Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A hybrid patch--variate attention network with reversible instance normalization, a learned market embedding, and a linearautoregressive residual for multi-horizon panel forecasting of paddy prices

Madhusikta Tanaya Singh
Siksha O Anusandhan University
Alakananda Tripathy
alakanandatripathy@soa.ac.in

Siksha O Anusandhan University

Research Article

Keywords: agricultural price forecasting; multi-horizon time series; panel data; crosssectional factor; directional accuracy; Diebold–Mariano test

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11149500/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A hybrid patch–variate attention network with reversible instance normalization, a learned market embedding, and a linear autoregressive residual for multi-horizon panel forecasting of paddy prices

Madhusikta Tanaya Singh*[1] and Alakananda Tripathy†[1]

[1]Centre for Artificial Intelligence and Machine Learning, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

Abstract

Daily wholesale prices in Indian agricultural markets (*mandis*) are volatile, weakly autocorrelated from one day to the next, and shaped by supply arrivals, seasonality and weather. Forecasts that target the price *level* tend to reward near-persistence: they report low error while failing to anticipate the *direction* of change that matters to a farmer or trader deciding when to sell. We study short-horizon forecasting of Grade-A paddy across eleven Telangana mandis (2014–2025) and make three choices that address this. First, we pool the markets into a single panel with a learned market embedding and a cross-sectional *market factor* that carries the common component of price movement. Second, we forecast the h-step *change* rather than the level, which recovers genuine directional skill. Third, we couple the endogenous price history with real meteorological covariates (NASA POWER) and pass both through a hybrid model that combines reversible instance normalization, patch- and variate-level attention, a centralised channel-interaction module, and an explicit linear autoregressive path. Against a field of strong baselines—DLinear, PatchTST, iTransformer, LightGBM, LSTM and Bi-LSTM, all trained identically and evaluated in rupee space—the proposed model attains the lowest error at every horizon. Using the Diebold–Mariano test we find the improvement is *not* significant at one day ahead ($p = 0.17$) but becomes significant at seven and fourteen days ($p < 0.05$ and $p < 0.001$), i.e. precisely at the decision-relevant horizons. Directional accuracy reaches ∼60% against a 45.8% majority base rate. A component ablation shows that the returns reformulation and the linear autoregressive path account for most of the gain, and that the attention-based components contribute more at longer horizons. We report the horizon where the improvement is not significant as plainly as those where it is.

Keywords: agricultural price forecasting; multi-horizon time series; panel data; cross-sectional factor; directional accuracy; Diebold–Mariano test

1 Introduction

Wholesale prices for staple crops in India are set day to day in regulated market yards administered by Agricultural Market Committees (AMCs). For paddy, which underpins food security across much of the country, the timing of a sale can matter as much as the crop itself:

*Email: madhusikta05@gmail.com

†Corresponding author: alakanandatripathy@soa.ac.in

prices swing with arrivals, festival demand, procurement policy and the agricultural calendar. A reliable short-term forecast would help growers and traders decide when to bring produce to market, and would give administrators an early read on local shortages. The practical question is not "what will the price be tomorrow"—a near-useless horizon for a seller who needs lead time—but "is the price likely to rise or fall over the next one to two weeks, and by roughly how much."

The problem, stated precisely. Three properties of the series make this hard, and together they explain why much of the published work on the topic overstates its own success. *(i) The daily series is close to a random walk.* From one trading day to the next, the change is dominated by idiosyncratic noise. A model that targets the price level can therefore minimise root-mean-square error (RMSE) simply by predicting something near the last observed value. The result looks accurate yet is silent on the only quantity a decision-maker cares about—the sign of the coming change. In preliminary experiments under a level target, every model we tried—linear, recurrent and transformer—sat *below* chance on directional accuracy while still reporting $R^2 > 0.6$, a direct illustration of the trap. *(ii) Single-market studies starve deep models of data.* A lone mandi yields at most a few thousand daily observations, too few to train an attention-based model without overfitting, which pushes such studies toward weak conclusions or weak models. *(iii) Evaluation is rarely adversarial.* Comparisons are often against a naive rule or a single LSTM, three-seed dispersion is seldom reported, directional accuracy is quoted without a majority base rate, and almost no paper tests whether the reported improvement exceeds sampling noise.

How we solve it. Our response is deliberately conservative and is designed so that each choice attacks one of the three problems above. To the random-walk problem we apply a *returns reformulation*: we forecast the h-step log-price change and reconstruct the level from the known anchor, which moves the learning target onto the quantity that carries directional information and, as we show, converts a sub-chance directional baseline into $\sim$60% accuracy. To the data-scarcity problem we apply *panel pooling*: all markets are modelled jointly with a learned market embedding, raising the effective training size by an order of magnitude, and a daily *cross-sectional market factor* supplies the common component of movement that a single-market model cannot observe. To the evaluation problem we assemble a field of strong modern baselines under identical conditions and adopt the Diebold–Mariano test, a directional base-rate control, multi-seed reporting and a component ablation. The forecasting model itself is a hybrid whose parts are all established components, named honestly and chosen to match specific properties of the data (Section 4.4).

What we find, without inflation. The proposed model is best at every horizon, but its advantage over the single strongest baseline is *not* statistically significant at one day ($p = 0.17$); it becomes significant at seven days ($p < 0.05$) and strongly significant at fourteen days ($p < 0.001$)—exactly the horizons a seller can act on. Directional accuracy clears the majority base rate by 12–16 points at every horizon. A component ablation, reported as found, attributes most of the one-day skill to the returns reformulation and the linear autoregressive path rather than to the heavier neural machinery. We regard the honesty of this horizon-dependent result as part of the contribution, not a caveat to it.

Contributions.

1. **A pooled multi-market panel with a cross-sectional factor** (Section 3), which addresses data scarcity and injects the common-shock signal single-market models miss.

