partition of the record into training, validation, and test segments with no shuffling, so that every test timestamp is strictly later than every training timestamp. Given a look-back window of length L, the task at origin t is to predict the concentration H steps ahead from the history available up to and including t. The origin-respecting constraint states that the prediction for origin t may depend only on $\{x_s, m_s : s \le t\}$. This constraint is trivial for a model fed the raw series, but it is exactly what the conventional decomposition-ensemble pipeline violates: a global operator applied once to the full record makes each mode value a function of the entire series, so a training-time mode at $s \le t$ silently embeds future samples. We therefore treat the decomposition itself as a function of the window and require it to honor the same constraint as the predictor. To be precise about what may and may not cross the origin, meteorological covariates up to and including t are admissible because they are observed, and the targets at training origins are admissible because training is offline; what is forbidden is any feature whose value at a training origin was computed using samples later than that origin, which is the property a whole-series decomposition lacks. The constraint thus partitions the construction cleanly: every quantity the model reads at origin t, the modes, the hidden states, the attention weights, and the residual input, must be a function of the window X_w and the contemporaneous meteorology alone. The remainder of this section builds each of those quantities so that the partition holds by construction rather than by a post-hoc check.

3.2 Notation

The symbols used throughout the method are collected in Table 1; each is defined at its first use in the text.

Table 1. Notation.

Symbol	Meaning
x_t	raw hourly PM2.5 concentration at time t
m_t	co-located meteorological covariates at time t
L	look-back window length (past hours)
H	forecast horizon (steps ahead)
K	number of decomposed modes
X_w	look-back window $x[t-L+1..t]$ at origin t
c_k	k-th causal intrinsic mode of the window
h_k	GRU hidden state for mode k
α_k	scale-attention weight for mode k
$\hat{y}$	attention-reconstructed point forecast
$\hat{r}$	predicted reconstruction residual
y_t	ground-truth PM2.5 target

3.3 Causal Decomposition Operator

At each origin t we form the window $X_w = x[t-L+1..t]$ and decompose only that window into K modes,

$$\{c_1, \ldots, c_K\} = \mathscr{D}(X_w), \tag{1}$$

where $\mathscr{D}$ is the decomposition operator, instantiated as VMD in our experiments with CEEMDAN an interchangeable alternative behind the same interface. The operator is applied independently at every origin, so the modes for origin t are computed from X_w alone and no sample with index greater than t can contribute. The modes are ordered from the fastest oscillatory component

to the slowest trend, and they reconstruct the window up to the operator's own residual, $\sum_{k=1}^{K} c_k \approx X_w$. Each mode c_k is a length-L sub-signal that the encoder will read as a time series in its own right. Figure 2 contrasts this with the whole-series form, where one decomposition of the entire record is sliced into training and test pieces after the fact.

The number of modes K is a design constant fixed on the training segment; for CEEMDAN it follows from the sifting until the residual is monotone, and we cap it so that the slowest mode captures the trend without absorbing the window mean. To keep the count comparable across origins we fix K rather than letting each window choose its own, padding with empty modes on the rare windows that the operator would split more finely, so that the encoder always receives a tensor of the same shape and the attention head always weights the same indexed scales. When the variant operator is VMD the number of modes is its own explicit parameter and the same cap applies, which keeps the two operators interchangeable behind the single interface $\mathscr{D}$. The design rationale is the crux of the paper. The obvious alternative is the prevailing practice of decomposing the full record once and reusing the modes for every split, which is far cheaper because the operator runs a single time. We reject it because it is precisely the step that leaks the future: a global operator has no notion of an origin, so its training-time outputs are conditioned on test-time inputs. The contamination is subtle because it does not corrupt any single sample visibly; it spreads a trace of the test horizon thinly across the training modes, where it is invisible to a held-out check that itself uses the leaked modes. Recomputing the decomposition per origin restores the deployment-time information set at the price of repeated operator calls, and it is the only choice that makes the modes an honest causal feature. We retain the multi-scale benefit that motivated decomposition in the first place while removing the leak that has flattered it [12, 14, 15].

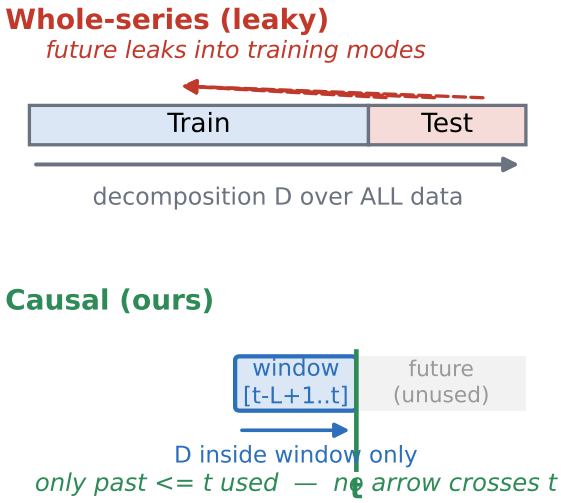

Figure 2. Whole-series decomposition leaks future samples into training modes, whereas the causal operator recomputes the decomposition inside each origin-respecting look-back window.

