

Beyond Persistence: Understanding Forecastability of Weekly Potato Prices Across Multiple Market Centres

Muktinath Vishwakarma

dt23cse014@students.vnit.ac.in

Visvesvaraya National Institute of Technology

Manish Kurhekar

Visvesvaraya National Institute of Technology

Article

Keywords: potato price forecasting, forecastability, persistence, rolling-origin evaluation, model complexity, multi-horizon forecasting

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11283618/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Beyond Persistence: Understanding Forecastability of Weekly Potato Prices Across Multiple Market Centres

Muktinath Vishwakarma^{1*} and Manish Kurhekar¹

^{1*}Department of Computer Science and Engineering, Visvesvaraya
National Institute of Technology Nagpur, South Ambajhari Road,
Nagpur, 440010, Maharashtra, India.

*Corresponding author(s). E-mail(s): dt23cse014@students.vnit.ac.in;
Contributing authors: manishkurhekar@cse.vnit.ac.in;

Abstract

Potato-price forecasting can support production, storage, and marketing decisions, yet it remains unclear when complex forecasting models provide useful improvement over simple price persistence. This study evaluates weekly potato prices from 17 Indian markets during 2005–2015 using 12 forecasting methods, including statistical, machine-learning, and deep-learning models. Forecasts were assessed at 1-, 4-, 8-, and 13-week horizons using a rolling-origin design, with Naive-1, which carries the latest observed price forward, as the primary benchmark. Naive-1 achieved the best average performance at one week, indicating strong short-term persistence. As the forecast horizon increased, improvements over persistence became more common, reaching 16 of 17 markets at 13 weeks. Exponential Smoothing (ETS) achieved the best average performance at 4, 8, and 13 weeks and significantly outperformed Naive-1 across markets at 13 weeks. Machine-learning and deep-learning models did not provide consistent gains. Cross-market analysis showed that persistence alone did not explain forecastability, whereas seasonal strength was increasingly associated with forecast skill at longer horizons. These findings show that potato-price forecastability depends more on forecast horizon and temporal structure than on model complexity alone.

Keywords: potato price forecasting, forecastability, persistence, rolling-origin evaluation, model complexity, multi-horizon forecasting

1 Introduction

Potato prices are important to both producers and consumers because changes in market prices directly affect farm income, marketing decisions, storage choices, and the cost of food. In India, potato markets are particularly exposed to price variation because production is seasonal, the crop is perishable, and the quantity arriving in markets changes substantially over the year. Previous studies of major Indian potato markets have reported pronounced seasonal movements and considerable variability in both market arrivals and prices (Rana et al. 2003; Saha et al. 2020). Such fluctuations make it difficult for farmers, traders, storage operators, and other market participants to decide when to sell, store, procure, or distribute the crop. Reliable information about likely future prices can therefore support better short- and medium-term decision-making.

Because of this practical importance, agricultural price forecasting has received considerable research attention. Existing methods range from traditional statistical models to machine learning (ML) and deep learning (DL) approaches, with recent reviews showing a clear movement toward increasingly flexible and hybrid forecasting systems (Sun et al. 2023). Potato-price studies in India have followed the same progression. Statistical models such as the Autoregressive Integrated Moving Average (ARIMA) model have been compared with neural-network approaches (Kumar et al. 2022), while more recent studies have investigated architectures such as Long Short-Term Memory (LSTM) networks and Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (N-BEATS) (Praveenkumar et al. 2025; Nayak et al. 2025). These studies demonstrate that sophisticated models can capture useful patterns in agricultural price data, but they also place considerable emphasis on identifying the model with the smallest forecasting error.

Selecting the model with the lowest error, however, does not by itself establish that the price series contains substantial information that is useful for forecasting. Weekly prices can be highly *persistent*, meaning that the price observed in one week often remains close to the price observed in the following week. When persistence is strong, a very simple forecast that carries the most recently observed price into the future can already be difficult to improve upon. This rule is commonly called the *Naive-1* forecast. Simple benchmarks of this type play an important role in forecasting research because a more complicated model provides practical forecasting value only when it performs better than an appropriate simple alternative (Makridakis et al. 2018, 2020; Petropoulos et al. 2022). The relevant question is therefore not only which model produces the smallest error, but whether additional modelling complexity extracts useful predictive information beyond what is already contained in the latest observed price.

This distinction becomes more important when forecasts are required at different distances into the future. The *forecast horizon* refers to how far ahead a prediction is made; for example, a one-week horizon predicts the following week, whereas a thirteen-week horizon predicts approximately one quarter ahead. A recent price may contain substantial information about the next week but progressively less information about prices several weeks later. At longer horizons, recurring structures such as seasonal patterns may become relatively more useful than short-term persistence. Forecast

performance may also differ across markets because individual price series can contain different amounts of persistence, seasonality, trend, and irregular variation. Evaluating only one horizon or reporting only an overall model ranking can therefore hide an important part of the forecasting problem: *when*, *where*, and *why* does a model improve upon a simple benchmark?

Answering this question matters for both forecasting research and agricultural applications. A complex model that reduces error only marginally, inconsistently, or only in a small number of markets may not provide the same evidence of predictability as a method that produces stable improvements across repeated forecasts. Forecast evaluation should consequently be based on observations that were not used to fit the model and should examine performance repeatedly through time. Rolling-origin evaluation, in which the forecast starting point is progressively moved forward and the models are evaluated at multiple historical points, provides a principled way to perform such comparisons (Tashman 2000). Moreover, numerical improvements should be distinguished from statistically reliable improvements, because a lower observed error can arise from ordinary sampling variation. These considerations shift the objective from simply finding a winning model toward understanding the *forecastability* of the underlying market, which in this study refers to the extent to which historical price information supports out-of-sample forecasts that improve upon the Naive-1 benchmark.

This study investigates forecastability using weekly potato prices from 17 market centres in India over the period 2005–2015. Twelve forecasting methods, ranging from simple benchmarks and statistical models to machine-learning and deep-learning approaches, are evaluated at forecast horizons of 1, 4, 8, and 13 weeks. All methods are compared with Naive-1 under a common rolling-origin experimental design so that improvement is measured relative to the same reference forecast. The analysis then moves beyond model comparison by examining whether differences in forecasting performance across markets are associated with measurable characteristics of the price series, including persistence, trend, seasonal structure, price variability, and spectral structure. Statistical tests and temporal robustness analyses are additionally used to distinguish descriptive improvements from evidence that is sufficiently consistent to support broader conclusions.

The contribution of this study is threefold. First, it reframes potato-price forecasting as a question of *forecastability beyond persistence* rather than solely as a competition among model architectures. Second, it evaluates whether the usefulness of additional modelling changes with the forecast horizon and across individual potato markets, thereby identifying situations in which simple and more structured forecasting approaches are informative. Third, it connects observed forecasting gains with characteristics of the underlying price series, providing an explanation for why some markets offer greater opportunities for prediction than others. In this way, the study seeks not only to determine which forecasting approaches perform well, but also to clarify the conditions under which additional forecasting complexity is justified.

The remainder of the paper is organised as follows. Section 2 reviews previous research on agricultural and potato-price forecasting and positions the present study

within the existing literature. Section 3 defines the forecasting problem, the benchmark, forecast horizons, and the meaning of forecastability used in this study. Section 4 describes the dataset, forecasting methods, evaluation protocol, statistical tests, and robustness procedures. Section 5 presents and analyses the forecasting results, including horizon-dependent performance and differences in forecastability across markets. Section 6 discusses the interpretation, practical implications, and limitations of the findings. Section 7 highlights the main research lessons and practical takeaways from the study. Finally, Section 8 summarises the main conclusions and directions for future potato-price forecasting research.

2 Related Work

2.1 Approaches to agricultural and potato-price forecasting

Agricultural price forecasting has progressed from classical statistical models to machine-learning, deep-learning, hybrid, and information-enriched approaches (Wang et al. 2020; Sun et al. 2023). Potato-price forecasting follows the same development, with recent studies increasingly comparing flexible nonlinear models with conventional statistical methods.

Several studies report gains from neural and deep-learning approaches. Kumar et al. (2025b) found Long Short-Term Memory (LSTM) to outperform statistical alternatives for Agra potato prices, while Praveenkumar et al. (2025) reported the strongest performance from an attention-based LSTM. Similarly, Nayak et al. (2025) found Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (N-BEATS) to outperform Convolutional Neural Network (CNN), LSTM, and Gated Recurrent Unit (GRU) models. These studies demonstrate the potential of flexible temporal models, although they mainly focus on identifying the method with the lowest forecasting error.

Hybrid and information-enriched approaches provide another route to improved prediction. Decomposition-based methods have shown benefits by separating complex price movements before forecasting (Shankar et al. 2025; Kumar et al. 2025a), while related-market prices and weather information have improved forecasts by adding information not contained in the target price series alone (Jaiswal et al. 2025; Nayak et al. 2024). Automated machine-learning approaches have also achieved strong performance through model selection and ensemble construction (Kumari et al. 2025).

These studies establish that sophisticated models can improve potato-price forecasts under particular settings. However, most emphasise predictive accuracy within a market rather than whether the improvement exceeds a strong persistence benchmark, how that improvement changes with forecast horizon, or why some markets are more forecastable than others. Table 1 summarises the main differences between representative studies and the present work.

2.2 Research gap

The comparison in Table 1 highlights a distinction between *model accuracy* and *forecastability*. Identifying the model with the smallest error does not necessarily show that

Table 1 Comparison of representative potato-price forecasting studies with the present study.

Study	Data and approach	Main finding	Gap relative to present study
Shankar et al. (2025)	Multiple northern Indian markets; statistical, neural, and EEMD-based hybrids.	Decomposition-based models performed strongly, with the best method varying by market.	Does not examine improvement over a common persistence benchmark across horizons.
Kumar et al. (2025b)	Weekly Agra prices (2003–2023); ARIMA, SARIMA, ANN, and LSTM.	LSTM achieved the lowest forecasting error.	Single-market analysis without cross-market forecastability assessment.
Kumari et al. (2025)	Daily Agra prices (2006–2023); AutoML, boosting, and stacked ensembles.	A stacked ensemble achieved the best performance.	Focuses on predictive optimisation rather than improvement over persistence.
Praveenkumar et al. (2025)	Weekly Indian prices (2006–2023); ARIMA, RNN, GRU, LSTM, Bi-LSTM, and AM-LSTM.	AM-LSTM achieved the strongest performance.	Focuses on model accuracy rather than horizon- and market-dependent forecastability.
Nayak et al. (2025)	Weekly Farrukhabad prices (2003–2023); CNN, LSTM, GRU, and N-BEATS.	N-BEATS achieved the lowest reported errors.	Single-market study without analysis of cross-market forecastability.
Kumar et al. (2025a)	Monthly prices from Agra, Delhi, and Bengaluru; SSA-TDNN and statistical models.	Decomposition improved forecasting performance.	Evaluates a hybrid method rather than broad model classes against persistence.
Jaiswal et al. (2025)	Delhi and Guwahati prices with related-market information; NARX.	Cross-market information improved forecast accuracy.	Uses external market information, unlike the present univariate price-only setting.
Present study	Weekly prices from 17 Indian markets; 12 methods evaluated at 1-, 4-, 8-, and 13-week horizons.	Evaluates improvement beyond Naive-1 and how forecastability changes across horizons and markets.	Relates forecast gains to market-level temporal characteristics to explain when and why improvement occurs.

it extracts substantial information beyond a simple persistence forecast. Benchmark-based and out-of-sample evaluation are therefore important for determining whether additional modelling provides genuine predictive value (Tashman 2000; Makridakis et al. 2018; Petropoulos et al. 2022).