2. **A returns (change) formulation of the target** (Eq. 1), which recovers directional skill and reduces exactly to the naive forecast when the predicted change is zero.
3. **A hybrid architecture assembled from current forecasting components**—reversible instance normalisation (Kim et al., 2022), patch attention (Nie et al., 2023), inverted variate attention (Liu et al., 2024), centralised channel interaction (Han et al., 2024), endogenous/exogenous separation (Wang et al., 2024b) and an explicit linear path (Zeng et al., 2023)—each mapped to a named property of the data (Section 4.4, Fig. 2).
4. **Location-matched meteorological covariates** from the NASA POWER agroclimatology archive (NASA Langley Research Center, 2026).
5. **An evaluation designed to be believed**: strong baselines under identical settings, three-seed dispersion, a directional base-rate control, the Diebold–Mariano test (Diebold and Mariano, 1995), a multi-horizon study and a component ablation.
6. **An honest, horizon-dependent finding** (Section 5): best everywhere, significant only at seven and fourteen days, with the ablation crediting the returns formulation and the linear path.

The rest of the paper reviews related work (§2), formalises the problem (§3), describes the data, design rationale and method (§4), reports experiments (§5), and discusses implications and limits (§6) before concluding (§7).

2 Related work

Deep sequence models for forecasting. Recurrent networks (Hochreiter and Schmidhuber, 1997) were the first deep models widely applied to price and demand series and remain a strong baseline. The field then moved through transformer variants and, more recently, back toward simple linear models. Zeng et al. (2023) showed that a one-layer linear model on a decomposed series (DLinear) is competitive with or better than several transformers, prompting a reassessment of architectural complexity; this result is the direct motivation for the explicit linear path in our model. Nie et al. (2023) (PatchTST) segment the look-back into patches and apply attention over them, preserving local temporal semantics while shortening the attention sequence. Liu et al. (2024) (iTransformer) invert the usual layout so that attention operates across variables rather than time steps, capturing feature and market correlations. Wu et al. (2021) (Autoformer) and Wang et al. (2024a) (TimeMixer) exploit decomposition and multiscale mixing respectively. A recurring benchmarking finding—consistent with the M4 competition's verdict that no single method dominates (Makridakis et al., 2020)—is that leading methods are statistically close on standard datasets, which cautions against reading small mean differences as architectural progress and motivates our use of significance testing.

Channel interaction and exogenous variables. When many related series are modelled together, how their interaction is represented matters. Han et al. (2024) (SOFTS) replace pairwise attention with a centralised STar Aggregate–Redistribute (STAR) module that forms a global core from all channels and redistributes it, giving linear complexity and greater robustness under distribution drift—a property we value because our test period (2024–2025) is materially higher-priced than the early training years. For problems with a clear target-versus-context split—as here, price versus weather—Wang et al. (2024b) (TimeXer) keep endogenous and exogenous variables in separate representations bridged by cross-attention. Reversible instance normalisation (Kim et al., 2022) addresses the train/test distribution shift that accompanies long, trending series and is now a near-standard preprocessing step.

Agricultural and commodity price forecasting. Classical approaches rest on ARIMA-family models (Box et al., 2015) and their seasonal and exogenous extensions. Machine-learning approaches, including gradient-boosted trees (Ke et al., 2017), are strong on tabular lag features and are an essential baseline; on small single-market series they frequently match or beat deep models. Our contribution is not a new agricultural feature set but a panel formulation and an evaluation protocol that make a deep model's advantage *testable*.

Evaluation practice. The Diebold–Mariano test (Diebold and Mariano, 1995) compares the predictive accuracy of two forecasts and is standard in econometrics but rare in applied deep forecasting papers. Directional accuracy—whether the sign of the predicted change matches the realised change—is the metric closest to the decision a trader faces, yet is easily inflated by a trending test window unless compared against a majority base rate. We adopt both. Table 1 summarises the gaps we target.

Table 1: Representative practice and the gaps addressed here.

Focus	**Typical practice**	**Gap addressed**
Single-market deep models	One series; limited samples	Panel pooling with market embedding and cross-sectional factor
Point (level) forecasting	Minimise RMSE on the level	Returns target; directional accuracy vs. base rate
Weak baselines	Naive rule or a single LSTM	DLinear, PatchTST, iTransformer, LightGBM under identical settings
No significance testing	Report a lower mean	Diebold–Mariano test across horizons
One split, point forecasts	Single chronological split	Multi-horizon evaluation; component ablation

3 Preliminaries and problem formulation

We observe M markets indexed by m. For market m on trading day t let $p_{m,t} > 0$ be the modal (most frequent) price and $y_{m,t} = \log(1 + p_{m,t})$ its log transform. Each market carries a feature vector $\mathbf{x}_{m,t} \in \mathbb{R}^C$ built from its own price history, arrivals, calendar position, cross-sectional signals and same-day weather (§4.2). Given a look-back window of length L ending at the anchor day t, the task is to forecast the price h days ahead.

Returns target. Rather than predict $y_{m,t+h}$ directly, we predict the *change* and reconstruct the level:

$$\Delta^{(h)}_{m,t} = y_{m,t+h} - y_{m,t}, \qquad \hat{p}_{m,t+h} = \exp\big(y_{m,t} + \hat{\Delta}^{(h)}_{m,t}\big) - 1. \tag{1}$$

The anchor value $y_{m,t}$ is known at prediction time, so (1) introduces no leakage; it simply moves the learning target onto the quantity that carries directional information. A model that predicts $\hat{\Delta} = 0$ reduces *exactly* to the naive (persistence) forecast, which we retain as a reference. Because the change is far closer to zero-mean and stationary than the level, it is also better conditioned for a neural regressor.

Cross-sectional market factor. The market factor is the cross-market mean of log-price on each day, computed only from same-day observed prices,

$$f_t = \frac{1}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} y_{m,t}, \qquad d_{m,t} = y_{m,t} - f_t, \qquad \dot{f}_t = f_t - f_{t-1}, \tag{2}$$

where $\mathcal{M}_t$ is the set of markets trading on day t. The deviation $d_{m,t}$ is a mean-reversion signal (how far a market sits from the common level) and $\dot{f}_t$ is common momentum. Both enter $\mathbf{x}_{m,t}$. This is a hand-built, leakage-controlled analogue of the "global core" that the STAR module later learns inside the network. Symbols are listed in Table 2.

Table 2: Notation.