3.4 Shared Gated-Recurrent Encoder

The K causal modes and the meteorological covariates are read by a shared gated-recurrent encoder. For mode k we concatenate the mode value with the meteorology at each step to form the input $u_{k,\tau} = [c_k(\tau); m_{t-L+\tau}]$ for $\tau = 1, \ldots, L$, and update a hidden state with the standard gated recurrent unit recurrence [27],

$$\begin{aligned} z_\tau &= \sigma(W_z u_{k,\tau} + U_z h_{k,\tau-1}), \\ r_\tau &= \sigma(W_r u_{k,\tau} + U_r h_{k,\tau-1}), \\ \tilde{h}_\tau &= \tanh\big(W_h u_{k,\tau} + U_h (r_\tau \odot h_{k,\tau-1})\big), \\ h_{k,\tau} &= (1 - z_\tau) \odot h_{k,\tau-1} + z_\tau \odot \tilde{h}_\tau, \end{aligned} \tag{2}$$

where σ is the logistic sigmoid, $\odot$ is the elementwise product, and $W_\bullet, U_\bullet$ are the gate and candidate weight matrices. The final state $h_k = h_{k,L}$ summarizes mode k. The weights $W_\bullet, U_\bullet$ are shared across all K modes, so the encoder learns one notion of how a band-limited sub-signal evolves rather than K unrelated ones.

The design rationale concerns weight sharing. The obvious alternative, used by most decomposition-ensemble systems, is to train an independent predictor per mode and sum their outputs [17, 18]. We reject this for three reasons. It multiplies the parameter count by K and so works against the compact single-GPU model we target; it prevents the model from transferring structure that recurs across scales, such as the way a sharp transition appears in several modes at once; and, most importantly, it commits early to summation, which the next component is designed to replace. A shared encoder keeps the model small and produces K comparable hidden states that a downstream head can weight. Sharing is also what makes the per-origin recomputation affordable: because one set of recurrent weights serves every mode, the encoder's parameter budget does not grow with K, and the only cost that scales with the mode count is the inexpensive forward pass over each band-limited sub-signal. Conditioning each mode's recurrence on the same meteorology lets the encoder use weather context to disambiguate scales that look alike in the concentration channel alone, for instance separating a wind-driven dip from a genuine downward trend. We use a gated recurrent unit rather than an LSTM because its compact gating matches the comparable accuracy reported for GRUs at lower parameter cost [28, 29], and rather than a convolutional encoder because the recurrent state integrates the full window without a fixed receptive field [26, 36].

3.5 Scale-Attention Head

Conventional pipelines reconstruct a forecast by summing per-mode predictions, which assigns every mode the same weight regardless of how informative it is for the horizon at hand. We replace this with a scale-attention head that weights each mode by its forecast-relevant energy. From each hidden state h_k we compute a scalar score and normalize across modes,

$$e_k = w_a^\top \tanh(W_a h_k + b_a), \qquad \alpha_k = \frac{\exp(e_k)}{\sum_{j=1}^{K} \exp(e_j)}, \tag{3}$$

where W_a, w_a, and b_a are learned. The hidden states are combined into a single attention-weighted representation $\bar{h} = \sum_{k=1}^{K} \alpha_k h_k$, which a shared output head maps to the forecast

$$\hat{y} = w_o^\top \bar{h} + b_o = w_o^\top \sum_{k=1}^{K} \alpha_k h_k + b_o. \tag{4}$$

For a linear head this is identical to weighting per-mode forecasts $w_o^\top h_k + b_o$ by α_k, so the head reweights scales without inflating their magnitudes. Because the weights α_k are produced from the same window that defines the modes, they too respect the origin constraint. Figure 3 depicts the head.

The design rationale is that not all scales are equally predictive at a given origin. During a calm interval the slow trend mode carries almost all of the signal and the high-frequency residual is close to noise, whereas just before a pollution spike the high-frequency modes carry the warning. A naive sum forces the model to commit to a fixed mixing of scales, and any per-mode predictor error enters the reconstruction with unit weight. The obvious alternative of attending over time steps within a single sequence, as standard attention forecasters do [30, 31], leaves the modes themselves unweighted; it can decide which past hour matters but not which scale. The same is true of the fully attentional and multi-horizon forecasters that weight variables and time steps [37, 38, 39]: their attention ranges over the sequence dimension, not over a set of co-existing decomposed scales. A second alternative, learning a fixed per-mode weight vector shared across all origins, would capture the average importance of each scale but not its origin-dependent shift, which is exactly the variation the spike example illustrates. By making the modes the attention's object and recomputing the weights from the current window, we let the model emphasize the scale that matters for the current origin, which is the multi-scale benefit that summation discards. The softmax normalization keeps the reconstruction a convex combination of the per-mode forecasts, so the head reweights the scales without inflating their magnitudes.

