PUBLIC HEALTH
OpenJournal
Openventio PUBLISHERS

Research

A Conditional Value Test of Exogenous Variables in Potato Price Forecasting

Raj Shekar M[1,2], Sujith Arokia Swamy B[1,2,*], Naveen Kumar P[3], Ramya Jagadish[4], Bhavana Kabber[3], Harish Nayak[5]

[1]The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India – 110012
[2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India – 110012
[3]University of Agricultural Sciences, Raichur, Karnataka, India – 584104
[4]Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India – 641003
[5]University of Agricultural Sciences, Dharwad, Karnataka, India – 580005

Corresponding Author
Sujith Arokia Swamy B
Email: sujith1907b@gmail.com

Article Information:
Received: 21 December 2025 | Revised: 13 January 2026 | Accepted: 15 January 2026 | Published: January 15, 2026

Cite this article:
Raj Shekar M, Sujith Arokia Swamy B, Naveen Kumar P, Ramya Jagadish, Bhavana Kabber and Harish Nayak(2026). A Conditional Value Test of Exogenous Variables in Potato Price Forecasting *Public Health Open Journal*. 11(S1):85–96. DOI: 10.17140/PHOJ.11.S1.85

Abstract
Whether weather variables improve agricultural price forecasts is usually answered on average, yet theory and some evidence suggest that weather matters only in particular regimes, such as drought. This study turns the question into one about conditional value: when does weather help, and can a rule detect those occasions in real time, Using weekly potato prices with temperature and precipitation (January 2003 – September 2023), three forecasters are compared in an expanding-window walk-forward design covering about 15 years of out-of-sample forecasts: a naive no-change forecast (A), a price-only machine-learning model (B) and a model that adds lagged temperature and precipitation (C). A confidence gate, which opens only when a model has shown consistent level-accuracy and turning-point improvements over its reference on recently realised origins, is applied to B versus A and to C versus B, and raw prices are compared with Kalman-filled prices as inputs. Price-only models clearly beat the naive forecast at four weeks ahead (relative MAE 0.85-0.90, Diebold-Mariano $p < 0.001$) but not significantly at one week. Weather did not improve accuracy on average at either horizon: relative to B, model C ranged from about 0.99 to 1.02 times B's error and was significantly worse for ridge regression at four weeks ($p = 0.017$). The gate between C and B stayed mostly closed (5-21% of origins), and its open and closed states were no better than chance at identifying periods in which weather later helped; the gate between B and A cost accuracy relative to always using B, because B was already reliably better than naive at four weeks. Descriptively, weather added value in the three drought years in the evaluation sample (MAE ratio 0.97 at one week versus 1.03 in other years), but the contrast was not statistically significant ($p = 0.064$). Kalman-filled inputs helped modestly for boosted trees at one week. The results support a conditional, regime-specific view of weather in these data but show that detecting the regime from recent forecast performance is hard, so a null result for the gate is informative about the limits of detecting conditional value.

Keywords: *potato prices, weather variables, forecast gating, Kalman filter, walk-forward evaluation, conditional predictive ability*

1. Introduction

Potato is a staple, perishable and storage-sensitive crop whose wholesale price swings sharply from season to season. Farmers, cold-storage operators and traders need forecasts to time sales and storage, and policymakers need early warning of price spikes that squeeze household budgets (Gilbert & Morgan, 2010; Tomek & Peterson, 2001; Wright, 2011). Because agricultural output depends on temperature and rainfall (Schlenker & Roberts, 2009; Lobell et al., 2011), it is natural to add weather variables to price-forecasting models, and a large applied literature does so on the premise that weather is an informative leading indicator of supply and therefore of price.

The empirical record is, however, mixed. Studies of forecasting competitions and benchmark comparisons show that added complexity and added predictors often fail to beat simple benchmarks out of sample (Makridakis et al., 2018, 2020; Petropoulos et al., 2022), and the equity-premium and exchange-rate literatures have long warned that predictors that look useful in sample frequently lose their value in forecasting (Meese & Rogoff, 1983; Goyal & Welch, 2008; Stock & Watson, 2003). In agricultural markets there is a further reason to expect weak average effects: weather affects prices mainly through extreme supply shortfalls, while in ordinary seasons storage, trade and policy absorb weather variation. If so, the value of weather is conditional: negligible in normal years and substantial in drought. Averaging over regimes then hides the effect and may lead a researcher to conclude, wrongly, that weather does not matter at all.