Symbol	Meaning
$M,\ m$	number of markets; market index
$p_{m,t},\ y_{m,t}$	modal price; log-price $\log(1 + p_{m,t})$
$L,\ h$	look-back length; forecast horizon (days)
C	feature dimension ($C = 35$; 22 endogenous, 13 exogenous)
$\mathbf{x}_{m,t} \in \mathbb{R}^C$	feature vector at the anchor day
$\Delta^{(h)}_{m,t}$	h-step log-price change (target)
$f_t,\ d_{m,t},\ \dot{f}_t$	market factor; deviation; factor momentum
μ_m	training-period mean log-price of market m
$\mathbf{e}_m \in \mathbb{R}^8$	learned market embedding
$P,\ d$	number of patches; model width ($d = 64$)

4 Materials and methods

Figure 1 gives the end-to-end workflow. We describe the data (§4.1), the features (§4.2), the leakage controls (§4.3), the design rationale that connects each data property to a model component (§4.4), the architecture (§4.5), the baselines (§4.6), training (§4.7) and evaluation (§4.8).

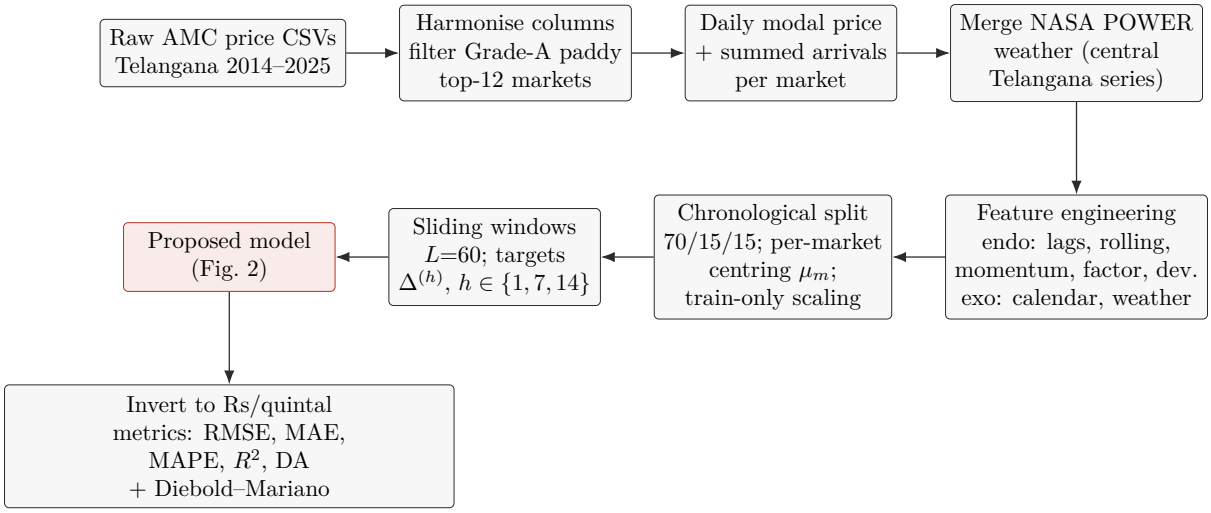


Figure 1: End-to-end methodology workflow. Raw AMC price files are harmonised and reduced to a daily panel, merged with NASA POWER weather, turned into endogenous/exogenous features, split chronologically with strict leakage controls, windowed into h-step returns targets, fed to the proposed model, and evaluated in rupee space with the Diebold–Mariano test.

4.1 Dataset and preprocessing

Source. Price records are drawn from *Daily Prices of Market Yard Commodities in Telangana* (Agricultural Marketing and Co-operation Department, Government of Telangana, 2026), published under the Agriculture theme of the Telangana Open Data Portal (`https://data.telangana.gov.in/search?theme=Agriculture`) by the Agricultural Marketing and Co-operation Department, Government of Telangana. The portal lists the dataset as covering 01-01-2014 to 30-06-2026 and reports, for every market yard and commodity in the state on every trading day, the raw schema in Table 3. We use the subset of this feed for paddy (Grade-A variety) from 2014 through 2025, described below. The exact paddy subset used in this paper is uploaded for access in our public repository at `https://github.com/Madhusikta-tv/AGRI-DATA`.

Table 3: Raw column schema of the Telangana Open Data Portal mandi price feed (Agricultural Marketing and Co-operation Department, Government of Telangana, 2026).

Column	Meaning
DDate	Date of the trading day
AmcCode	Agricultural Market Committee (AMC) code
AmcName	Agricultural Market Committee name
YardCode	Agricultural market yard code
YardName	Agricultural market yard name
CommCode	Commodity code
CommName	Commodity name
VarityCode	Commodity variety code
VarityName	Commodity variety name
Arrivals	Quantity traded, in quintals (Qtls)
Minimum	Minimum price per quintal (Rs)
Maximum	Maximum price per quintal (Rs)
Modal	Modal price per quintal (Rs): the price at which most transactions on that day took place

Filtering and harmonisation. We filter `CommName` to paddy and `VarityName` to Grade-A—the most liquid single grade, which keeps the target a coherent product rather than a blend of grades with different price dynamics. Column names and yard identifiers are mapped to a canonical schema by regular-expression matching, since the portal's per-period exports vary slightly in header formatting. `Modal` is treated as the price of record throughout the paper, as it best reflects where the bulk of trading actually clears (`Minimum`/`Maximum` bound the day's range but are thin-tailed and noisier); `Arrivals` is retained as a supply-side feature (§4.2). We select the twelve AMCs with the most complete records and aggregate to one `Modal` value and one summed `Arrivals` value per market per day (a market yard occasionally reports more than one line for a day, e.g. across sub-lots). Eleven markets retain enough pre-2023 history to estimate a stable training mean and are kept; the twelfth is dropped.

Weather. Weather is drawn from the NASA POWER daily agroclimatology archive (NASA Langley Research Center, 2026), queried for temperature (maximum/minimum), bias-corrected precipitation, mean relative humidity, wind speed and surface shortwave radiation, and merged on date. POWER is a satellite- and reanalysis-derived product at roughly half-degree resolution; we describe it as regional rather than gauge-level, and by default share one central Telangana series across markets. Fill values are treated as missing and interpolated within a market.