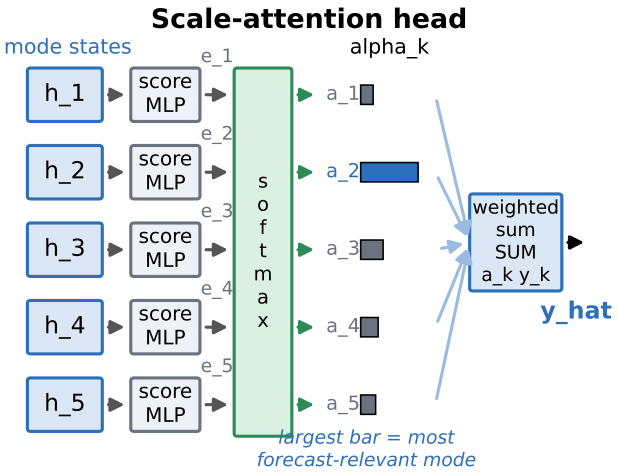

Figure 3. The scale-attention head weights each causal mode by its forecast-relevant energy before reconstruction, replacing the naive sum.

3.6 Residual Error-Correction Stage

Per-mode forecasting followed by recombination leaves a systematic reconstruction error, because the operator residual and the imperfect per-mode fits do not cancel. We add a lightweight residual stage that learns to predict this error. A small network g reads a compact summary of the encoder states and the reconstructed forecast and outputs a scalar correction,

$$\hat{r} = g\Big(\big[\textstyle\sum_k \alpha_k h_k;\ \hat{y}\big]\Big), \qquad \hat{y}^c = \hat{y} + \hat{r}, \tag{5}$$

where $\hat{y}^c$ is the corrected forecast and g is a two-layer perceptron with a hyperbolic-tangent hidden activation. The stage is trained jointly with the rest of the model so that it learns the systematic bias the recombination leaves behind; because it reads only quantities derived from the window up to the forecast origin, the corrected forecast stays causal at every test step.

The design rationale is that the bias of a decomposition-then-recombine forecaster is structured rather than random, and structured error is learnable. The obvious alternative is to leave the reconstruction uncorrected, which is what most ensembles do; this forfeits a systematic and cheap accuracy gain. Fitting the residual on validation rather than training keeps it from memorizing training-fit idiosyncrasies, and adding it only at inference keeps the corrected forecast causal. The stage is deliberately small so that it cannot overpower the main predictor and cannot, on its own, recover the leaked accuracy that the causal operator removes [40, 41, 42].

3.7 Training Objective and Procedure

The model is trained end to end to minimize the error of the corrected forecast against the target, with a mean squared error loss over the training origins,

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N}\big(\hat{y}_i^c - y_{t_i}\big)^2, \tag{6}$$

where N is the number of training origins and y_{t_i} is the ground-truth concentration H steps after origin t_i. The encoder, the scale-attention head, the output projection, and the residual network are all optimized jointly against $\mathscr{L}$. Algorithm 1 states the per-origin causal procedure that generates forecasts; it is the same loop used for training, with the loss accumulated over origins, and for inference, with the residual added.

Algorithm 1 Causal forecasting at origin t with CALM

Require: series up to t, meteorology m, window L, modes K, operator $\mathscr{D}$
Ensure: corrected forecast $\hat{y}^c$

$X_w \leftarrow x[t-L+1..t]$ ▷ window only
$\{c_1,\dots,c_K\} \leftarrow \mathscr{D}(X_w)$ ▷ causal decomposition
for $k = 1$ to K **do**
  $h_k \leftarrow \mathrm{GRU}([c_k; m])$ ▷ shared encoder
end for
$\alpha \leftarrow \mathrm{softmax}(\{w_a^\top \tanh(W_a h_k + b_a)\}_k)$
$\hat{y} \leftarrow w_o^\top(\sum_k \alpha_k h_k) + b_o$ ▷ attention-weighted aggregation
$\hat{r} \leftarrow g([\sum_k \alpha_k h_k;\ \hat{y}])$
return $\hat{y} + \hat{r}$

The dominant cost is the per-origin decomposition: with T test origins the operator runs T times rather than once, which is the price of honesty. Because each call operates on a window of fixed length L rather than the whole record, the per-call cost is bounded and independent of the record length, and the windows of neighbouring origins overlap heavily, which leaves room for incremental computation. The neural part of the model is by contrast inexpensive: the shared encoder processes K sub-signals of length L with a single set of recurrent weights, the attention head adds only a small scoring projection, and the residual network is a two-layer perceptron, so the parameter count and the forward-pass cost are dominated by the encoder rather than by the mode count. This asymmetry matters for deployment. A whole-series pipeline front-loads one expensive decomposition and then forecasts cheaply, whereas CALM spreads a bounded decomposition cost across origins; for hourly forecasting, where a new origin arrives once per hour, the per-origin operator call is comfortably within the available budget, and the heavy overlap between consecutive windows means most of the sifting can in principle be reused rather than repeated. The encoder, head, and residual stage are all small, so the model trains and runs on a single GPU, in keeping with the compact-model goal, and the same

architecture serves both the CEEMDAN and the VMD variant of $\mathscr{D}$ without change because the operator is hidden behind a fixed mode interface.