A second gap concerns forecast horizon and cross-market variation. A model that is useful several weeks ahead may provide little improvement for near-term prediction, and performance may also differ across markets with different temporal structures. Existing studies show that model performance varies among locations (Shankar et al. 2025; Kumar et al. 2025a), but comparatively little attention has been given to explaining these differences through properties such as persistence, seasonality, trend, and variability.

The present study addresses these gaps by evaluating 12 forecasting methods across 17 markets and four forecast horizons under a common rolling-origin protocol. Performance is measured relative to the same Naive-1 benchmark, statistical testing distinguishes numerical from reliable improvements, and market-level characteristics are examined to explain differences in forecast skill. The focus therefore shifts from

identifying only the “best model” to understanding *when, where, and why* forecasting beyond persistence becomes possible.

3 Problem Definition and Forecasting Framework

3.1 Forecasting task and horizons

Let $P_{i,t}$ denote the weekly potato price at market i in week t , where $i = 1, \dots, 17$. Each market is treated as a separate univariate time series, so forecasts use only that market’s historical prices. External information such as weather, arrivals, storage, or neighbouring-market prices is not included.

At forecast origin t , model m predicts the price h weeks ahead using only information available up to week t :

$$\hat{P}_{i,t+h|t}^{(m)} = f_m(P_{i,1}, P_{i,2}, \dots, P_{i,t}). \quad (1)$$

The study evaluates four forecast horizons,

$$\mathcal{H} = \{1, 4, 8, 13\} \text{ weeks}, \quad (2)$$

representing approximately one week, one month, two months, and one quarter ahead. Evaluating multiple horizons is important because the usefulness of recent prices may decrease as the prediction target moves further into the future.

3.2 Naive-1 benchmark

The primary benchmark is the *Naive-1* forecast, which assumes that the most recently observed price remains unchanged:

$$\hat{P}_{i,t+h|t}^{(\text{Naive})} = P_{i,t}, \quad h \in \mathcal{H}. \quad (3)$$

Naive-1 is a strong benchmark when prices are persistent, because the latest price already contains substantial short-term information. A more complex model therefore provides additional forecasting value only if it improves upon this simple reference (Makridakis et al. 2018; Petropoulos et al. 2022).

3.3 Forecast skill and forecastability

For model m , market i , and horizon h , Mean Absolute Error (MAE) is

$$\text{MAE}_{i,h,m} = \frac{1}{|\mathcal{O}_h|} \sum_{t \in \mathcal{O}_h} \left| P_{i,t+h} - \hat{P}_{i,t+h|t}^{(m)} \right|, \quad (4)$$

where $\mathcal{O}_h$ denotes the evaluation origins available at horizon h .

Forecast skill relative to Naive-1 is defined as

$$S_{i,h,m} = 1 - \frac{\text{MAE}_{i,h,m}}{\text{MAE}_{i,h,\text{Naive}}}. \quad (5)$$

A positive value indicates lower MAE than Naive-1, zero indicates equal performance, and a negative value indicates worse performance. For example, $S = 0.10$ corresponds to a 10% reduction in MAE relative to the benchmark.

In this study, *forecastability beyond persistence* refers to the extent to which historical prices allow forecasting models to improve upon Naive-1 under out-of-sample evaluation. For each market and horizon, the best observed skill is

$$S_{i,h}^* = \max_{m \neq \text{Naive}} S_{i,h,m}. \quad (6)$$

A positive $S_{i,h}^*$ means that at least one evaluated model improves upon persistence, whereas a non-positive value means that none does. Because $S_{i,h}^*$ depends on the models and information considered in this study, it should not be interpreted as the maximum forecastability theoretically achievable.

A positive skill score represents a numerical improvement and does not automatically imply statistical significance. Statistical testing is therefore performed separately, as described in Section 4.

3.4 Explaining cross-market forecastability

The framework also examines why forecastability differs across markets. Five characteristics are considered: short-term persistence, trend strength, seasonal strength, week-to-week price variability, and spectral structure. Their formal definitions are provided in Section 4.

These characteristics may affect forecasting in different ways. Strong persistence can make Naive-1 difficult to beat, while recurring seasonal structure may become more useful at longer horizons when the latest observed price is less informative. The analysis therefore asks two connected questions: whether forecasting models improve upon persistence, and whether differences in that improvement can be associated with measurable properties of the underlying price series.

4 Experimental Setup

4.1 Dataset and evaluation design

The experiments use weekly potato prices from 17 Indian market centres. Each series contains 539 observations from 3 January 2005 to 27 April 2015, expressed in Indian Rupees per kilogram (INR/kg), with no missing observations. Table 2 summarises the price distributions. The dataset is used as a historical forecasting panel rather than for current operational price prediction.

A chronological 70:30 split is used to preserve the direction of time. The first 377 observations form the development region and the remaining 162 observations form the held-out evaluation region. Model selection and hyperparameter tuning are performed only on the development data, following standard out-of-sample forecasting practice (Tashman 2000; Petropoulos et al. 2022).

Forecast accuracy is assessed using an expanding rolling-origin design. At each origin, only observations available up to that week are used for fitting, preprocessing, and prediction. Forecasts are evaluated at 1-, 4-, 8-, and 13-week horizons. To ensure

Table 2 Descriptive characteristics of the weekly potato-price series. All centres contain 539 observations.

Centre	Mean	SD	Minimum	Maximum
Agartala	13.83	5.67	5.00	33.00
Ahmedabad	12.40	4.98	5.80	26.00
Bengaluru	15.02	5.39	6.00	32.25
Bhopal	10.21	5.23	4.00	30.00
Bhubaneshwar	10.16	4.82	4.00	35.00
Chennai	14.50	5.49	6.00	35.67
Delhi	13.64	6.30	4.00	37.00
Guwahati	10.89	5.08	5.00	28.00
Hyderabad	14.84	6.03	8.00	36.00
Jaipur	9.27	4.88	4.00	25.60
Kolkata	9.20	4.34	3.00	22.00
Lucknow	10.82	5.40	3.00	30.00
Mumbai	15.88	6.40	7.00	32.00
Patna	9.95	4.49	4.00	26.20
Shillong	11.93	5.36	5.00	27.00
Shimla	13.77	6.23	5.00	36.00
T.Puram	20.41	7.41	6.96	43.00

that all horizons are compared on identical dates, 149 common forecast origins are retained per market. Neural models are refitted every four origins for computational feasibility; no future observation is used in producing an earlier forecast.

4.2 Forecasting models

Twelve methods are compared under the same evaluation protocol: three simple benchmarks, two statistical models, two machine-learning models, and five deep-learning architectures. Their principal configurations are summarised in Table 3.

The simple benchmarks are Naive-1, Seasonal Naive, and Drift. Naive-1 carries the latest observed price forward and serves as the primary persistence benchmark. Seasonal Naive uses the observation from the corresponding position in the previous 52-week cycle, while Drift extrapolates the average historical change.

The statistical models are Exponential Smoothing (ETS) and Autoregressive Integrated Moving Average (ARIMA). ETS selects combinations of level, trend, and seasonal components using the corrected Akaike Information Criterion (AICc) (Hyndman and Khandakar 2008). ARIMA, including seasonal components with a 52-week period, is selected from bounded candidate orders using AICc (Box et al. 2015).

Ridge regression provides a regularised linear machine-learning benchmark (Hoerl and Kennard 1970), while Extreme Gradient Boosting (XGBoost) provides a non-linear tree-based alternative (Chen and Guestrin 2016). The deep-learning group contains Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (N-BEATS)

(Oreshkin et al. 2020), Patch Time-Series Transformer (PatchTST) (Nie et al. 2023), inverted Transformer (iTransformer) (Liu et al. 2024), and Temporal Convolutional Network (TCN) (Bai et al. 2018). The iTransformer is evaluated in a univariate setting and therefore does not use the cross-variable attention mechanism of its original multivariate formulation.

Table 3 Forecasting models and principal configuration choices.

Model	Family	Configuration or selection rule
Naive-1	Benchmark	Latest observed price carried forward.
Seasonal Naive	Benchmark	Corresponding observation from the previous 52-week cycle.
Drift	Benchmark	Extrapolates the average historical change.
ETS	Statistical	Error, trend, and seasonal components selected using AICc.
ARIMA	Statistical	Bounded order selection using AICc; seasonal period = 52.
Ridge	Machine learning	Lookback $L \in \{13, 26, 52\}$; regularisation selected internally.
XGBoost	Machine learning	Lookback $L \in \{13, 26, 52\}$; 50 Optuna trials.
LSTM	Deep learning	Selected lookback; one layer, 64 hidden units, dropout = 0.1.
N-BEATS	Deep learning	Selected lookback; two stacks, one block per stack, width = 64.
PatchTST	Deep learning	Lookback = 26; patch/stride = 8; embedding = 64; four heads; two layers.
iTransformer	Deep learning	Lookback = 26; embedding = 64; four heads; two layers.
TCN	Deep learning	Lookback = 26; three convolutional layers; 32 channels.

4.3 Input construction, tuning, and training

Machine-learning and deep-learning models use historical windows of

$$\mathcal{L} = \{13, 26, 52\} \quad (7)$$

weeks, representing approximately one quarter, one half-year, and one year of past prices. For a lookback L , the input sequence is

$$\mathbf{x}_t = [P_{t-L+1}, \dots, P_t], \quad (8)$$

and the models directly predict the following 13 weekly prices,

$$\mathbf{y}_t = [P_{t+1}, \dots, P_{t+13}]. \quad (9)$$

Only horizons 1, 4, 8, and 13 are retained for evaluation. Price inputs are standardised using the mean and standard deviation calculated from the available training data only, and forecasts are transformed back to the original price scale before evaluation.

Hyperparameter selection is performed entirely within the development region using average Mean Absolute Error (MAE) over the four forecast horizons as the tuning objective. Ridge considers 50 logarithmically spaced regularisation values between 10^{-4} and 10^4 . XGBoost is tuned using 50 Optuna trials (Akiba et al. 2019). LSTM and N-BEATS select their lookback from the development data, while PatchTST, iTransformer, and TCN use the fixed configurations in Table 3.

Neural models are trained using the Adaptive Moment Estimation (Adam) optimiser (Kingma and Ba 2015), with learning rate 10^{-3} , batch size 32, a maximum

of 200 epochs, and early-stopping patience of 20 epochs. Each neural architecture is evaluated using five random seeds,

$$\mathcal{S} = \{42, 43, 44, 45, 46\}, \quad (10)$$

and the resulting metrics are averaged for the primary comparison.

4.4 Forecast evaluation

Forecast accuracy is evaluated using MAE, Root Mean Squared Error (RMSE), Symmetric Mean Absolute Percentage Error (sMAPE), and Mean Absolute Scaled Error (MASE). MAE and forecast skill relative to Naive-1 were defined in Section 3. RMSE gives greater weight to large errors, while sMAPE expresses error relative to the magnitude of the observed and predicted prices.