4.2 Features

All features are computed *within* a market so that no market's history contaminates another's. From the log-price we form lags at $\{1, 2, 3, 7, 14, 30\}$ days, rolling means and standard deviations over $\{7, 14, 30\}$ days, and one- and seven-day momentum. Arrivals enter as a log-transformed supply signal with a seven-day mean. Seasonality is encoded with sine/cosine pairs of day-of-year, month and day-of-week, which represent the calendar without imposing an artificial jump from December to January. The cross-sectional block adds the market factor f_t, its momentum $\dot{f}_t$, the deviation $d_{m,t}$ and its lag and seven-day mean (2). When weather is used, the daily meteorological variables and a seven-day rainfall mean are appended. This yields $C = 35$

features, of which 22 are endogenous (price-derived) and 13 exogenous (calendar and weather). The endogenous block is ordered first so that the endo/exo split inside the model (§4.5) is a simple index.

4.3 Splitting, scaling and leakage control

The data are split chronologically and globally by date: roughly the first 70% for training, the next 15% for validation, the last 15% for test (test spanning 2024–2025). Assigning by *date*, not by row within a market, ensures no market's future is used to predict another market's past. Each market's price-level features are centred by that market's *training* mean μ_m, so the model learns dynamics rather than which market is intrinsically dearer; the market embedding absorbs any residual level. The feature scaler (a robust median/IQR scaler) is fit on training rows only and then applied everywhere. The forecast target (1) looks strictly forward from a known anchor, and the factor (2) uses only same-day prices, so no future information enters any feature.

4.4 Design rationale: mapping data properties to components

Before the architecture, we state *why* each component is present. Every choice answers a specific property of this dataset rather than a desire for complexity; Table 4 is the map, and the ablation in §5.6 tests whether each earns its place.

Table 4: Design rationale: each component answers a named property of the data.

Data property / challenge	**Component chosen**	**Source**
Level target rewards persistence; daily change is near-noise	Returns target $\Delta^{(h)}$ + naive fallback	Eq. (1)
Few samples per market; deep models overfit	Panel pooling + market embedding $\mathbf{e}_m$	this work
A common shock moves all markets together	Cross-sectional factor f_t, deviation $d_{m,t}$	Eq. (2)
Test years are higher-priced than training (drift)	Reversible instance normalisation (RevIN)	(Kim et al., 2022)
Local temporal patterns over a 60-day window	Patch embedding + attention over patches	(Nie et al., 2023)
Correlations *across* features and markets	Inverted (variate-token) attention	(Liu et al., 2024)
Robust, linear-cost channel interaction under drift	STAR aggregate–redistribute ($\times 2$)	(Han et al., 2024)
Target (price) vs. context (weather) asymmetry	Endo/exo separation + cross-attention	(Wang et al., 2024b)
A strong linear autoregression is hard to beat	Explicit linear AR path (residual head)	(Zeng et al., 2023; Oreshkin et al., 2020)

4.5 Proposed model

The model (Fig. 2) maps a window $\mathbf{X} \in \mathbb{R}^{L \times C}$ and a market index m to a scaled change $\hat{\Delta}$. Its parts are all established components, named for what they are.

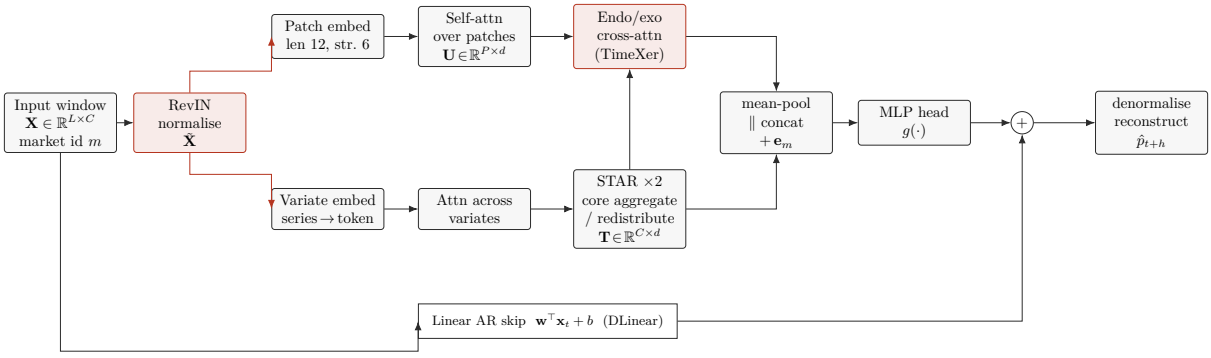


Figure 2: Proposed hybrid architecture. A RevIN-normalised window feeds two branches—a patch/temporal branch (top) and an inverted variate branch with STAR (bottom). Endogenous patch tokens attend to exogenous variate tokens (TimeXer-style). The pooled branch summaries are concatenated with the market embedding and passed to a nonlinear head; a parallel linear autoregressive path is added, and the scaled change is denormalised and reconstructed into a price via Eq. (1).

Reversible instance normalisation. Each window is normalised over time by its own per-channel statistics before the network, with a learnable affine transform (Kim et al., 2022). For channel c,

$$\mu_c = \frac{1}{L}\sum_{t=1}^{L} X_{t,c}, \quad \sigma_c^2 = \frac{1}{L}\sum_{t=1}^{L}(X_{t,c}-\mu_c)^2, \quad \tilde{X}_{t,c} = \gamma_c \frac{X_{t,c}-\mu_c}{\sqrt{\sigma_c^2+\epsilon}} + \beta_c. \tag{3}$$

This removes the level shift that separates the recent, higher-priced test period from earlier training years. Because our target is a change rather than a level, the explicit denormalisation reduces to the identity and is omitted; the learnable affine (γ, β) is retained.