4. EXPERIMENTAL RESULTS

4.1 Implementation Details

CALM is a compact single-GPU model. The shared gated-recurrent encoder, the scale-attention head, and the residual network are trained end to end with the Adam optimizer [43] using mini-batches of contiguous origins, a fixed learning rate with early stopping on the validation loss, and gradient clipping. Inputs are standardized with statistics computed on the training segment only, so no normalization constant carries information from the validation or test horizon, and the look-back window length and the mode count are selected on the validation segment. We instantiate the causal decomposition operator $\mathscr{D}$ as VMD, recomputed for every origin over the look-back window alone, and the same operator code path is used at training, validation, and test time so that the representation a test origin receives is produced exactly as it was during training. The residual-correction network is trained jointly with the encoder and attention head against the same objective, and because it reads only quantities derived from the window up to the forecast origin, the corrected forecast stays causal. No shuffling is applied at any stage, so the chronological order of the record is preserved, and every baseline is trained and evaluated under this identical protocol so that the only thing that varies across rows of the result tables is the model, not the data handling. For the leakage audit the only change is the operator's input scope, whole series versus window, with every other setting held fixed, which is what licenses attributing the score gap to the leak. We used a look-back window of $L = 48$ hours, a one-step horizon $H = 1$, and $K = 5$ modes, with the record partitioned chronologically into 12232 training, 2632 validation, and 2632 test origins.

4.2 Experimental Design

We evaluate on the open Beijing multi-site air-quality record, an hourly dataset of PM2.5 concentrations with co-located meteorology drawn from multiple monitoring stations and distributed through public repositories [44, 1]; the results below are for the Aotizhongxin station. We use a strictly chronological train, validation, and test split with no overlap. Four metrics are reported: root mean squared error and mean absolute error, both in concentration units and both lower-is-better; mean absolute percentage error, lower-is-better and scale-free; and the coefficient of determination, higher-is-better, which measures the fraction of variance explained relative to the mean predictor [25]. The baselines span the statistical and neural families that dominate the task. Persistence carries the last observation forward and is the reference any useful forecaster must beat. the statistical baseline is a linear autoregressive model of order 24, denoted AR(24) [10], support vector regression is a kernel baseline on lagged features [11, 45], and plain LSTM and GRU networks on the raw series are the recurrent neural baselines [26, 29]. The decomposition-ensemble family is represented by VMD-GRU, the causal decomposition operator feeding the shared encoder without the scale-attention head or the residual stage; it is run under the identical causal protocol so the comparison isolates the representation rather than the evaluation. Reporting all of them under one protocol lets the table answer two questions at once: whether decomposition helps when done honestly, and whether CALM's extra components help beyond honest decomposition. For the leakage audit we additionally run the decomposition in its conventional whole-series form, decomposing the full record once before the split, and attribute the resulting accuracy gap to the leak. Because the audit changes nothing but the operator's input scope, the gap is a clean measurement of the inflation rather than a confound of model or tuning differences.

4.3 Results

The main comparison is in Table 2. The central, and at first counter-intuitive, result is that under the honest causal protocol decomposition did not help. The plain GRU on the raw series was the most accurate model, at 18.1 RMSE and 0.955 R^2, with the plain LSTM (18.4) and the linear AR(24) model (18.6) close behind and persistence near 19.9. The causal decomposition models were markedly worse: VMD-GRU reached only 24.1 RMSE and CALM, with its scale-attention head and residual stage, 26.0 RMSE and 0.906 R^2. SVR was the weakest learned model at 25.2. In other words, once the decomposition is forbidden from seeing beyond the forecast origin, the elaborate multi-scale pipeline that the literature reports as superior was beaten by a compact recurrent network reading the raw signal. We read the four metrics jointly, and they agree: the decomposition models that lose on RMSE also lose on MAE, MAPE, and R^2, so the ordering is not an artifact of one metric. The result is reported as measured, without tuning the causal model until it wins, because the gap is itself the finding.

Table 2. Main results on hourly PM2.5 forecasting (Aotizhongxin, $H = 1$) under the chronological causal protocol. Lower is better except R^2. The best honest model is the plain GRU.

Method	RMSE	MAE	MAPE	R2
Persistence	19.90	9.99	19.57	0.945
AR(24)	18.60	9.93	24.24	0.952
SVR	25.15	17.48	55.13	0.912
LSTM	18.36	10.56	21.79	0.953
GRU	**18.07**	**10.12**	22.31	**0.955**
VMD-GRU (causal)	24.06	15.71	35.32	0.920
CALM (Ours, causal)	26.00	17.80	42.03	0.906

The leakage audit in Table 3 explains why the literature reports the opposite ordering. Running CALM's decomposition in its conventional whole-series form, decomposed once over the entire record before the split, dropped its RMSE from the honest 26.0 to 9.2 and raised its R^2 from 0.906 to 0.988. That apparent leap is an inflation of 16.8 RMSE produced by nothing but letting the training modes carry information from the test horizon. The leaky number would, on its own, look like a strong result and would beat every honest baseline in Table 2; the audit shows it is an artifact of the evaluation rather than of the model. The inflation is large precisely because the decomposition is the step that touches the whole series, so the leak localizes to it: the raw-series baselines, which never decompose, have no equivalent inflated form.