MASE is the primary scale-independent metric for comparing performance across markets (Hyndman and Koehler 2006):

$$\text{MASE} = \frac{\frac{1}{N} \sum_{t=1}^N |y_t - \hat{y}_t|}{\frac{1}{T-1} \sum_{t=2}^T |y_t - y_{t-1}|}. \quad (11)$$

Lower values indicate better forecasts. The main measure of forecastability remains the skill relative to Naive-1 because it directly quantifies improvement over persistence.

4.5 Market characteristics and explanatory analysis

To explain why forecastability differs across markets, five characteristics are calculated from the 377 development observations: lag-1 autocorrelation, trend strength, seasonal strength, normalised week-to-week volatility, and spectral entropy. Stationarity is also examined using the Augmented Dickey–Fuller (ADF) test (Dickey and Fuller 1979) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (Kwiatkowski et al. 1992).

Trend and seasonal strength are obtained from Seasonal-Trend decomposition using LOESS (STL) with a 52-week seasonal period (Cleveland et al. 1990). For

$$P_t = T_t + S_t + R_t, \quad (12)$$

the corresponding strengths are

$$F_T = \max \left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)} \right), \quad (13)$$

and

$$F_S = \max \left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t + R_t)} \right). \quad (14)$$

Short-term volatility is measured as the standard deviation of weekly price changes and normalised by the market’s mean price. Spectral entropy summarises whether periodic variation is concentrated in a small number of frequencies or dispersed across many frequencies.

For each horizon, Spearman rank correlation is used to relate these five characteristics to the best observed forecast skill $S_{i,h}^*$. Uncertainty is summarised using

95% nonparametric bootstrap confidence intervals, and the five p -values within each horizon are corrected using Holm’s procedure (Holm 1979). These relationships are interpreted as associations rather than causal effects.

4.6 Statistical comparison of forecasting performance

At the individual-market level, each challenger is compared with Naive-1 using the Diebold–Mariano test (Diebold and Mariano 1995) with the Harvey–Leybourne–Newbold small-sample correction (Harvey et al. 1997). With 17 markets and 11 challengers, this produces 187 comparisons within each horizon. Their p -values are adjusted jointly using Holm’s procedure (Holm 1979).

Across markets, model rankings are compared using the Friedman test (Friedman 1937). When the overall test is significant, pairwise Wilcoxon signed-rank tests are performed (Wilcoxon 1945). The resulting 66 pairwise p -values at each horizon are again Holm-corrected. The centre-level tests therefore examine whether a challenger reliably beats Naive-1 within a particular market, whereas the across-market tests examine systematic differences among models over the collection of markets.

4.7 Forecastability classification and robustness

Each market–horizon combination is assigned to one of three evidence classes. Class A (*Persistence-dominated*) indicates that no challenger achieves positive skill over Naive-1. Class B (*Weakly improvable*) indicates positive numerical skill without statistical significance after multiplicity correction. Class C (*Forecastable beyond persistence*) requires both positive skill and a statistically significant improvement.

Two robustness checks are performed. First, temporal stability is examined by dividing the 149 evaluation origins into early, middle, and late periods and recomputing best skill at the one-week horizon. Because the shorter periods are insufficient for reliable centre-level significance testing, this analysis is descriptive. Skill is divided into three bands: A for $S^* \leq 0$, B for $0 < S^* \leq 0.05$, and C for $S^* > 0.05$.

Second, neural-model stability is evaluated across the five random seeds. For each neural architecture and horizon, the mean and standard deviation of MASE are calculated. This determines whether the main neural-model conclusions depend strongly on random initialisation.

Together, these procedures ensure that all forecasting methods are evaluated using the same markets, forecast horizons, held-out origins, and persistence benchmark, while separately assessing predictive accuracy, statistical evidence, cross-market structure, and robustness.

5 Results and Analysis

5.1 Temporal characteristics of the potato-price series

Before comparing forecasting models, it is important to understand the temporal structure of the price series that the models are attempting to predict. Figure 1 presents the weekly potato prices for all 17 market centres. Although the centres differ substantially in their average price levels and the magnitude of price fluctuations, prices generally

evolve gradually from one week to the next. This behaviour suggests that the most recently observed price contains substantial information about the near future and provides an initial explanation for why a persistence-based forecast may be difficult to improve upon.

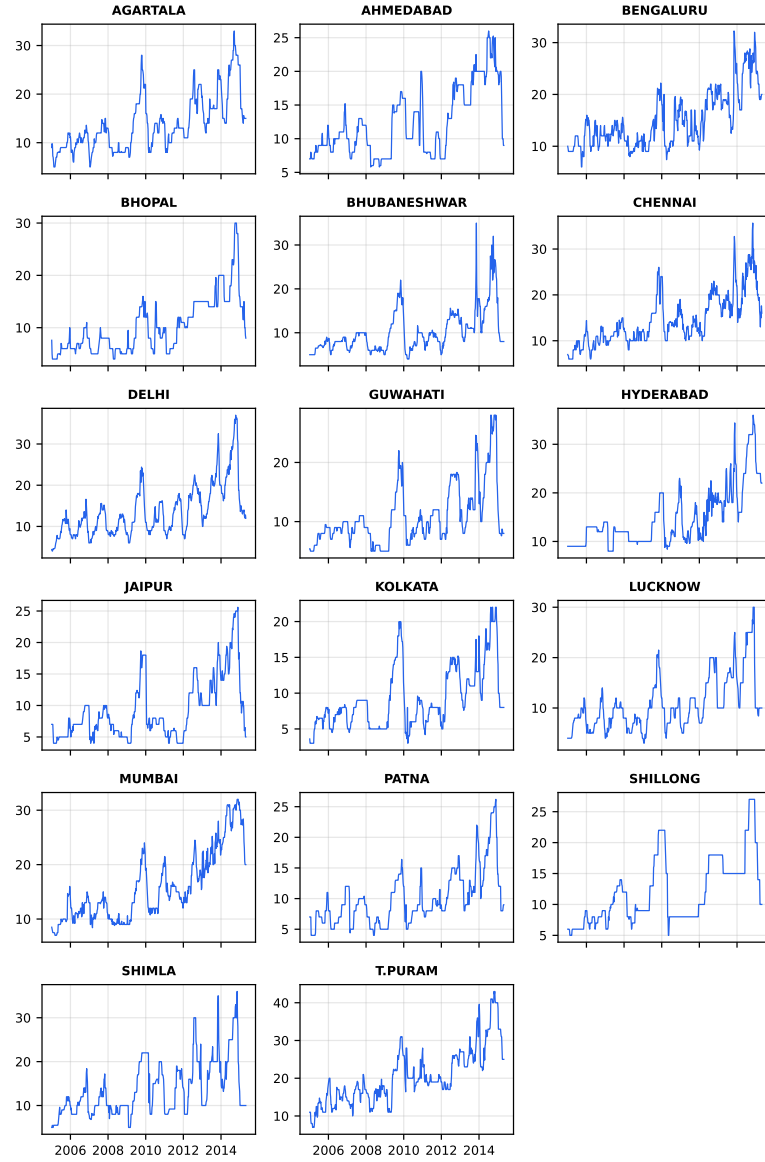


Fig. 1 Weekly potato prices across the 17 market centres from January 2005 to April 2015. The markets differ in price levels and temporal behaviour, but consecutive weekly prices generally remain strongly related.

The visual persistence in Figure 1 is confirmed by the lag-1 autocorrelation values reported in Table 4. Lag-1 autocorrelation measures how strongly the price in one week resembles the price observed in the immediately preceding week. Values range from 0.901 in Bengaluru to 0.986 in Shillong, showing that all 17 markets exhibit strong short-term persistence. Thus, the strength of Naive-1 as a benchmark is supported directly by the observed structure of the data.

The markets nevertheless differ considerably in their longer-term temporal structure. Seasonal strength ranges from essentially no detectable annual seasonal structure in Shillong ($F_S = 0.000$) and Hyderabad ($F_S = 0.032$) to a strong seasonal pattern in Delhi ($F_S = 0.735$). Lucknow ($F_S = 0.513$), Shimla ($F_S = 0.479$), and Patna ($F_S = 0.410$) also exhibit comparatively strong seasonality. These differences are important because two markets may both be highly persistent in the short term while containing very different amounts of recurring information that can potentially be exploited at longer forecast horizons.

Table 4 Time-series characteristics of the 17 market centres computed from the development period.

Centre	ACF(1)	Trend	Seasonal	Norm. volatility	Spectral entropy
Agartala	0.976	0.330	0.280	0.073	0.530
Ahmedabad	0.968	0.338	0.106	0.075	0.494
Bengaluru	0.901	0.230	0.142	0.107	0.685
Bhopal	0.957	0.385	0.303	0.104	0.514
Bhubaneswar	0.978	0.211	0.120	0.075	0.481
Chennai	0.952	0.599	0.388	0.088	0.633
Delhi	0.969	0.490	0.735	0.080	0.437
Guwahati	0.973	0.253	0.186	0.076	0.520
Hyderabad	0.956	0.159	0.032	0.072	0.615
Jaipur	0.972	0.249	0.125	0.099	0.560
Kolkata	0.976	0.039	0.166	0.091	0.479
Lucknow	0.964	0.384	0.513	0.096	0.541
Mumbai	0.972	0.399	0.093	0.063	0.575
Patna	0.958	0.402	0.410	0.094	0.551
Shillong	0.986	0.306	0.000	0.067	0.439
Shimla	0.961	0.454	0.479	0.102	0.558
T.Puram	0.954	0.619	0.252	0.081	0.626

The stationarity diagnostics provide further evidence that the 17 markets do not follow one common temporal process. The Augmented Dickey–Fuller (ADF) test rejects the unit-root null at the 5% level for 14 of the 17 centres, whereas the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test rejects trend stationarity for 8 centres. These results should not be interpreted as contradictory because the two tests begin from different null hypotheses. Instead, they indicate heterogeneous temporal behaviour across markets despite their uniformly high short-term persistence. This heterogeneity motivates evaluating forecasting performance separately by market and forecast horizon.

5.2 Overall forecasting performance

The first forecasting comparison asks which methods achieve the lowest prediction error when performance is averaged across all 17 markets. Table 5 reports Mean Absolute Scaled Error (MASE) at the four forecast horizons. Lower MASE indicates better forecasting performance, while the final column summarises the average model ranking across the evaluation.

Table 5 Overall forecasting performance across the 17 market centres. Values are mean MASE; lower values indicate better forecasting performance. Neural-model metrics are averaged across five random seeds within each centre before cross-centre aggregation.

Model	H = 1	H = 4	H = 8	H = 13	Mean rank
<i>Simple benchmarks</i>					
Naive-1	2.03	5.46	8.57	11.72	2.22
Seasonal Naive	13.37	13.33	13.29	13.38	9.06
Drift	2.06	5.52	8.67	11.89	3.16
<i>Statistical models</i>					
ETS	2.27	5.42	7.64	9.65	2.19
ARIMA	2.19	5.70	8.81	11.81	3.25
<i>Machine-learning models</i>					
Ridge	2.42	6.55	10.28	13.67	5.25
XGBoost	6.15	10.18	12.89	15.20	8.56
<i>Deep-learning models</i>					
LSTM	5.58	9.90	13.15	15.64	8.82
N-BEATS	4.86	9.88	13.67	16.49	8.76
PatchTST	9.32	12.00	14.46	16.14	10.46
iTransformer	7.19	10.52	13.15	15.18	8.51
TCN	3.55	9.06	13.03	16.25	7.75

At the one-week horizon, Naive-1 achieves the lowest average MASE of 2.03. Drift and Autoregressive Integrated Moving Average (ARIMA) are close, with MASE values of 2.06 and 2.19, respectively, while Exponential Smoothing (ETS) reaches 2.27. Thus, when the target is only one week ahead, attempting to estimate additional temporal structure does not improve average performance over simply retaining the most recent price. The result is consistent with the strong week-to-week persistence identified in Table 4.