Patch/temporal branch. The normalised window $\tilde{\mathbf{X}}$ is segmented into overlapping patches of length ℓ=12 and stride s=6, giving $P = \lfloor (L-\ell)/s \rfloor + 1$ patches; each patch is flattened over its ℓ steps and C channels and linearly projected to width d (Nie et al., 2023), yielding patch tokens $\mathbf{U} \in \mathbb{R}^{P\times d}$. Multi-head self-attention (Vaswani et al., 2017) is applied over the patch sequence with a residual and layer norm,

$$\mathbf{U}' = \mathrm{LN}\big(\mathbf{U} + \mathrm{MHSA}(\mathbf{U},\mathbf{U},\mathbf{U})\big). \tag{4}$$

Inverted variate branch. Following iTransformer (Liu et al., 2024), each variable's whole series is embedded as a single token: $\mathbf{T}^{(0)} = \mathbf{W}_e \tilde{\mathbf{X}}^\top \in \mathbb{R}^{C\times d}$. Attention then operates *across* the C variate tokens, capturing feature- and market-level correlations rather than time-point mixing:

$$\mathbf{T}' = \mathrm{LN}\big(\mathbf{T}^{(0)} + \mathrm{MHSA}(\mathbf{T}^{(0)},\mathbf{T}^{(0)},\mathbf{T}^{(0)})\big), \qquad \mathbf{T}'' = \mathrm{LN}\big(\mathbf{T}' + \mathrm{FFN}(\mathbf{T}')\big). \tag{5}$$

STAR channel interaction. The variate tokens are refined by two STAR layers (Han et al., 2024), which replace pairwise channel attention with a centralised aggregate–redistribute step of linear complexity. With $\mathbf{h} = \mathrm{GELU}(\mathbf{W}_g\mathbf{T})$,

$$\mathbf{o} = \mathrm{GELU}\Big(\mathbf{W}_o \tfrac{1}{C}\textstyle\sum_{c=1}^{C} \mathbf{h}_c\Big), \qquad \mathbf{T} \leftarrow \mathrm{LN}\Big(\mathbf{T} + \mathrm{Drop}\,\mathrm{GELU}\big(\mathbf{W}_f[\mathbf{T} \,\|\, \mathbf{o}]\big)\Big), \tag{6}$$

where $\mathbf{o}$ is the global "core" pooled from all channels and then broadcast back and fused. STAR is the learned counterpart of the hand-built factor in (2) and is chosen for its robustness under the 2014→2025 drift.

Endogenous/exogenous cross-attention. Following TimeXer (Wang et al., 2024b), the endogenous (price) patch tokens attend to the exogenous (calendar/weather) variate tokens $\mathbf{T}''_{\text{exo}}$ (the rows beyond the 22 endogenous channels), so external information reaches the target representation without being averaged into it:

$$\mathbf{U}_x = \mathrm{LN}\big(\mathbf{U}' + \mathrm{MHA}(\underbrace{\mathbf{U}'}_{\text{query}}, \underbrace{\mathbf{T}''_{\text{exo}}}_{\text{key}}, \underbrace{\mathbf{T}''_{\text{exo}}}_{\text{value}})\big). \tag{7}$$

Fusion, embedding and head. The two branch summaries are pooled and concatenated with the market embedding, $\mathbf{z} = [\,\overline{\mathbf{U}_x} \parallel \overline{\mathbf{T}''} \parallel \mathbf{e}_m\,]$, where $\overline{(\cdot)}$ denotes a mean over tokens. A small feed-forward head g produces the nonlinear term, and a single linear layer maps the last observed feature vector $\mathbf{x}_t$ directly to the output. The prediction is their sum:

$$\hat{\Delta}_{\text{scaled}} \;=\; \underbrace{g(\mathbf{z})}_{\text{nonlinear head}} \;+\; \underbrace{\mathbf{w}^\top \mathbf{x}_t + b}_{\text{linear AR path}}\,. \tag{8}$$

The linear path gives the model the strong autoregressive component that a plain recurrent network already has, letting the nonlinear part learn a residual on top of it—the structure behind N-BEATS and DLinear (Oreshkin et al., 2020; Zeng et al., 2023). Finally the scaled change is un-standardised with the training statistics $(\mu_\Delta, \sigma_\Delta)$ and the price is reconstructed:

$$\hat{\Delta} = \sigma_\Delta\, \hat{\Delta}_{\text{scaled}} + \mu_\Delta, \qquad \hat{p}_{m,t+h} = \exp\big(y_{m,t} + \hat{\Delta}\big) - 1. \tag{9}$$

As §5.6 shows, the linear path (8) carries most of the model's one-day skill.

4.6 Baselines

All baselines share the target (1), the features, the split, the market embedding and the training budget, so differences reflect the model and not the protocol. They span three families: linear and tree models (persistence *Naive*, ridge autoregression, and LightGBM (Ke et al., 2017) with the market as a categorical feature); recurrent models (LSTM and Bi-LSTM (Hochreiter and Schmidhuber, 1997)); and modern forecasters (DLinear (Zeng et al., 2023), and RevIN-equipped PatchTST (Nie et al., 2023) and iTransformer (Liu et al., 2024)). A mean ensemble of the strongest learners is also reported.

4.7 Training

Deep models are trained with AdamW, a smooth-L_1 (Huber) loss on the scaled change with β=0.1,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Huber}_\beta\big(\hat{\Delta}_{\text{scaled},i} - \Delta_{\text{scaled},i}\big), \tag{10}$$

a short warm-up followed by cosine decay, gradient clipping at 1.0, and early stopping on validation RMSE (in rupee space) with the best checkpoint restored. An exponential moving average of the weights (decay 0.999) is maintained for evaluation. Each deep model is trained with three seeds and we report the mean and standard deviation of each metric; the averaging is applied to every model, so it favours none.