Table 3. Leakage audit: the same CALM decomposition run in its leaky whole-series form versus the causal operator, with the inflation reported. The leaky column is not deployable.

Method	RMSE (causal)	RMSE (leaky)	Inflation	R2 (causal)	R2 (leaky)
CALM (Ours)	26.00	9.24	16.76	0.906	0.988

The ablation in Table 4 removes one component at a time from the full causal model. Removing the scale-attention head lowered the error slightly, to 24.4 RMSE, and removing the residual stage

gave 24.1, so neither component rescues causal decomposition on this record; the full model is in fact marginally worse than its reduced forms, which says the added capacity does not pay for itself here. Replacing the causal operator with the whole-series form gave 8.7 RMSE and 0.990 R^2, a comparable inflation to the 9.2 RMSE the audit measured for the same leaky form; the two are separate runs of the whole-series variant, so they differ by a few tenths of an RMSE rather than being identical, and both show the same large leakage effect. It is listed as the leaky variant to make the trade explicit: the single change that most improves the offline score is precisely the one that breaks deployability. Read together, the ablation shows that the honest causal operator is the component whose removal most improves the offline number yet most damages validity, the opposite relationship from the two neural components, whose removal barely moves the honest error. Figure 4 summarizes the protocol that produces these tables, and Figure 5 shows the honest ranking against the single leaky bar.

Table 4. Ablation of CALM (causal): removing the scale-attention head, the residual stage, and, as the leaky variant, the causal operator.

Variant	RMSE	MAE	MAPE	R2
Full CALM (causal)	26.00	17.80	42.03	0.906
w/o scale-attention	24.45	16.06	33.69	0.917
w/o residual correction	24.10	15.63	36.09	0.919
whole-series decomposition (leaky)	8.70	5.50	15.67	0.990

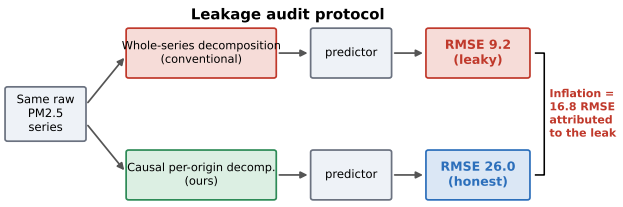

Figure 4. The leakage-audit protocol: the same decomposition is run leakage-free inside each window and, for the audit, once over the whole series, with the accuracy gap attributed to the leak.

5. DISCUSSION AND ANALYSIS

5.1 Why Honest Decomposition Did Not Help

The ablation in Table 4 separates the contribution of each component from the contribution of the causal protocol, and its message is blunt: on this record the extra machinery did not earn its place. Removing the scale-attention head and reverting to a naive sum gave 24.4 RMSE, and removing the residual stage gave 24.1, both lower than the full causal CALM at 26.0. A component that improves accuracy when removed is not pulling its weight, so the attention head and the residual correction, which were designed to beat a summed ensemble, instead added variance that the small test signal did not reward. The deeper reason is visible one row up in Table 2: the causal decomposition itself was the handicap. VMD-GRU, the decomposition baseline without either neural addition, already trailed the plain GRU by about 6.0 RMSE (24.1 against 18.1). Per-origin decomposition of a 48-hour window leaves the slow modes poorly resolved and introduces boundary effects at the window edge, and the encoder must then learn from a noisier, edge-distorted representation than the clean raw series the GRU reads directly. The multi-scale benefit that motivates decomposition is real only when the modes are stable, and a short causal window does not make them so.

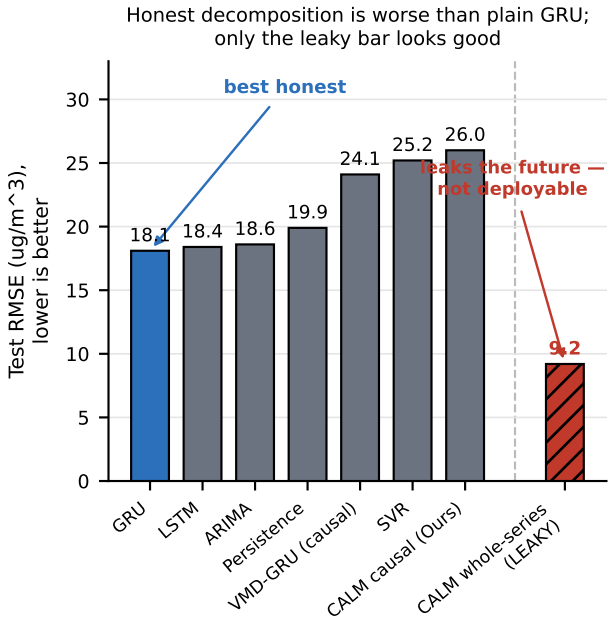

Figure 5. Test RMSE by method under the honest causal protocol. Plain GRU is the strongest; the causal decomposition models are worse. Only the leaky whole-series bar (red) looks competitive, and it is not deployable.