A conditional effect poses a second, operational question. Even if weather helps only in particular regimes, a forecaster in real time does not know the regime in advance. A natural solution is a gate that lets the weather-augmented model speak only when it has recently shown that it is more accurate than a weather-free one. Related ideas exist in forecast combination, where weights shift toward recent winners (Bates & Granger, 1969; Timmermann, 2006), and in conditional predictive ability testing (Giacomini & White, 2006; Giacomini & Rossi, 2010), but to our knowledge, nobody has asked whether such a gate can actually detect the regimes in which weather helps, nor whether the dual criteria of level accuracy and turning-point skill, which matter to decision makers, can be assessed on a short window of recent forecasts.

This paper therefore asks not whether weather helps on average but when it helps and whether that can be detected. Three forecasters are compared on weekly potato prices: a naive no-change forecast, a price-only machine-learning model and a price-plus-weather model. The study examines whether price-only machine learning beats the naive forecast, whether adding lagged temperature and precipitation improves on the price-only model on average and within drought-like and other regimes, and whether a confidence gate applied to the price-only versus naive comparison and to the weather versus price-only comparison opens at the right times. It also tests whether feeding the models a Kalman-filled price series instead of the raw series, which contains gaps, substitutes for gating or alters the conclusions. A null result, in which the gate rarely opens or opens at random, is itself informative about the conditional value of weather and about the limits of recent-performance rules. The remainder of the paper reviews the literature (Section 2), describes the data (Section 3), sets out the methodology (Section 4), reports results (Section 5), discusses them (Section 6) and concludes (Section 7).

2. Review of Literature

Agricultural price forecasting has moved from time-series models to machine learning and hybrid methods. Classical approaches include exponential smoothing and ARIMA families (Hyndman & Athanasopoulos, 2021), and neural and hybrid statistical–neural models have been applied to Indian agricultural prices with reported gains over linear benchmarks (Jha & Sinha, 2014; Zhang, 2003; Zhang et al., 1998). Many of these studies evaluate one or a few series without a persistence benchmark scored under the same walk-forward protocol, so the generality of reported gains is unclear. Large forecasting competitions caution that statistical and simple methods are often hard to beat, that gains from machine learning depend on series characteristics and covariates, and that combinations are robust (Makridakis et al., 2018, 2020, 2022; Petropoulos et al., 2022). Tree ensembles are particularly competitive on tabular features (Friedman, 2001; Grinsztajn et al., 2022; Ke et al., 2017), which motivates their use here, with ridge regression (Hoerl & Kennard, 1970) as a linear contrast.

The role of weather is well established for crop yields. Temperature extremes and rainfall deficits reduce yields in a nonlinear way (Schlenker & Roberts, 2009; Lobell et al., 2011), and weather shocks are widely used as exogenous supply shifters in studies of agricultural supply and demand (Roberts & Schlenker, 2013). The macroeconomic literature finds that weather phenomena such as El Nino affect commodity and food prices, with effects that vary across countries and episodes (Cashin et al., 2017). Translating these supply effects into short-horizon price forecasts is less straightfor-

ward. Prices reflect stocks, trade, policy and expectations, so the pass-through of weather to prices may be weak, delayed or regime-dependent. In forecasting practice, weather variables often add little to price-only models, a pattern consistent with the broader finding that predictors with strong in-sample links frequently fail out of sample (Goyal & Welch, 2008; Stock & Watson, 2003). The possibility that the value of a predictor is concentrated in particular regimes, and therefore that average tests understate it, is a natural implication of instability research (Giacomini & Rossi, 2010; Rossi, 2013).