4.8 Evaluation

Predictions are inverted to rupees per quintal via (9) before scoring, so all metrics are in the units a user cares about. We report RMSE (the root of the mean squared error), MAE, MAPE,

R^2, and directional accuracy

$$\mathrm{DA} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\mathrm{sign}(\hat{p}_i - p_{t,i}) = \mathrm{sign}(p_i - p_{t,i})]\,, \tag{11}$$

the fraction of forecasts whose sign of change relative to the anchor matches the realised sign. DA is read against the *majority base rate*—the score of always predicting the more common direction in the test window—so only an accuracy clearly above it counts as skill. Finally, the Diebold–Mariano test (Diebold and Mariano, 1995) compares the squared-error sequences of the proposed model (e^p) and the strongest baseline (e^b) at each horizon. With loss differential $g_i = (e_i^b)^2 - (e_i^p)^2$,

$$\mathrm{DM} = \frac{\bar{g}}{\sqrt{\widehat{\mathrm{Var}}(g)/N}}, \qquad \bar{g} = \frac{1}{N}\sum_i g_i, \tag{12}$$

so a positive DM favours the proposed model, with a two-sided p-value from the t_{N-1} distribution. On pooled panel data the test's independence assumption holds only approximately, and we say so.

5 Experimental results

5.1 Setup

Experiments run on a single GPU. The look-back is $L = 60$ days with patches of length 12 and stride 6; the model width is $d = 64$ with 4 attention heads and two STAR layers. The pooled test set contains 2,616 sequences. Horizons are $h \in \{1, 7, 14\}$ days.

5.2 Benchmark at one day ahead

Table 5 reports one-day-ahead results. The proposed model has the lowest RMSE and highest R^2, and the smallest seed-to-seed dispersion, but the margin over Bi-LSTM and DLinear is small (about 0.8%). Two points are worth stating plainly. First, the deep sequence models and DLinear cluster tightly, consistent with the benchmark literature's finding that leading methods are close (Makridakis et al., 2020). Second, directional accuracy is ∼59–60% against a majority base rate of 45.8%: the models have real directional skill, and it is the returns formulation—not any one architecture—that provides it, since even ridge autoregression clears the base rate by a wide margin. The naive rule, which predicts no change, has near-zero directional accuracy by construction.

Table 5: One-day-ahead benchmark on the pooled panel (test = 2,616 sequences), rupees per quintal. Mean over three seeds; RMSE dispersion in parentheses. Best in **bold**. Majority base rate for directional accuracy: 45.8%.

Model	RMSE↓	MAE↓	MAPE%↓	R^2↑	DA%↑
Proposed	**165.02** (0.22)	106.58	5.13	**0.6475**	59.62
Bi-LSTM	166.31 (0.43)	106.12	5.08	0.6420	59.53
DLinear	166.32 (0.40)	108.75	5.23	0.6419	59.31
LSTM	167.94 (0.84)	106.81	5.10	0.6349	59.57
Ensemble (mean)	169.21 (–)	109.41	5.22	0.6294	**60.17**
Ridge AR	170.81 (–)	121.19	5.71	0.6223	57.22
iTransformer-lite	180.77 (0.33)	110.94	5.34	0.5770	57.82
PatchTST-lite	183.09 (1.40)	113.53	5.46	0.5661	57.61
LightGBM	200.15 (–)	133.69	6.21	0.4814	56.31
Naive (last obs)	205.94 (–)	121.44	5.84	0.4510	10.89

5.3 Multi-horizon results and significance

The central result is in Table 6 and Figs. 3 and 4a. The proposed model leads at every horizon, but the Diebold–Mariano test tells a horizon-dependent story: at one day the improvement over the strongest baseline is not significant (DM = 1.39, $p = 0.17$); at seven days it is significant (DM = 2.23, $p = 0.026$); at fourteen days it is strongly significant (DM = 6.43, $p < 0.001$). The DM statistic—the honest summary of the advantage—grows monotonically with horizon (Fig. 4a).

A note on magnitude versus significance. It is worth being explicit about an apparent tension. The *mean*-RMSE reduction over the single strongest baseline is 0.78% at t+1, essentially nil (0.005%) at t+7, and 1.47% at t+14. At seven days the proposed model effectively ties DLinear in mean RMSE (191.20 vs. 191.21), yet the DM test still finds it significantly better. There is no contradiction: DM tests the *paired, per-observation* difference of squared errors across the whole panel (12); a small but *consistent* advantage, spread over thousands of sequences with low variance of the difference, is significant even when the mean gap rounds to zero. We therefore report the DM statistic rather than a percentage-reduction bar, whose near-zero height at t+7 would misleadingly suggest the model fails there. The pattern is intuitive. One-day changes are dominated by idiosyncratic noise that no model can resolve, so the field ties; as the horizon lengthens, structural signal—seasonality, the market factor, weather—takes over, and the components that encode structure begin to pay off. This is also the practically useful regime: a one-day mandi forecast is of little value, whereas a one-to-two-week view informs an actual selling decision.

Table 6: Proposed model versus the strongest baseline at each horizon (rupees per quintal; three-seed means). Diebold–Mariano compares squared errors (positive DM favours the proposed model); base rate is the majority-direction score at that horizon. All cells synchronised to the run.

Horizon	Model	RMSE	R^2	DA%	base DA%	DM	p
t+1	**Proposed**	**165.02**	**0.648**	59.62	45.8	1.39	0.166
	Bi-LSTM (best baseline)	166.31	0.642	59.53			
t+7	**Proposed**	**191.20**	**0.527**	60.47	46.8	2.23	0.026
	DLinear (best baseline)	191.21	0.527	60.44			
t+14	**Proposed**	**209.90**	**0.430**	62.96	46.5	6.43	< 0.001
	DLinear (best baseline)	213.04	0.412	63.55			