The ranking changes as the forecast extends further into the future. ETS obtains the lowest mean MASE at 4, 8, and 13 weeks, with values of 5.42, 7.64, and 9.65, respectively. At 13 weeks, the mean MASE of ETS is approximately 17.7% lower than the corresponding Naive-1 value of 11.72. The increasing separation between the two methods suggests that the latest observed price becomes progressively less

informative as the target moves further away, creating greater opportunity for a model that represents level, trend, and seasonal structure.

Deep-learning models do not show the same improvement. Temporal Convolutional Network (TCN) is the strongest neural method at one week, with a mean MASE of 3.55, but its error remains considerably above the Naive-1 value of 2.03. Every evaluated neural architecture also has a higher mean MASE than Naive-1 at all four horizons. Greater representational capacity therefore does not automatically translate into greater forecast accuracy for these relatively short, univariate weekly price series.

5.3 Forecastability beyond the Naive-1 benchmark

Average forecasting error can conceal substantial differences among individual markets. To determine whether forecasting models actually extract useful information beyond persistence, Figure 2 shows the best observed forecast skill $S_{i,h}^*$ relative to Naive-1 for every market–horizon combination. Positive values mean that at least one competing method reduces Mean Absolute Error (MAE) relative to Naive-1, whereas non-positive values indicate that none of the evaluated challengers improves upon the persistence benchmark.

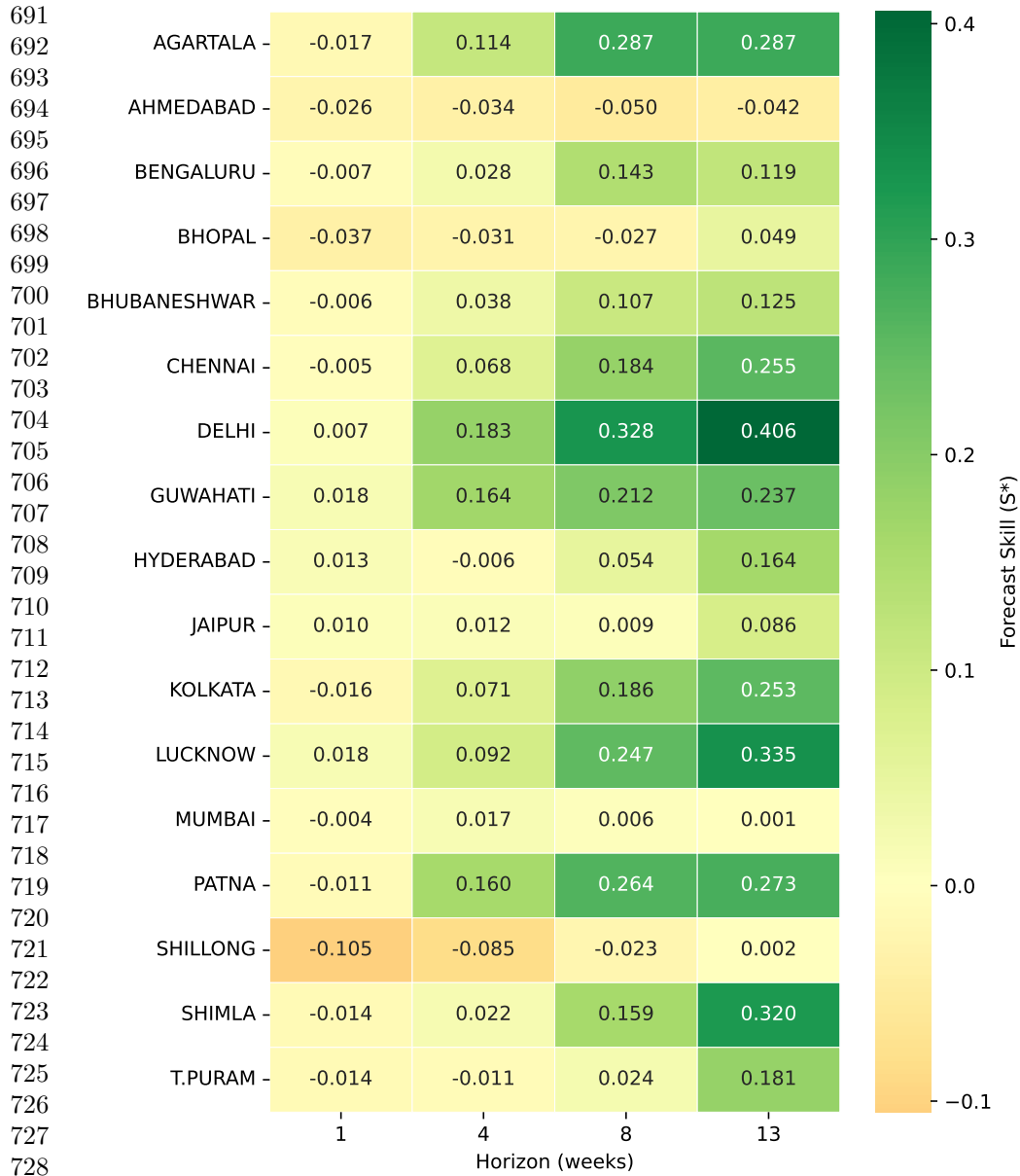


Fig. 2 Best observed forecast skill relative to Naive-1 across the 17 market centres and four forecast horizons. Positive values indicate lower MAE than Naive-1, while negative values indicate deterioration relative to the persistence benchmark.

Figure 2 reveals a clear change with forecast horizon. At one week, only 5 of the 17 centres (29%) exhibit positive numerical skill. These centres are Delhi, Guwahati,

Hyderabad, Jaipur, and Lucknow, and the largest gain is no more than 1.8%. In the remaining 12 centres, none of the evaluated alternatives improves upon Naive-1. The near absence of improvement at this horizon reinforces the conclusion that the current weekly price already contains most of the readily exploitable information about the following week.

The opportunity for improvement becomes much larger as the horizon increases. Positive skill is observed in 12 of 17 centres (71%) at four weeks, 14 of 17 (82%) at eight weeks, and 16 of 17 (94%) at thirteen weeks. Mean best skill similarly increases from -0.012 at one week to 0.180 at thirteen weeks. Forecastability beyond persistence is therefore strongly horizon-dependent in these data: the further the prediction extends into the future, the more frequently a structured model improves upon carrying the latest price forward.

Delhi provides the clearest example of this transition. Its best observed skill is only 0.007 at one week, representing less than a 1% reduction in MAE relative to Naive-1. The improvement increases to 0.183 at four weeks, 0.328 at eight weeks, and 0.406 at thirteen weeks. At the longest horizon, the best model therefore reduces MAE by 40.6% relative to persistence. This increasing advantage is especially notable because Delhi also has the strongest measured seasonal structure among the 17 markets.

Ahmedabad shows the opposite pattern. Its best skill remains negative at every horizon: -0.026 , -0.034 , -0.050 , and -0.042 at 1, 4, 8, and 13 weeks, respectively. For this market, none of the evaluated models improves upon simply retaining the most recent price. The contrast between Delhi and Ahmedabad demonstrates why market-level analysis is necessary: an overall model ranking cannot reveal whether the same forecasting opportunity exists in every market.

Table 6 provides representative centre-level results together with the adjusted Diebold–Mariano–Harvey–Leybourne–Newbold (DM-HLN) p -values. The table separates the magnitude of a numerical improvement from the strength of the statistical evidence supporting that improvement.

Despite the increasing numerical gains at longer horizons, none of the positive centre-specific improvements remains statistically significant after the within-horizon Holm correction. Among the best challengers for all 68 centre–horizon combinations, the smallest adjusted p -value is 0.256. Consequently, the individual market results provide evidence of numerical improvement, but not sufficient evidence to classify any individual market–horizon pair as statistically forecastable beyond persistence under the adopted multiple-testing procedure.

This distinction is important because numerical magnitude and statistical certainty answer different questions. A skill of 0.406 in Delhi at 13 weeks indicates a large observed reduction in MAE, but the centre-specific statistical test asks whether the improvement is sufficiently consistent across repeated forecast origins to remain convincing after accounting for the many simultaneous comparisons. The Results section therefore uses the term *positive skill* for observed improvements and reserves *statistically significant improvement* for cases supported by the corresponding corrected test.

Table 6 Representative centre-level forecastability results. Class A indicates that no challenger improves upon Naive-1; class B indicates positive numerical skill without statistical significance after Holm correction. No centre–horizon pair is classified as C.

Centre	Horizon	Best model	Best skill	Class
Agartala	1	Drift	−0.017	A
	4	ETS	0.114	B
	8	ETS	0.287	B
	13	ETS	0.287	B
Ahmedabad	1	Drift	−0.026	A
	4	Drift	−0.034	A
	8	Drift	−0.050	A
	13	Drift	−0.042	A
Delhi	1	ARIMA	0.007	B
	4	ETS	0.183	B
	8	ETS	0.328	B
	13	ETS	0.406	B
Guwahati	1	ETS	0.018	B
	4	ETS	0.164	B
	8	ETS	0.212	B
	13	ETS	0.237	B
Lucknow	1	ARIMA	0.018	B
	4	ETS	0.092	B
	8	ETS	0.247	B
	13	ETS	0.335	B
Shillong	1	Drift	−0.105	A
	4	Drift	−0.085	A
	8	ARIMA	−0.023	A
	13	ARIMA	0.002	B

5.4 Effect of forecast horizon

The preceding results show that forecast horizon fundamentally changes the forecasting problem. Positive-skill counts increase from 5/17 markets at one week to 12/17, 14/17, and 16/17 at 4, 8, and 13 weeks, respectively. Mean best skill simultaneously increases from −0.012 to 0.180. Thus, the ability of the evaluated models to improve upon Naive-1 becomes progressively greater as the target moves further into the future.

The intuition behind this pattern follows directly from the structure of the benchmark. At one week, Naive-1 uses a price observed only seven days before the target, and the strong persistence shown in Table 4 makes this a highly informative prediction. At thirteen weeks, the same forecast assumes that a price observed approximately three months earlier remains unchanged. The latest observed price is therefore a progressively weaker representation of future conditions, creating more opportunity for models that can use longer-term structure.

The transition is gradual rather than abrupt. Improvement is already visible in many markets at four weeks and becomes stronger at eight and thirteen weeks. This behaviour suggests that short-term persistence and longer-term temporal structure coexist in the price series: persistence dominates when the target is close, whereas other recurring patterns become increasingly useful as the forecasting distance expands.

5.5 Does greater model complexity improve forecasting?

The fact that forecastability increases with horizon does not necessarily imply that increasingly complex models are responsible for the improvement. Figure 3 provides a visual comparison of each model’s win/loss behaviour against Naive-1 across markets and horizons, while Table 7 reports the exact win/tie/loss counts.