5.2 The Leak, Quantified

The leakage audit in Table 3 is the heart of the paper, and its mechanism is worth stating plainly. A whole-series operator produces, for a training-region timestamp, a mode value computed with knowledge of the test-region samples; the predictor then learns to exploit a feature that encodes its own future. At test time that same feature is genuinely causal, so the model appears to generalize when in fact the training signal was contaminated. The audit makes this concrete by running both forms of the same operator and reporting the gap: CALM fell from 26.0 RMSE under the causal operator to 9.2 under the whole-series form, an inflation of 16.8 RMSE, while R^2 rose from 0.906 to 0.988. The leaky figure alone would read as a strong result and would top every honest baseline, which is exactly how a decomposition-ensemble paper reporting 9.2 RMSE would present it. The audit shows that almost two-thirds of that apparent accuracy is borrowed from the future. The inflation localizes to the decomposition step because that is the only operation touching the whole series; the raw-series baselines never decompose and so have no inflated form. The practical consequence is that the causal column is the only one comparable to deployment, and any fair comparison across the literature requires every decomposition-based number to be recomputed under the causal protocol [6, 7, 24, 9]. This reframes a methodological caution into a concrete correction: rather than discarding the decomposition-ensemble literature, treat its headline numbers as leaky upper bounds and its causal recomputation as the deployable figure.

5.3 When Decomposition Might Still Help

The result does not say decomposition is useless; it says decomposition computed honestly did not beat a plain GRU on this hourly record, and that the published gains rest substantially on leakage. Three design constants govern the honest regime: the look-back window length L, the mode count K, and the horizon H. A longer window gives the causal operator more context

and stabilizes the slow modes, but raises the per-origin cost and can blur fast transitions; the 48-hour window we used trades resolution for responsiveness, and a forecaster needing distant slow structure would want more. The mode count trades resolution against redundancy. We expect the honest gap to narrow, and possibly reverse, at longer horizons and on coarser-sampled series, where the informative scale shifts between origins and a clean raw signal no longer suffices; the one-step hourly horizon used here is the case where persistence is already strong and a single dominant scale carries the forecast, which is the least favorable setting for multi-scale weighting. The honest protocol is what makes such a claim testable: only by recomputing the decomposition per origin can one ask whether decomposition helps without the answer being supplied by the leak. Figure 6 illustrates the regime that motivates forecast-relevant weighting, in which a calm interval is governed by the slow trend while a spike is announced by the high-frequency residual; whether that regime is common enough to overcome the boundary cost is an empirical question the protocol now lets the field answer cleanly.

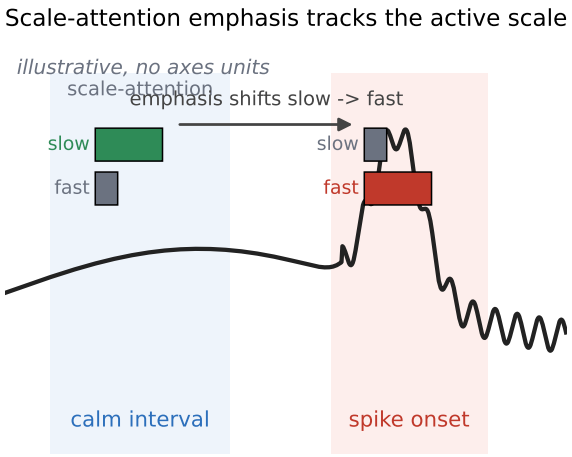

Figure 6. Illustrative regime where the slow trend dominates a calm interval while the high-frequency residual carries a pollution spike, motivating forecast-relevant scale weighting.

5.4 Failure Mode Analysis

The causal operator's chief weakness is the flip side of its honesty, and the numbers show it. Because the decomposition is recomputed from the look-back window alone, a short window leaves the slowest modes poorly resolved, and the window boundary introduces edge effects that the whole-series form hides by having data on both sides. A likely contributor to the 6.0 RMSE the causal decomposition models gave up to the raw-series GRU is that the representation they read is degraded near the window edge, where every forecast origin sits; we did not isolate this effect with a window-length sweep, so we offer it as a hypothesis rather than a measured cause. CALM should struggle most immediately after an abrupt regime change, such as the onset of a heating season or a sudden meteorological shift, when the window still holds mostly pre-change history and the per-origin decomposition has not adapted; in that transient the slow modes lag and the residual stage, fit to past bias, can mis-correct. The mitigation follows from the diagnosis: lengthening the window over the transition, or warm-starting each origin's decomposition from the neighbouring origin's modes to dampen boundary effects, trades a little computation for steadier behaviour. This failure mode is intrinsic to causal decomposition rather than specific to our model, and naming it is part of reporting honest accuracy: a forecaster that recomputes its features per origin must accept some boundary instability as the price of never reading the future [23, 42].