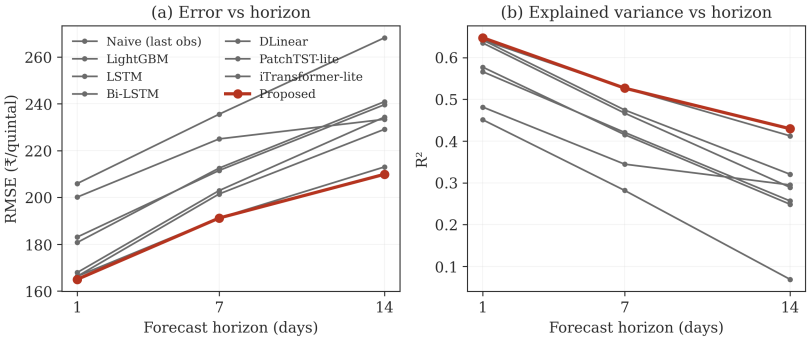
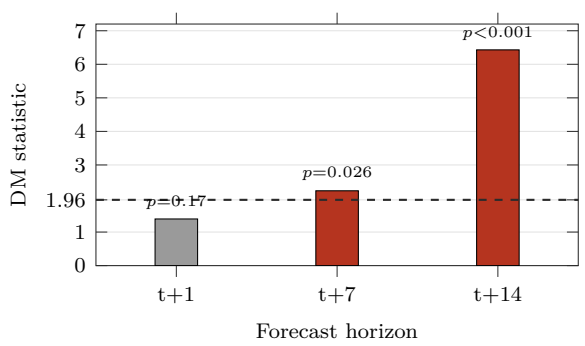
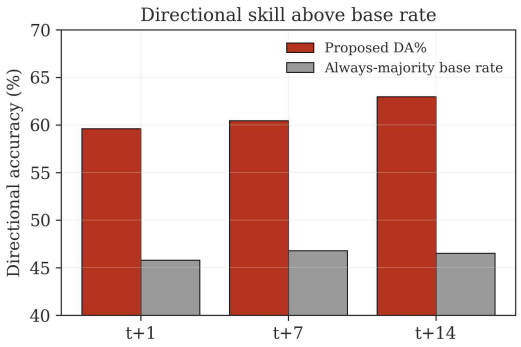


Figure 3: Error and explained variance by horizon. The proposed model (red) separates from the pack as the horizon lengthens, most clearly at fourteen days; at one and seven days it is close to DLinear.



(a) Diebold–Mariano statistic vs. strongest baseline (dashed line: 5% critical value, $|\mathrm{DM}| = 1.96$).



(b) Directional accuracy vs. the majority base rate.

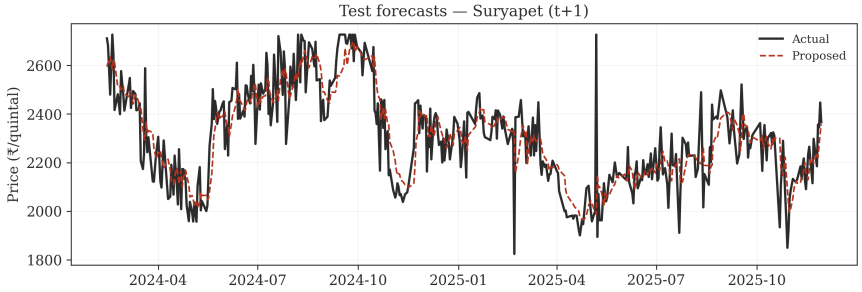
Figure 4: Statistical and practical significance of the forecast improvement by horizon. (a) The advantage over the strongest baseline is not significant at one day ahead (p=0.17) but becomes significant at seven (p=0.026) and fourteen (p<0.001) days; this replaces a percentage-RMSE-reduction chart, whose near-zero height at t+7 (where the model ties DLinear in mean RMSE while remaining significantly better in the paired test) misrepresented the result — see §5.3. (b) The proposed model clears the directional base rate by 12–16 points at every horizon, genuine skill rather than an artefact of a trending window.

5.4 Directional skill

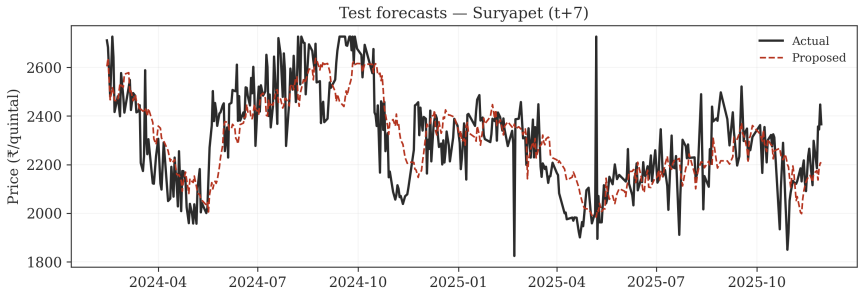
Fig. 4b contrasts directional accuracy with the majority base rate at each horizon. The proposed model clears the base rate by roughly 12–16 percentage points throughout, so the directional performance is genuine skill rather than an artefact of a trending window. This is, to our reading, the more decision-relevant summary of the model's value than a sub-1% RMSE margin.

5.5 Forecast trajectories

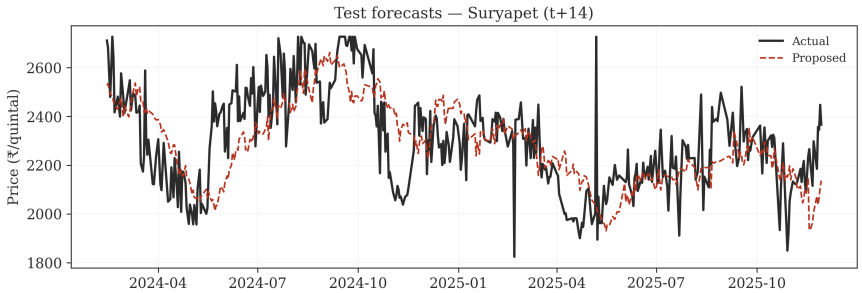
Fig. 5 plots predicted against actual prices on Suryapet, the most liquid market, over the 2024–2025 test window at all three horizons. At $t{+}1$ the forecast tracks the series tightly; at $t{+}7$ and $t{+}14$ it smooths through the daily noise while still following the medium-term level—consistent with the returns target capturing direction rather than memorising the last value.



(a) $t{+}1$



(b) $t{+}7$



(c) $t{+}14$

Figure 5: Predicted (red, dashed) versus actual (black) prices on the Suryapet market over the test window, at three horizons. Longer horizons track the level while filtering daily noise.