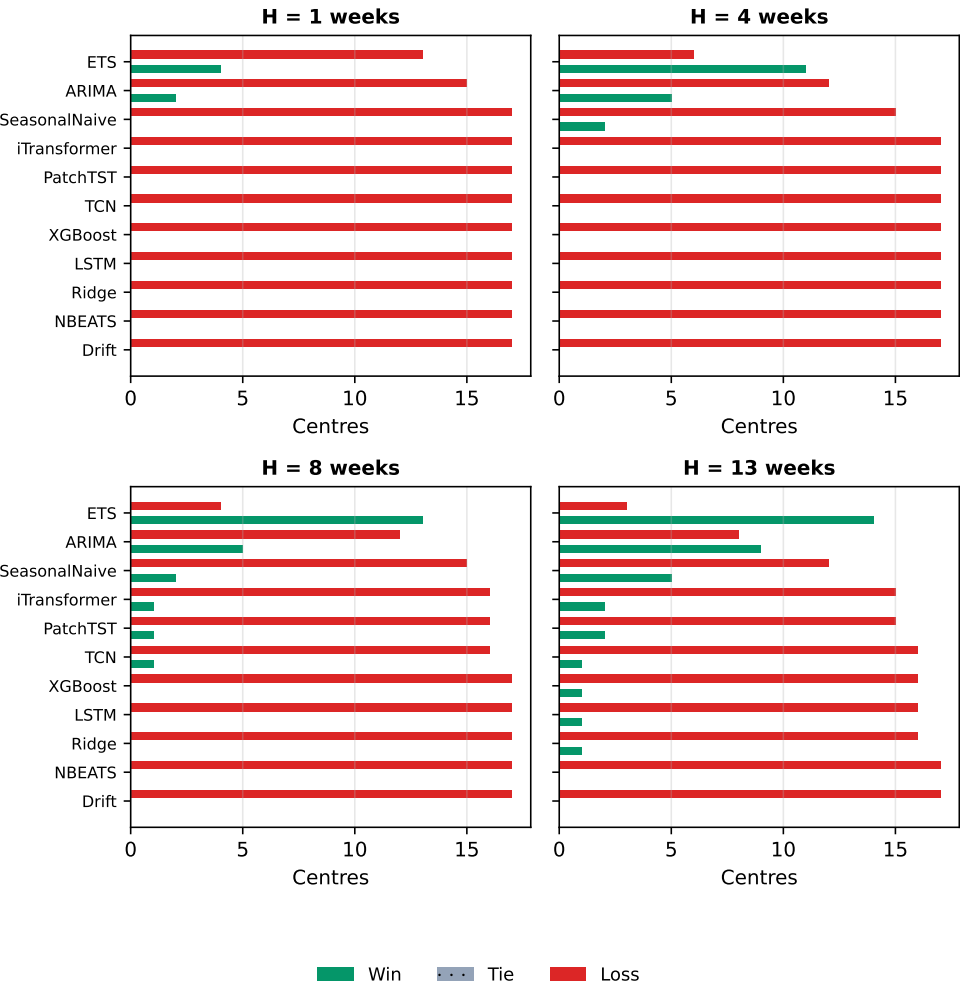


Fig. 3 Win/loss record of each forecasting model relative to Naive-1 across the 17 market centres and four forecast horizons. A win indicates lower centre-level MAE than Naive-1.

The visual pattern in Figure 3 shows that ETS is the only method whose advantage becomes widespread as the forecast horizon increases. At one week, ETS improves upon Naive-1 in only two markets. Its win count rises to 11 markets at four weeks,

13 at eight weeks, and 14 at thirteen weeks. The increasing prevalence of ETS wins closely mirrors the horizon-dependent increase in best observed forecast skill.

Table 7 Win/tie/loss counts relative to Naive-1 based on centre-level MAE. A win indicates that the corresponding model has lower MAE than Naive-1 for that centre.

Model	H = 1	H = 4	H = 8	H = 13
<i>Simple benchmarks</i>				
Seasonal Naive	0/0/17	0/0/17	1/0/16	5/0/12
Drift	0/0/17	2/0/15	2/0/15	0/0/17
<i>Statistical models</i>				
ETS	2/0/15	11/0/6	13/0/4	14/0/3
ARIMA	4/0/13	5/0/12	5/0/12	9/0/8
<i>Machine-learning models</i>				
Ridge	0/0/17	0/0/17	1/0/16	1/0/16
XGBoost	0/0/17	0/0/17	0/0/17	1/0/16
<i>Deep-learning models</i>				
LSTM	0/0/17	0/0/17	0/0/17	1/0/16
N-BEATS	0/0/17	0/0/17	0/0/17	0/0/17
PatchTST	0/0/17	0/0/17	0/0/17	2/0/15
iTransformer	0/0/17	0/0/17	1/0/16	2/0/15
TCN	0/0/17	0/0/17	0/0/17	1/0/16

ARIMA also becomes more competitive as the forecast horizon increases, reaching 9 wins at thirteen weeks. The increasing success of ETS and, to a lesser extent, ARIMA indicates that the useful information emerging at longer horizons can be captured by conventional statistical models. The improvement therefore appears to arise from exploitable temporal structure rather than from model complexity itself.

The Seasonal Naive benchmark provides an important qualification. If annual seasonality alone were sufficient, directly copying the corresponding price from the previous year should perform strongly. Instead, Seasonal Naive loses to Naive-1 in all 17 markets at one and four weeks and wins in only 5 markets at thirteen weeks. The longer-horizon success of ETS therefore appears to depend on its ability to adapt seasonal information together with changing level and trend rather than simply repeating the previous year's price.

Machine-learning models provide little systematic advantage. Ridge regression produces no wins at one or four weeks and only one win at each of eight and thirteen weeks. Extreme Gradient Boosting (XGBoost) records only one win, occurring at thirteen weeks. Thus, allowing nonlinear relationships or additional parameter flexibility does not by itself produce consistent improvement over the persistence benchmark.

The deep-learning results are even more striking. Long Short-Term Memory (LSTM), Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (N-BEATS), Patch Time-Series Transformer (PatchTST), inverted Transformer (iTransformer), and TCN all fail to produce systematic gains. N-BEATS does not beat Naive-1 in any of the 68 market-horizon comparisons, and the other neural models achieve only isolated wins. For the implementations and information set considered here, additional architectural complexity therefore does not translate into greater forecastability.

A plausible explanation is the combination of limited sample size and limited input information. Each market provides only 377 observations in the development region, and the models receive only the market’s own historical price sequence. Flexible models can learn complex relationships when such relationships are present and sufficiently represented in the data, but additional parameters cannot create predictive information that is absent from the inputs. The observed results are therefore consistent with the useful structure in these series being comparatively simple and better matched to parsimonious statistical models. This interpretation is plausible rather than causal, because the experiment was not designed to isolate the independent effect of training-sample size or model capacity.

5.6 Statistical comparison across markets

The centre-level tests evaluate whether a model reliably improves upon Naive-1 within one particular market. A complementary question is whether models differ systematically when performance is considered across all 17 markets. The Friedman test rejects equal model performance at every forecast horizon, with test statistics of 179.36, 166.62, 141.08, and 124.08 at 1, 4, 8, and 13 weeks, respectively. All corresponding p -values are below 10^{-20} , providing strong evidence that the forecasting methods do not have equivalent performance across markets.

Pairwise Wilcoxon comparisons reveal how the relationship between ETS and Naive-1 changes with horizon. At one week, ETS performs significantly worse than Naive-1 after Holm correction ($p_{\text{adj}} = 0.001$). At four weeks ($p_{\text{adj}} = 1.000$) and eight weeks ($p_{\text{adj}} = 0.205$), the difference is not statistically significant. At thirteen weeks, the direction reverses and ETS performs significantly better than Naive-1 ($p_{\text{adj}} = 0.022$). Statistical evidence therefore supports the same horizon-dependent transition observed in the descriptive forecasting results.

The across-market finding does not contradict the absence of significant centre-specific improvements. The two analyses ask different questions. The centre-level DM-HLN analysis requires a sufficiently consistent improvement within one market and additionally corrects 187 challenger comparisons within each horizon. The across-market Wilcoxon analysis instead evaluates whether ETS tends to outperform Naive-1 across the collection of 17 markets. The results therefore support a systematic cross-market advantage for ETS at thirteen weeks without establishing statistically significant improvement for any individual centre after the more conservative centre-level correction.

967 5.7 Why does forecastability differ across markets?

968 The preceding results show that forecasting gains vary substantially from one mar-
969 ket to another. The next analysis therefore examines whether these differences can be
970 associated with measurable properties of the historical price series. The central ques-
971 tion is no longer which model wins, but what type of market structure appears to
972 provide information beyond simple persistence.
973

974 5.7.1 Persistence alone does not explain cross-market 975 forecastability 976

977 Because Naive-1 is a persistence-based forecast, one natural possibility is that mar-
978 kets with weaker persistence should be easier for alternative models to improve
979 upon. Figure 4 examines this possibility by plotting lag-1 autocorrelation against best
980 observed forecast skill at the four horizons.
981

982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012

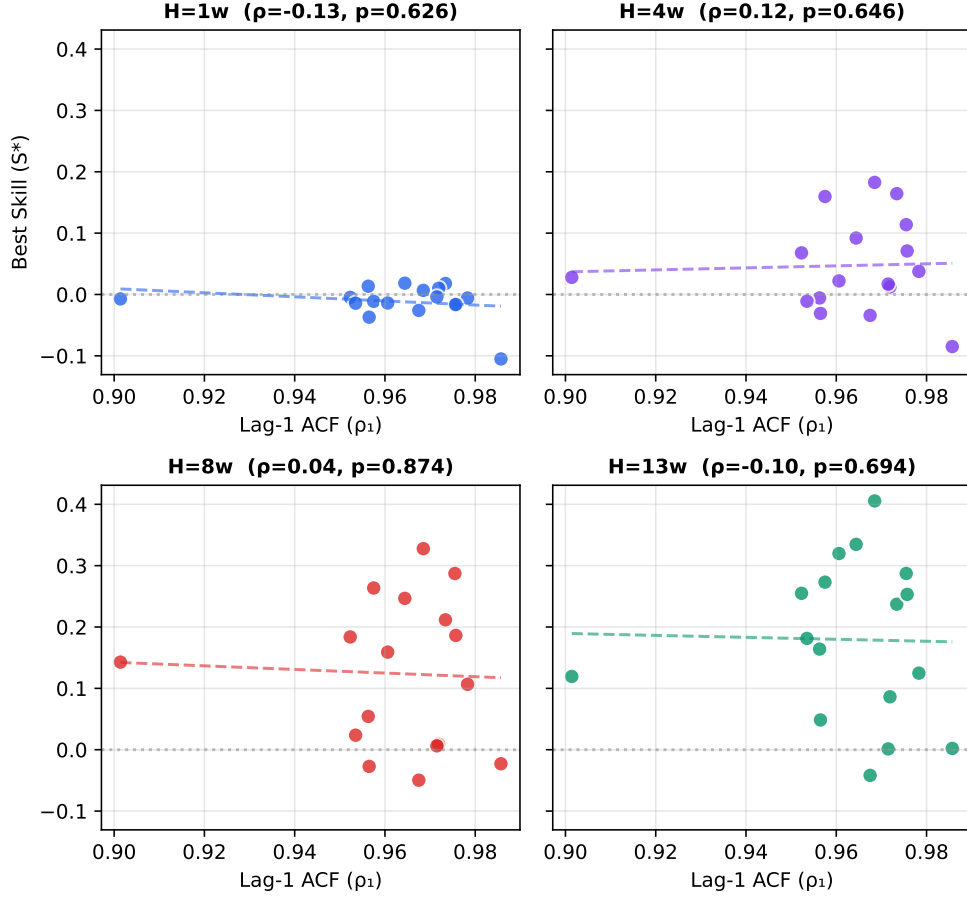


Fig. 4 Relationship between lag-1 autocorrelation and best observed forecast skill relative to Naive-1 across the four forecast horizons. Lag-1 autocorrelation represents short-term price persistence.