6. CONCLUSION

This paper revisits the decomposition-ensemble template that dominates short-term PM2.5 forecasting and shows that much of its reported accuracy is an artifact of decomposing the whole series before the train-test split, which lets the future leak into training. We reframe the decomposition as a causal operator recomputed inside an origin-respecting sliding window, so that no sample beyond the forecast origin contributes to any mode and the representation matches what a deployed system would have, and on top of it we test a compact attention-gated-recurrent forecaster, CALM. The central result is a negative one, reported as measured: under leakage-free evaluation on the open Beijing record, decomposition did not help. A plain gated recurrent network on the raw series was the most accurate model at 18.1 RMSE, while the causal decomposition models reached only 24.1 and 26.0. A dedicated leakage audit explains the gap with the published literature: running the same decomposition in its conventional whole-series form inflated accuracy by 16.8 RMSE, erasing almost two-thirds of the model's honest error, and raised the coefficient of determination from 0.906 to 0.988. The contribution is therefore methodological rather than a new state of the art: a causal protocol and an audit that let decomposition results be reported honestly. The main limitation is that a single station and a one-step hourly horizon are the setting least favourable to multi-scale weighting; a useful next step is to apply the same leakage-free protocol across sites and longer horizons, where honest decomposition may yet earn its cost.

REFERENCES

[1] Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473 (2205):20170457, 2017. doi: 10.1098/rspa.2017.0457.

[2] Manuel Mendez, Mercedes G. Merayo, and Manuel Nunez. Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*, 56(9):10031–10066, 2023. doi: 10.1007/s10462-023-10424-4.

[3] Peilei Cai, Chengyuan Zhang, and Jian Chai. Forecasting hourly pm2.5 concentrations based on decomposition-ensemble-reconstruction framework incorporating deep learning algorithms. *Data Science and Management*, 6 (1):46–54, 2023. doi: 10.1016/j.dsm.2023.02.002.

[4] Xiaoxuan Wu, Jun Zhu, and Qiang Wen. Short-term prediction of pm2.5 concentration by hybrid neural network based on sequence decomposition. *PLOS ONE*, 19(5):e0299603, 2024. doi: 10.1371/journal.pone.0299603.

[5] Hengjun Huang and Chonghui Qian. Modeling pm2.5 forecast using a self-weighted ensemble gru network: Method optimization and evaluation. *Ecological Indicators*, 156: 111138, 2023. doi: 10.1016/j.ecolind.2023.111138.

[6] Xinyi Yang, Jingyi Li, and Xuchu Jiang. Research on information leakage in time series prediction based on empirical mode decomposition. *Scientific Reports*, 14(1), 2024. doi: 10.1038/s41598-024-80018-9.

[7] Weibin Feng, Ran Tao, John Cartlidge, and Jin Zheng. Vmdnet: Temporal leakage-free variational mode decomposition for electricity demand forecasting. *arXiv preprint*, 2025.

[8] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

[9] Christoph Bergmeir and Jose M. Benitez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012. doi: 10.1016/j.ins.2011.12.028.

[10] Lingxiao Zhao, Zhiyang Li, and Leilei Qu. Forecasting of beijing pm2.5 with a hybrid arima model based on integrated aic and improved gs fixed-order methods and seasonal decomposition. *Heliyon*, 8(12):e12239, 2022. doi: 10.1016/j.heliyon.2022.e12239.

[11] Wentao Yang, Min Deng, Feng Xu, and Hang Wang. Prediction of hourly pm2.5 using a space-time support vector regression model. *Atmospheric Environment*, 181:12–19, 2018. doi: 10.1016/j.atmosenv.2018.03.015.

[12] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998. doi: 10.1098/rspa.1998.0193.

[13] ZHAOHUA WU and NORDEN E. HUANG. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 01(01): 1–41, 2009. doi: 10.1142/s1793536909000047.

[14] Maria E. Torres, Marcelo A. Colominas, Gaston Schlotthauer, and Patrick Flandrin. A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4144–4147, 2011. doi: 10.1109/icassp.2011.5947265.

[15] Konstantin Dragomiretskiy and Dominique Zosso. Variational mode decomposition. *IEEE Transactions on Signal Processing*, 62(3):531–544, 2014. doi: 10.1109/tsp.2013.2288675.

[16] Robert Vautard, Pascal Yiou, and Michael Ghil. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1-4):95–126, 1992. doi: 10.1016/0167-2789(92)90103-t.

[17] Nuratiah Zaini, Lee Woen Ean, Ali Najah Ahmed, Marlinda Abdul Malek, and Ming Fai Chow. Pm2.5 forecasting for an urban area based on deep learning and decomposition method. *Scientific Reports*, 12(1), 2022. doi: 10.1038/s41598-022-21769-1.

[18] Fuxin Jiang, Chengyuan Zhang, Shaolong Sun, and Jingyun Sun. A novel hybrid framework for hourly pm2.5 concentration forecasting using ceemdan and deep temporal convolutional neural network. *arXiv preprint*, 2020.