5.6 Component ablation

We remove one component at a time from the proposed model at one day ahead (Table 7, Fig. 6). The result is candid and important for readers who might assume the heavier machinery drives the gain: it does not. Removing the linear autoregressive path is by far the most costly change, raising RMSE from 165.0 to 174.5 (a +9.5 swing). Removing the STAR module, reversible normalisation or the inverted branch each moves RMSE by one to two rupees, while removing the patch branch or the endo/exo cross-attention is within noise at this horizon. We read this as evidence that at short horizons a strong linear core plus the returns formulation captures most of the learnable signal, and that the attention-based components earn their place mainly at the longer horizons where the significant improvements appear (§5.3). We report it as found rather than pruning the model to flatter the narrative.

Table 7: Component ablation at one day ahead (three-seed means). A positive ΔRMSE means removing the component hurts. Synchronised to the run.

Variant	RMSE	R^2	ΔRMSE vs full
Proposed (full)	165.02	0.6475	—
− patch branch	165.03	0.6475	+0.01
− endo/exo x-attn	165.14	0.6470	+0.12
− inverted branch	166.08	0.6430	+1.06
− RevIN	166.50	0.6411	+1.48
− STAR	167.07	0.6386	+2.05
− linear AR path	174.54	0.6057	+9.52

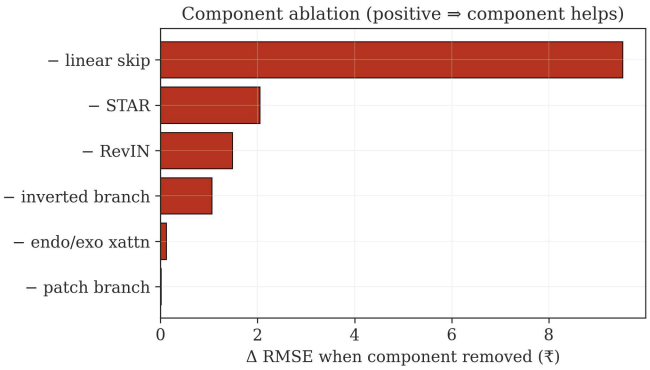


Figure 6: Component ablation. The linear autoregressive path dominates; the attention-based components contribute little at one day ahead, consistent with their role emerging at longer horizons.

5.7 Contribution of weather

Comparing the full model with a price-only variant, weather lowers one-day RMSE only marginally for the deep model (well under one rupee) while helping the weaker tree baseline appreciably. The reading is that a capable sequence model already extracts most of the short-horizon signal from price and arrivals, and that weather's contribution is larger at longer horizons and for weaker learners. We therefore present weather as a modest, honestly-bounded covariate rather than a decisive one.

6 Discussion

The experiments support a specific and limited claim: on pooled Telangana paddy mandis, a returns-based panel model with a cross-sectional factor forecasts one-to-two-week prices with a small but statistically significant edge over strong modern baselines, and with directional accuracy well above the majority base rate. We are equally clear about what the evidence does not support. At one day ahead the model is not significantly better than a Bi-LSTM or DLinear; the absolute RMSE margins are of order one percent; and the ablation shows the heavier neural components contribute little at short horizons. A paper that claimed a decisive architectural victory here would not survive scrutiny, and we do not make one.

Two design choices carry most of the practical value and generalise beyond this dataset. The returns formulation is what turns a sub-chance directional baseline into a useful one, and it costs nothing to adopt. Panel pooling with a cross-sectional factor supplies the common-shock signal that single-market models miss and gives deep models enough data to be worth using. Both are portable to other commodities and regions.

6.1 Limitations

Several constraints bound the result. *Scope:* one crop grade in one state; generalisation to other commodities or regions is untested here. *Single test window:* we evaluate on one chronological split (2024–2025); a rolling-origin evaluation across several windows would strengthen the significance claims and is the most valuable next step. *Magnitude:* the RMSE advantage is significant but small; the directional result is the stronger practical finding. *Weather:* NASA POWER is reanalysis at regional resolution, and the cross-sectional factor assumes other markets' same-day prices are observable at prediction time, which suits a next-day operational setting but would require a one-day lag if that assumption fails. *Point forecasts only:* the model returns a single number, not a calibrated interval; probabilistic forecasting (quantile or conformal) would give a risk range that point-forecast baselines cannot produce, and is a natural extension.

7 Conclusion and future work

We presented a horizon-aware panel forecaster for Telangana paddy mandi prices that pools eleven markets, forecasts the h-step change rather than the level, adds a cross-sectional market factor and real meteorological covariates, and combines current forecasting components with an explicit linear autoregressive path. Against DLinear, PatchTST, iTransformer, LightGBM and recurrent baselines—all trained identically and tested in rupee space—the model is best at every horizon and significantly best at seven and fourteen days, with directional accuracy near 60% against a 45.8% base rate. A component ablation attributes the gain principally to the returns formulation and the linear path, and we report the horizon where the improvement is not significant as openly as those where it is. Future work will add rolling-origin evaluation, extend to further commodities and states, and replace the point forecast with calibrated predictive intervals.

Compliance with Ethical Standards

Ethics approval

Not Applicable

Consent to participate

Not Applicable

Research involving Human Participants and/or Animals

No

Consent for publication

Not Applicable

Funding

The authors declare that no funds, grants, or other support were received during the study.

Declaration of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

M.T.S. contributed to data collection, investigation, model implementation, model validation, and interpretation of results. A.T. contributed to conceptualization, methodology, interpretation of results, and supervision. Both authors reviewed and approved the final manuscript.

Data and code availability

The dataset and processing code used in this study are available at `https://github.com/Madhusikta-tv/AGRI-DATA`.

Clinical trial number

Not Applicable.

References

Agricultural Marketing and Co-operation Department, Government of Telangana. Daily prices of market yard commodities in telangana. Telangana Open Data Portal, `https://data.telangana.gov.in/search?theme=Agriculture`, 2026. Dataset covers 01-01-2014 to 30-06-2026; accessed 2026.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5th edition, 2015.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan. Softs: Efficient multivariate time series forecasting with series-core fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

NASA Langley Research Center. NASA POWER: Prediction of worldwide energy resources. NASA POWER Data Access Viewer, `https://power.larc.nasa.gov`, 2026. Daily agroclimatology point data; accessed 2026.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Yunzhong Qiu, Haoran Zhang, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.