Figure 4 does not show a clear monotonic relationship between persistence and forecast skill. At one week, the Spearman correlation is $\rho_s = -0.128$, with a 95% bootstrap confidence interval of $[-0.586, 0.424]$ and Holm-adjusted $p = 1.000$. The correlations remain weak at four, eight, and thirteen weeks, with values of 0.120, 0.042, and -0.103 , respectively, and all corresponding adjusted p -values equal 1.000.

This result does not mean that persistence is unimportant. Persistence is extremely strong in almost every market and explains why Naive-1 is such a competitive general benchmark, particularly at short horizons. However, lag-1 autocorrelation varies only between 0.901 and 0.986. Because all markets are already highly persistent, differences within this narrow range provide little ability to distinguish markets that are more or less forecastable. Persistence therefore explains a common property of the markets but not their cross-market differences in forecast skill.

5.7.2 Seasonal strength is associated with longer-horizon forecastability

Table 8 extends the explanatory analysis to trend strength, seasonal strength, normalised week-to-week volatility, and spectral entropy. The values are Spearman rank correlations between each market characteristic and the best observed forecast skill S^* .

Table 8 Spearman rank correlations between market characteristics and best observed forecast skill S^* . Bold values indicate associations with Holm-adjusted $p < 0.05$.

Characteristic	H = 1	H = 4	H = 8	H = 13
ACF(1)	-0.128	0.120	0.042	-0.103
Trend strength	-0.081	0.066	0.118	0.316
Seasonal strength	0.157	0.618	0.681	0.826
Norm. diff. volatility	0.010	0.110	0.132	0.257
Spectral entropy	0.311	-0.130	-0.086	-0.039

Seasonal strength is the only measured characteristic that shows a statistically reliable association with forecastability. At one week, its correlation with best skill is weak ($\rho_s = 0.157$) and not statistically significant. The association becomes substantially stronger at four weeks ($\rho_s = 0.618$, $p_{\text{adj}} = 0.041$), eight weeks ($\rho_s = 0.681$, $p_{\text{adj}} = 0.013$), and thirteen weeks ($\rho_s = 0.826$, $p_{\text{adj}} < 0.001$). The relationship therefore strengthens as the forecast target moves further into the future.

The horizon dependence of this association provides an intuitive explanation for the earlier forecasting results. At one week, the current price remains a highly informative reference, so recurring annual structure offers relatively little additional advantage. At longer horizons, the latest price becomes less representative of the future target, while knowledge of how prices typically evolve through the year can become increasingly useful. This interpretation is consistent with the growing advantage of ETS, a model explicitly capable of adapting level, trend, and seasonal structure.

The association should nevertheless not be interpreted as proof of causation. Only 17 markets are available for the cross-market analysis, and S^* is selected as the maximum observed skill among the evaluated challenger models. The result therefore shows that markets with stronger measured seasonality tend to exhibit larger longer-horizon gains; it does not establish that seasonality alone causes those gains.

The remaining characteristics provide much weaker explanations. Trend strength reaches a correlation of 0.316 at thirteen weeks, but the association is not statistically significant after correction ($p_{\text{adj}} = 0.865$). Normalised volatility and spectral entropy are also not significantly related to forecast skill at any horizon. Among the measured characteristics, seasonal strength therefore provides the clearest empirical distinction between markets in which structured forecasting increasingly improves upon persistence and those in which it does not.

5.7.3 Contrasting market patterns

Delhi illustrates a market with strong seasonal structure and increasing longer-horizon forecastability. Its seasonal strength is $F_S = 0.735$, the highest among all 17 centres. Its best forecast skill increases from only 0.007 at one week to 0.183 at four weeks, 0.328 at eight weeks, and 0.406 at thirteen weeks. ETS provides the best forecast at each of the three longer horizons. Delhi therefore illustrates how a market can remain strongly persistent at short horizons while still containing recurring structure that becomes useful further into the future.

Lucknow, Shimla, and Patna provide similar examples of markets with relatively strong seasonal structure. Their seasonal-strength values are 0.513, 0.479, and 0.410, respectively. These centres contribute to the broader positive relationship between seasonal strength and longer-horizon forecast skill, although the preferred model is not necessarily identical in every market. For example, Seasonal Naive becomes the best method for Shimla at thirteen weeks, showing that the form in which seasonal information is useful can vary between markets.

Ahmedabad represents a contrasting case. Despite strong short-term persistence, its seasonal strength is only 0.106, and none of the evaluated forecasting methods improves upon Naive-1 at any horizon. Shillong provides another useful contrast: it has the highest lag-1 autocorrelation in the dataset (0.986) but no detectable annual seasonal structure under the adopted measure ($F_S = 0.000$). It remains persistence-dominated through eight weeks and shows only a negligible positive skill of 0.002 at thirteen weeks. These examples reinforce the distinction between persistence and seasonality: persistence makes Naive-1 strong, whereas seasonal structure appears to determine whether meaningful additional information becomes available at longer horizons.

Taken together, the explanatory results suggest that the 17 markets are not separated primarily by whether they are persistent—almost all of them are. They are separated more clearly by whether their historical series contain recurring seasonal structure that can complement persistence as the forecasting horizon increases. This finding provides the main empirical explanation for why ETS becomes increasingly competitive at longer horizons.

5.8 Robustness of the findings

5.8.1 Temporal stability of short-horizon forecastability

The first robustness analysis examines whether the observed level of short-horizon forecastability remains stable over different parts of the evaluation period. The 149 common forecast origins are divided chronologically into early, middle, and late sub-periods. Because these shorter segments contain too few observations for reliable centre-level DM-HLN testing, the analysis uses descriptive skill bands and is restricted to the one-week horizon.

Table 9 classifies each centre into band A when $S^* \leq 0$, band B when $0 < S^* \leq 0.05$, and band C when $S^* > 0.05$. These bands describe the magnitude of numerical skill within each subperiod and should not be confused with the inferential A/B/C classification used for the full evaluation sample.

Table 9 Temporal robustness at the one-week horizon using descriptive skill bands. A: $S^* \leq 0$; B: $0 < S^* \leq 0.05$; C: $S^* > 0.05$.

Centre	Early	Middle	Late	Stable?
AGARTALA	A	A	A	Yes
AHMEDABAD	A	A	A	Yes
BENGALURU	A	C	B	No
BHOPAL	A	A	A	Yes
BHUBANESHWAR	A	B	B	No
CHENNAI	A	B	C	No
DELHI	A	B	C	No
GUWAHATI	A	A	C	No
HYDERABAD	A	C	B	No
JAIPUR	B	B	A	No
KOLKATA	A	A	A	Yes
LUCKNOW	B	C	A	No
MUMBAI	A	B	A	No
PATNA	A	B	A	No
SHILLONG	A	A	A	Yes
SHIMLA	A	A	A	Yes
T.PURAM	A	B	A	No

Only 6 of the 17 centres remain in the same descriptive band across all three subperiods: Agartala, Ahmedabad, Bhopal, Kolkata, Shillong, and Shimla. All six remain in band A. The remaining 11 centres move between bands, indicating that the magnitude of one-week forecastability changes over time rather than remaining a fixed property of each market.

The corresponding agreement rate is

$$A_{\text{temporal}} = \frac{6}{17} = 0.353. \quad (15)$$

Thus, only 35.3% of the centres retain exactly the same one-week descriptive skill band throughout the three historical subperiods. The result cautions against permanently labelling a market as either forecastable or persistence-dominated on the basis of one historical period.

Some markets show particularly clear temporal changes. Delhi moves from band A in the early period to B in the middle period and C in the late period, whereas Lucknow moves from B to C and then to A. The present experiment cannot determine whether these changes arise from production conditions, storage behaviour, weather, policy changes, market integration, or other structural factors because those variables are not included in the dataset. The robustness result therefore establishes temporal variation in numerical forecastability, but not its economic cause.

This analysis also has an important scope limitation. It is performed only for $h = 1$, so it cannot establish whether the longer-horizon relationship between seasonality and forecast skill is equally stable across historical subperiods. The appropriate conclusion

is that *short-horizon numerical forecastability is time-dependent*; temporal robustness across all four horizons is not tested here.

5.8.2 Stability across neural-network random seeds

The second robustness analysis examines whether the weak neural-network performance could be caused by an unfavourable random initialisation. Neural models may produce slightly different fitted parameters when their initial weights or training batches change. Table 10 therefore reports the mean and standard deviation of MASE over the five random seeds used for each neural architecture.

Table 10 Seed stability of the neural models averaged across the 17 market centres. Values are mean $\pm$ standard deviation of MASE over five random seeds.

Model	H = 1	H = 4	H = 8	H = 13
LSTM	5.58 $\pm$ 0.13	9.90 $\pm$ 0.10	13.15 $\pm$ 0.16	15.64 $\pm$ 0.15
N-BEATS	4.86 $\pm$ 0.11	9.88 $\pm$ 0.13	13.67 $\pm$ 0.19	16.49 $\pm$ 0.07
PatchTST	9.32 $\pm$ 0.03	12.00 $\pm$ 0.16	14.46 $\pm$ 0.17	16.14 $\pm$ 0.13
iTransformer	7.19 $\pm$ 0.10	10.52 $\pm$ 0.12	13.15 $\pm$ 0.11	15.18 $\pm$ 0.09
TCN	3.55 $\pm$ 0.11	9.06 $\pm$ 0.10	13.03 $\pm$ 0.18	16.25 $\pm$ 0.11

Variation across random seeds is small relative to the difference between the neural models and Naive-1. At one week, for example, TCN is the strongest neural model with 3.55 $\pm$ 0.11 MASE, whereas Naive-1 achieves 2.03. The difference of approximately 1.52 MASE units is substantially larger than the 0.11 standard deviation produced by changing the random seed. The same general pattern occurs at the remaining horizons.

The neural-model conclusion is therefore not driven by a single unusually poor random training run. This robustness result does not imply that the tested architectures have reached their maximum possible performance. A wider hyperparameter search, larger architectures, additional training observations, or external explanatory variables could produce different outcomes. It establishes only that the conclusions reported for the specific neural implementations used in this study are stable across the five evaluated random seeds.

5.9 Summary of the empirical findings

The results reveal a clear horizon-dependent forecasting pattern. At one week, Naive-1 is the strongest average method, and only 5 of 17 markets show even a numerical improvement over persistence. As the horizon increases, positive skill becomes progressively more common, reaching 16 of 17 markets at thirteen weeks. ETS simultaneously becomes the strongest average forecasting method and beats Naive-1 in 14 of the 17 centres at the longest horizon.