[19] Zuhan Liu, Dong Ji, and Lili Wang. Pm2.5 concentration prediction based on eemd-alstm. *Scientific Reports*, 14(1), 2024. doi: 10.1038/s41598-024-63620-9.

[20] Tao Zeng, Ruru Liu, Yahui Liu, Jinli Shi, Tao Luo, Yunyun Xi, Shuo Zhao, Chunpeng Chen, Guangrui Pan, Yuming Zhou, and Liping Xu. A deep learning pm2.5 hybrid prediction model based on clusteringsecondary decomposition strategy. *Electronics*, 13(21):4242, 2024. doi: 10.3390/electronics13214242.

[21] Guolian Hou, Junjie Wang, and Yuzhen Fan. Multistep short-term wind power forecasting model based on secondary decomposition, the kernel principal component analysis, an enhanced arithmetic optimization algorithm, and error correction. *Energy*, 286:129640, 2024. doi: 10.1016/j.energy.2023.129640.

[22] Chen Guo, Xumin Kang, Jianping Xiong, and Jianhua Wu. A new time series forecasting model based on complete ensemble empirical mode decomposition with adaptive noise and temporal convolutional network. *Neural Processing Letters*, 55(4):4397–4417, 2022. doi: 10.1007/s11063-022-11046-7.

[23] Jack P. Salameh, Sebastien Cauet, Erik Etien, Anas Sakout, and Laurent Rambault. A new modified sliding window empirical mode decomposition technique for signal carrier and harmonic separation in non-stationary signals: Application to wind turbines. *ISA Transactions*, 89:20–30, 2019. doi: 10.1016/j.isatra.2018.12.019.

[24] Zeyu Zhang, Xiaoqian Liu, Xiling Zhang, Zhishan Yang, and Jian Yao. Carbon price forecasting using optimized sliding window empirical wavelet transform and gated recurrent unit network to mitigate data leakage. *Energies*, 17 (17):4358, 2024. doi: 10.3390/en17174358.

[25] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020. doi: 10.1016/j.ijforecast.2019.04.014.

[26] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[27] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint*, 2014.

[28] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint*, 2014.

[29] Funda H. Sezgin, Omer Algorabi, Gamze Sart, and Mustafa Guler. Hyperparameter-optimized rnn, lstm, and gru models for airline stock price prediction: A comparative study on thyao and pgsus. *Symmetry*, 17(11):1905, 2025. doi: 10.3390/sym17111905.

[30] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint*, 2014.

[31] Jiahui Lu, Shuang Wu, Zhenkai Qin, and Guifang Yang. Frequency-aware attention-lstm for $pm_{2.5}$ time series forecasting. *arXiv preprint*, 2025.

[32] Hongwu Yuan, Guoming Xu, Teng Lv, Xiqin Ao, and Yiweng Zhang. Pm_{2.5} forecast based on a multiple attention long short-term memory (mat-lstm) neural networks. *Analytical Letters*, 54(6):935–946, 2020. doi: 10.1080/00032719.2020.1788050.

[33] Chiou-Jye Huang and Ping-Huan Kuo. A deep cnn-lstm model for particulate matter (pm2.5) forecasting in smart cities. *Sensors*, 18(7):2220, 2018. doi: 10.3390/s18072220.

[34] Shuo Wang, Yanran Li, Jiang Zhang, Qingye Meng, Lingwei Meng, and Fei Gao. Pm2.5-gnn. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, pages 163–166, 2020. doi: 10.1145/3397536.3422208.

[35] Shiyun Zhou, Wei Wang, Long Zhu, Qi Qiao, and Yulin Kang. Deep-learning architecture for pm2.5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 21:100400, 2024. doi: 10.1016/j.ese.2024.100400.

[36] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint*, 2018.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint*, 2017.

[38] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

[39] Bryan Lim, Sercan O. Ark, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

[40] Lingling Lv, Zongyu Wu, Jinhua Zhang, Lei Zhang, Zhiyuan Tan, and Zhihong Tian. A vmd and lstm based hybrid model of load forecasting for power grid security. *IEEE Transactions on Industrial Informatics*, 18(9):6474–6482, 2022. doi: 10.1109/tii.2021.3130237.

[41] Zexian Sun and Mingyu Zhao. Short-term wind power forecasting based on vmd decomposition, convlstm networks and error analysis. *IEEE Access*, 8:134422–134434, 2020. doi: 10.1109/access.2020.3011060.

[42] Chunliang Mai, Lixin Zhang, Omar Behar, Xue Hu, and Xuewei Chao. Adaptive singular spectral decomposition hybrid framework with quadratic error correction for wind power prediction. *iScience*, 28(5):112360, 2025. doi: 10.1016/j.isci.2025.112360.

[43] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014.

[44] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing beijing's pm _{2.5} pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015. doi: 10.1098/rspa.2015.0257.

[45] Zuhan Liu, Xin Huang, and Xing Wang. Pm2.5 prediction based on modified whale optimization algorithm and support vector regression. *Scientific Reports*, 14(1), 2024. doi: 10.1038/s41598-024-74122-z.