The longer-horizon gains are not associated with increasing model complexity. Machine-learning and deep-learning models rarely improve upon Naive-1, whereas the

comparatively parsimonious ETS model captures most of the observed gains. Across-market statistical testing confirms that ETS is significantly better than Naive-1 at thirteen weeks, although none of the individual market-level improvements remains significant after the more conservative Holm-adjusted centre-level testing procedure.

The explanatory analysis shows why this transition is likely to occur. All markets exhibit strong short-term persistence, but differences in persistence do not explain differences in forecastability across markets. Seasonal strength provides a much clearer distinction: its association with best observed skill increases from $\rho_s = 0.157$ at one week to $\rho_s = 0.826$ at thirteen weeks and becomes statistically significant from the four-week horizon onward. The empirical evidence therefore indicates a transition from short-horizon forecasting dominated by persistence toward longer-horizon forecasting in which recurring seasonal structure becomes increasingly useful.

Finally, the robustness analyses place important limits on this conclusion. One-week numerical forecastability changes across historical subperiods for most markets, showing that forecastability should not be treated as a permanent market property. At the same time, the weak average performance of the neural models is stable across random initialisations. These findings provide the empirical basis for the broader interpretation, practical implications, and limitations discussed in Section 6.

6 Discussion

6.1 Forecastability depends on the forecast horizon

The main finding of this study is that potato-price forecastability depends strongly on the forecast horizon. At one week, Naive-1 is difficult to improve upon, whereas structured forecasting methods become more useful at 4-, 8-, and 13-week horizons. Forecastability should therefore not be treated as a fixed property of a market; it should be interpreted relative to both the forecast horizon and the benchmark used for comparison.

The strong short-horizon performance of Naive-1 is consistent with the high lag-1 autocorrelation observed across all 17 markets. When consecutive weekly prices are similar, the most recent price already provides a strong estimate of the immediate future. As the horizon increases, however, carrying the same price forward becomes increasingly restrictive, creating more opportunity for models that represent broader temporal structure. This supports the wider forecasting principle that complex methods should be evaluated against strong simple benchmarks rather than judged only by absolute error (Makridakis et al. 2018, 2020; Petropoulos et al. 2022).

6.2 Seasonality explains longer-horizon differences

Seasonal strength provides the clearest explanation for why some markets become more forecastable than others at longer horizons. Persistence is high across nearly all centres and therefore explains why Naive-1 is generally strong, but it does not explain cross-market differences in forecast skill. In contrast, the association between seasonal strength and forecast improvement becomes progressively stronger from four to thirteen weeks.

This distinction is important because persistence and seasonality represent different forms of temporal structure. Persistence reflects similarity between neighbouring weeks, whereas seasonality represents recurring patterns over longer periods. Delhi, for example, combines strong persistence with strong seasonality and shows substantial longer-horizon gains, while Shillong exhibits very strong persistence but almost no measured annual seasonality and little improvement over Naive-1. These contrasts suggest that recurring structure becomes useful once the latest observed price is no longer sufficient.

This interpretation is also plausible in the context of Indian potato markets, where production and market arrivals show substantial seasonal variation (Saha et al. 2020). However, the relationship observed here is associative rather than causal. The analysis shows that stronger seasonality and higher forecast skill occur together, but identifying the economic mechanisms behind this relationship would require additional information on arrivals, production, storage, weather, and other market conditions.

6.3 Greater model complexity does not guarantee better forecasts

The longer-horizon improvements are concentrated mainly in Exponential Smoothing (ETS), with some contribution from Autoregressive Integrated Moving Average (ARIMA), whereas machine-learning and deep-learning models rarely outperform Naive-1. Thus, within the information setting considered here, greater model complexity does not translate directly into greater forecasting value.

This result should not be interpreted as evidence that machine learning or deep learning is unsuitable for agricultural forecasting. Each market contains only 377 development observations, and the models use only the historical price of that market. Highly flexible architectures require sufficient observations and informative inputs to estimate complex relationships reliably. In this study, the useful longer-horizon structure appears sufficiently regular for a parsimonious statistical model to capture it effectively.

This finding differs from several potato-price studies in which neural, hybrid, or automated methods achieved the lowest reported errors (Praveenkumar et al. 2025; Nayak et al. 2025; Kumar et al. 2025a; Kumari et al. 2025). The difference does not necessarily represent a contradiction because those studies use different markets, data frequencies, evaluation designs, and comparison baselines. The present study asks a different question: whether additional modelling complexity provides consistent improvement over a strong persistence benchmark across markets and forecast horizons.

The weak performance of complex univariate models also suggests that future improvements may depend more on additional information than on additional architecture. Previous studies have shown that related-market prices and exogenous variables such as weather can improve agricultural-price forecasts (Jaiswal et al. 2025; Nayak et al. 2024). Future work should therefore examine whether market arrivals, production, storage, weather, transportation, or neighbouring-market prices provide predictive information that is absent from a market's own historical price sequence.

1335 6.4 Statistical evidence and temporal stability

1336 The statistical analysis provides an important qualification to the numerical improve-
1337 ments. Although several centres show substantial positive skill at longer horizons, none
1338 of the individual centre–horizon gains remains statistically significant after correcting
1339 for multiple comparisons. In contrast, the across-market analysis shows that ETS per-
1340 forms significantly better than Naive-1 at thirteen weeks. The evidence is therefore
1341 stronger for a general longer-horizon pattern across markets than for claims about any
1342 single centre.

1343 This distinction highlights the importance of reporting both effect size and statis-
1344 tical evidence. A large numerical improvement may still be estimated with substantial
1345 uncertainty, particularly when many model–market comparisons are tested simul-
1346 taneously. Positive forecast skill should therefore be interpreted as an observed
1347 improvement, whereas statistical significance indicates stronger evidence that the
1348 improvement is reproducible.

1349 Forecastability also varies over time. Only 6 of the 17 centres remain in the same
1350 one-week descriptive skill band across the early, middle, and late evaluation periods.
1351 A market that is persistence-dominated during one period may therefore become more
1352 forecastable during another. Because this robustness analysis is restricted to the one-
1353 week horizon, it does not establish whether longer-horizon forecastability is equally
1354 time-varying.

1356 6.5 Practical implications

1357 The results suggest that model selection should begin with the forecast horizon rather
1358 than with model complexity. For one-week forecasting, Naive-1 provides a strong,
1359 transparent benchmark that should be difficult to justify replacing unless a more
1360 complex method demonstrates a clear and repeatable advantage.

1361 At longer horizons, structured statistical models deserve greater attention, partic-
1362 ularly in markets with visible seasonal behaviour. ETS provides the most consistent
1363 gains in the present study, suggesting that practitioners should first examine per-
1364 sistence and seasonality before deciding whether additional model complexity is
1365 warranted.

1366 Forecasting systems should also be evaluated repeatedly through time. Because
1367 forecastability can change across historical periods, a model that performs well once
1368 should not be assumed to remain optimal. Rolling evaluation against simple bench-
1369 marks provides a practical way to monitor whether additional modelling continues to
1370 provide value.

1371 These implications should be interpreted within the historical scope of the dataset,
1372 which ends in April 2015. The study therefore does not identify the currently
1373 best operational forecasting model for Indian potato markets. Its contribution is
1374 methodological: forecasting methods should be judged relative to strong benchmarks,
1375 appropriate horizons, and the temporal structure actually present in the data.

1376
1377
1378
1379
1380

6.6 Limitations and future research

The study has several limitations. First, the dataset contains 539 weekly observations per market from 2005–2015, with only 377 observations available for initial model development. This limits the amount of information available to data-intensive models and does not capture recent changes in Indian potato markets. Future studies should repeat the analysis using longer and more recent datasets.

Second, the forecasting problem is deliberately univariate. Weather, arrivals, production, storage, transportation, and prices from related markets are excluded. The study therefore measures forecastability from historical prices alone rather than the maximum predictability obtainable from all available information.

Third, the neural models are evaluated using reproducible but not exhaustive architecture searches. Their weak performance should therefore be interpreted as evidence about the tested implementations rather than proof that every possible neural configuration would perform poorly.

Fourth, the cross-market explanatory analysis contains only 17 centres. The strong association between seasonal strength and forecast skill should therefore be validated using a larger market sample and, where possible, multivariable analysis.

Finally, the study evaluates point-forecast accuracy rather than predictive uncertainty or economic decision value. Future work should examine prediction intervals and determine whether improvements in statistical accuracy translate into better storage, marketing, procurement, or other agricultural decisions.

Overall, the results show that the main challenge in weekly potato-price forecasting is not simply selecting a more powerful model. Short-horizon forecasts are dominated by persistence, whereas longer-horizon gains are associated more strongly with seasonal structure and are captured most consistently by a relatively simple statistical model. Forecasting value therefore depends on matching model complexity to the information structure, forecast horizon, and market context.

7 Key Research Takeaways

The first takeaway is that potato-price forecastability should not be treated as a fixed property of a market. Its usefulness depends strongly on the forecast horizon. At very short horizons, the most recently observed price already contains substantial predictive information, making Naive-1 difficult to improve upon. At longer horizons, however, additional temporal structure becomes increasingly useful. Forecast horizon should therefore be considered explicitly when selecting and evaluating a forecasting method.

The second takeaway is that a complex forecasting model is not necessarily a better forecasting model. In the present experiments, machine-learning and deep-learning approaches did not provide consistent advantages over the simple persistence benchmark, whereas Exponential Smoothing (ETS) became increasingly effective at longer horizons. Model selection should therefore be driven by demonstrated improvement over a strong benchmark rather than by architectural complexity alone.

The third takeaway is that persistence and seasonality play different roles in forecastability. Strong short-term persistence explains why Naive-1 performs well across most markets, but differences in persistence do not explain why some markets become

1427 easier to forecast than others. Seasonal strength provides a clearer distinction, par-
1428 ticularly at longer horizons, where markets with stronger recurring seasonal structure
1429 tend to show larger forecasting gains.

1430 The fourth takeaway concerns how forecasting performance should be evaluated.
1431 A lower numerical error does not automatically establish a reliable forecasting advan-
1432 tage. Improvements should be assessed using repeated out-of-sample forecasts, strong
1433 benchmarks, and appropriate statistical testing. The distinction between numerical
1434 improvement and statistically supported improvement is therefore important when
1435 interpreting model rankings.

1436 Finally, the results suggest that future progress in potato-price forecasting may
1437 depend more on adding informative variables than on increasing model complexity.
1438 Once historical prices are adequately represented through persistence and seasonal
1439 structure, further gains may require information such as market arrivals, production,
1440 storage conditions, weather, or prices from connected markets. The broader lesson is
1441 that forecasting performance improves when the model, forecast horizon, and available
1442 information are matched to the structure of the market being studied.

1443

1444 8 Conclusion

1445

1446 This study examined whether weekly potato prices can be forecast beyond a simple
1447 persistence benchmark across 17 Indian markets and four forecast horizons. Twelve
1448 forecasting methods were evaluated under a common rolling-origin framework, with
1449 emphasis on whether additional modelling provides meaningful improvement over
1450 Naive-1.

1451 Forecastability was strongly horizon-dependent. Naive-1 remained the strongest
1452 average method at one week, whereas positive forecast skill became more common at
1453 longer horizons. At 13 weeks, 16 of the 17 markets showed numerical improvement over
1454 Naive-1, and Exponential Smoothing (ETS) achieved the best average performance.

1455 The gains were not driven by model complexity. Machine-learning and deep-
1456 learning methods did not consistently outperform the persistence benchmark, while
1457 ETS captured most of the longer-horizon improvement. Cross-market analysis fur-
1458 ther showed that short-term persistence did not explain differences in forecastability,
1459 whereas seasonal strength was increasingly associated with forecast skill at longer
1460 horizons.

1461 Statistical testing provided an important qualification: no individual centre-
1462 horizon improvement remained significant after multiple-comparison correction,
1463 although ETS performed significantly better than Naive-1 across markets at the 13-
1464 week horizon. The evidence therefore supports a general longer-horizon advantage
1465 more strongly than market-specific claims.

1466 Overall, the study reframes potato-price forecasting as a problem of *forecastability*
1467 *beyond persistence*. Forecasting value depends on the interaction among horizon, mar-
1468 ket structure, and model choice, and additional complexity should be adopted only
1469 when it produces clear gains over strong simple benchmarks. Future work should eval-
1470 uate this framework using more recent data, additional markets, and external variables
1471 such as arrivals, weather, storage, production, and related-market prices.

1472

Declarations	1473
Funding	1474
The authors received no specific funding for this work.	1475
	1476
Competing interests	1477
	1478
The authors have no relevant financial or non-financial interests to disclose.	1479
	1480
Data and code availability	1481
	1482
The analysed weekly potato-price panel, complete experimental pipeline, rolling-origin	1483
evaluation code, saved predictions and metrics, significance-test outputs, environment	1484
specification, and figure-generation scripts are provided with this article as zip file for	1485
the analysis and verification of the work. The archive documents all random seeds and	1486
model configurations needed to reproduce the reported analysis.	1487
	1488
Author contributions	1489
	1490
Muktinath Vishwakarma: Conceptualization, methodology, software, formal analysis,	1491
data curation, visualization, and writing—original draft. Manish Kurhekar: Supervi-	1492
sion, validation, and writing—review and editing. Both authors read and approved	1493
the final manuscript.	1494
	1495
Declaration of Generative AI Use	1496
	1497
Generative AI was used solely for language polishing and improving the clarity and	1498
readability of the manuscript. All scientific content, analyses, results, and conclusions	1499
were produced and verified by the authors, who take full responsibility for the final	1500
manuscript.	1501
	1502
Ethics approval	1503
	1504
Not applicable. The study does not involve human participants, human data, or	1505
animals.	1506
	1507
	1508
	1509
	1510
	1511
	1512
	1513
	1514
	1515
	1516
	1517
	1518

1519 References

- 1520
 1521 Akiba T, Sano S, Yanase T, et al (2019) Optuna: A next-generation hyperparameter
 1522 optimization framework. In: Proceedings of the 25th ACM SIGKDD International
 1523 Conference on Knowledge Discovery and Data Mining, pp 2623–2631, [https://doi.](https://doi.org/10.1145/3292500.3330701)
 1524 [org/10.1145/3292500.3330701](https://doi.org/10.1145/3292500.3330701)
- 1525
 1526 Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional
 1527 and recurrent networks for sequence modeling. arXiv preprint arXiv:180301271 URL
 1528 <https://arxiv.org/abs/1803.01271>
- 1529
 1530 Box GEP, Jenkins GM, Reinsel GC, et al (2015) Time Series Analysis: Forecasting
 1531 and Control, 5th edn. John Wiley & Sons
- 1532
 1533 Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings
 1534 of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and
 1535 Data Mining, pp 785–794, <https://doi.org/10.1145/2939672.2939785>
- 1536
 1537 Cleveland RB, Cleveland WS, McRae JE, et al (1990) STL: A seasonal-trend
 1538 decomposition procedure based on loess. Journal of Official Statistics 6(1):3–73
- 1539
 1540 Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time
 1541 series with a unit root. Journal of the American Statistical Association 74(366):427–
 1542 431. <https://doi.org/10.1080/01621459.1979.10482531>
- 1543
 1544 Diebold FX, Mariano RS (1995) Comparing predictive accuracy. Journal of Busi-
 1545 ness & Economic Statistics 13(3):253–263. [https://doi.org/10.1080/07350015.1995.](https://doi.org/10.1080/07350015.1995.10524599)
 1546 [10524599](https://doi.org/10.1080/07350015.1995.10524599)
- 1547
 1548 Friedman M (1937) The use of ranks to avoid the assumption of normality implicit
 1549 in the analysis of variance. Journal of the American Statistical Association 32(200):675–701. <https://doi.org/10.1080/01621459.1937.10503522>
- 1550
 1551 Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean
 1552 squared errors. International Journal of Forecasting 13(2):281–291. [https://doi.org/](https://doi.org/10.1016/S0169-2070(96)00719-4)
 1553 [10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4)
- 1554
 1555 Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation
 1556 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- 1557
 1558 Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogo-
 1559 nal problems. Technometrics 12(1):55–67. [https://doi.org/10.1080/00401706.1970.](https://doi.org/10.1080/00401706.1970.10488634)
 1560 [10488634](https://doi.org/10.1080/00401706.1970.10488634)
- 1561
 1562 Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian
 1563 Journal of Statistics 6(2):65–70
- 1564

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: The forecast package for R. <i>Journal of Statistical Software</i> 27(3):1–22. https://doi.org/10.18637/jss.v027.i03	1565 1566 1567 1568
Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. <i>International Journal of Forecasting</i> 22(4):679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001	1569 1570 1571 1572
Jaiswal R, Jha GK, Kumar RR, et al (2025) NARX model for potato price prediction utilising multimarket information. <i>Potato Research</i> https://doi.org/10.1007/s11540-025-09846-0	1573 1574 1575 1576
Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR)	1577 1578 1579
Kumar P, Jha GK, Kumar RR, et al (2025a) Leveraging singular spectrum analysis and time delay neural network for improved potato price forecasting. <i>Potato Research</i> 68(2):1803–1820. https://doi.org/10.1007/s11540-024-09806-0	1580 1581 1582 1583
Kumar R, Lad YA, Kumari P (2025b) Forecasting potato prices in Agra: Comparison of linear time series statistical vs. neural network models. <i>Potato Research</i> 68(3):2523–2544. https://doi.org/10.1007/s11540-024-09838-6	1584 1585 1586 1587
Kumar S, Arora K, Singh P, et al (2022) Performance comparison of ARIMA and time delay neural network for forecasting of potato prices in india. <i>Agricultural Economics Research Review</i> 35(2):119–134. https://doi.org/10.5958/0974-0279.2022.00035.0	1588 1589 1590 1591
Kumari P, M SK, Vekariya P, et al (2025) Predicting potato prices in Agra, UP, india: An H2O AutoML approach. <i>Potato Research</i> 68(1). https://doi.org/10.1007/s11540-024-09726-z	1592 1593 1594 1595
Kwiatkowski D, Phillips PCB, Schmidt P, et al (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. <i>Journal of Econometrics</i> 54(1–3):159–178. https://doi.org/10.1016/0304-4076(92)90104-Y	1596 1597 1598 1599
Liu Y, Hu T, Zhang H, et al (2024) iTransformer: Inverted transformers are effective for time series forecasting. In: International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=JePfAI8fah	1600 1601 1602 1603
Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and machine learning forecasting methods: Concerns and ways forward. <i>PLoS ONE</i> 13(3):e0194889. https://doi.org/10.1371/journal.pone.0194889	1604 1605 1606 1607
Makridakis S, Spiliotis E, Assimakopoulos V (2020) The M4 competition: 100,000 time series and 61 forecasting methods. <i>International Journal of Forecasting</i> 36(1):54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014	1608 1609 1610

1611 Nayak GHH, Alam MW, Singh KN, et al (2024) Exogenous variable driven deep
 1612 learning models for improved price forecasting of TOP crops in india. Scientific
 1613 Reports 14:17203. <https://doi.org/10.1038/s41598-024-68040-3>
 1614
 1615 Nayak GHH, Alam MW, Avinash G, et al (2025) N-BEATS deep learning architec-
 1616 ture for agricultural commodity price forecasting. Potato Research 68(2):1437–1460.
 1617 <https://doi.org/10.1007/s11540-024-09789-y>
 1618
 1619 Nie Y, Nguyen NH, Sinthong P, et al (2023) A time series is worth 64 words:
 1620 Long-term forecasting with transformers. In: International Conference on Learning
 1621 Representations (ICLR), URL <https://openreview.net/forum?id=B9tMrx5nYE>
 1622
 1623 Oreshkin BN, Carpow D, Chapados N, et al (2020) N-BEATS: Neural basis expan-
 1624 sion analysis for interpretable time series forecasting. In: International Confer-
 1625 ence on Learning Representations (ICLR), URL <https://openreview.net/forum?id=r1ecqn4YwB>
 1626
 1627 Petropoulos F, et al (2022) Forecasting: Theory and practice. International Journal of
 1628 Forecasting 38(3):705–871. <https://doi.org/10.1016/j.ijforecast.2021.11.001>
 1629
 1630 Praveenkumar A, Jha GK, Madival SD, et al (2025) Deep learning approaches
 1631 for potato price forecasting: Comparative analysis of LSTM, Bi-LSTM, and
 1632 AM-LSTM models. Potato Research 68(2):1941–1965. <https://doi.org/10.1007/s11540-024-09823-z>
 1633
 1634
 1635 Rana RK, Pandey NK, Kumar NR, et al (2003) Seasonal price indices of potato for
 1636 major indian markets. Potato Journal 30(1–2):201–202
 1637
 1638 Saha N, Kar A, Jha GK, et al (2020) Market arrival and price behaviour analysis
 1639 of potato in four major markets in india. Economic Affairs 65(4):529–533. <https://doi.org/10.46852/0424-2513.4.2020.8>
 1640
 1641
 1642 Shankar SV, Chandel A, Gupta RK, et al (2025) Comparative study on key time
 1643 series models for exploring the agricultural price volatility in potato prices. Potato
 1644 Research 68(2):1189–1207. <https://doi.org/10.1007/s11540-024-09776-3>
 1645
 1646 Sun F, Meng X, Zhang Y, et al (2023) Agricultural product price forecasting methods:
 1647 A review. Agriculture 13(9):1671. <https://doi.org/10.3390/agriculture13091671>
 1648
 1649 Tashman LJ (2000) Out-of-sample tests of forecasting accuracy: An analysis and
 1650 review. International Journal of Forecasting 16(4):437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)
 1651
 1652 Wang L, Feng J, Sui X, et al (2020) Agricultural product price forecasting methods:
 1653 Research advances and trend. British Food Journal 122(7):2121–2138. <https://doi.org/10.1108/BFJ-09-2019-0683>
 1654
 1655
 1656

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bulletin	1657
1(6):80–83. https://doi.org/10.2307/3001968	1658
	1659
	1660
	1661
	1662
	1663
	1664
	1665
	1666
	1667
	1668
	1669
	1670
	1671
	1672
	1673
	1674
	1675
	1676
	1677
	1678
	1679
	1680
	1681
	1682
	1683
	1684
	1685
	1686
	1687
	1688
	1689
	1690
	1691
	1692
	1693
	1694
	1695
	1696
	1697
	1698
	1699
	1700
	1701
	1702