A second strand concerns data quality. Weekly agricultural price series frequently contain missing observations, repeated values from unrecorded market days and outliers. State-space models and the Kalman filter (Kalman, 1960; Harvey, 1989; Durbin & Koopman, 2012) are a standard tool for filling gaps and smoothing noise, and Kalman-based imputation is widely used for time series with missing values (Moritz & Bartz-Beielstein, 2017). Whether smoothing or imputation improves forecasting is not obvious: a smoother can remove noise that is irrelevant to prediction but can also remove information, and a two-sided smoother can leak future information into past values if applied to the full sample. The present study keeps this distinction explicit.

A third strand concerns how to evaluate and combine forecasts. Out-of-sample evaluation must respect time order (Tashman, 2000; Bergmeir & Benitez, 2012; Cerqueira et al., 2020), and equal predictive accuracy is typically tested with the Diebold-Mariano test and its small-sample correction (Diebold & Mariano, 1995; Harvey et al., 1997), with special care when models are nested (Clark & McCracken, 2001). Accuracy is usually measured by absolute or squared errors, scaled for comparability (Hyndman & Koehler, 2006), but decision-makers also care about directional performance and turning points (Pesaran & Timmermann, 1992; Granger & Pesaran, 2000). Forecast combination can shrink noisy forecasts toward robust benchmarks, and fixed, equal weights often outperform estimated weights because of estimation error (Smith & Wallis, 2009; Claeskens et al., 2016), a point that matters for any rule that switches between models on the basis of a short window of performance.

Taken together, the literature suggests that weather should matter for supply, that its marginal forecasting value may nonetheless be small on average and concentrated in specific regimes, and that a rule to detect those regimes from recent performance faces a severe estimation problem. No study, to our knowledge, has set up the test directly: price-only versus weather-augmented forecasts on a single clean dataset, a gate applied to both the machine-learning-versus-naive and the weather-versus-price-only comparisons, regime-specific evaluation, and a check on whether Kalman-filled prices substitute for gating. This paper fills that gap.

3. Data Description

The main data set contains 1,080 consecutive weekly observations (dated on Sundays, from 12 January 2003 to 17 September 2023) of potato price, a Kalman-processed price, mean temperature and precipitation. Prices are used in the units supplied, which appear to be a scaled unit rather than rupees per quintal (a daily modal-price file supplied alongside, covering 2006–2023, has weekly means about 8.7 times larger, with a correlation of 0.78 with the weekly series), so results are reported both in supplied units and, for comparability, relative to the naive forecast. Temperature is taken to be in degrees Celsius and precipitation in millimetres per week. The location of the weather station and the market are not documented in the files and should be specified by the author before submission.

Raw price is missing in 161 weeks (14.9%), all between 2003 and 2011; from 2012 onward the raw series is complete. The Kalman price is identical to the raw price wherever the raw price is observed (the two series differ only in the 161 missing weeks) and is therefore a gap-filled series rather than a smoothed one. Two further data features matter. First, the filled values in some early gaps are implausibly low (the minimum Kalman price is 0.64, against a raw minimum of 13.4), so the imputation is imperfect in the early period. Second, about 15% of weekly prices repeat the previous week's value, which suggests carried-forward quotations. In addition, the first year contains an isolated spike (138 in March 2003 between values of 34 and 17). It is not known whether the Kalman procedure was a one-sided filter or a two-sided smoother over the whole sample; if the latter, filled values inside gaps could contain information from later observations, a limitation discussed in Section 6. A third file with the Kalman price alone runs to December 2023 and agrees exactly with the Kalman column over the 1,080 overlapping weeks, but has no weather data for its extra weeks and was not used.

Table 1 and Figure 1 summarise the data. Prices are strongly right-skewed (skewness 1.55 for raw and 1.64 for Kalman price) with a pronounced peak in 2020, when the maximum of 438.8 was reached in December. Precipitation is highly skewed (skewness 5.8) and zero

Table 1. Descriptive Statistics of the Weekly Potato Data

Variable	n	Mean	SD	Min	Max	Skewness
Price (raw, observed weeks)	919	107.5	66.5	13.43	438.81	1.55
Price (Kalman)	1080	98.8	65.4	0.64	438.81	1.64
Temperature	1080	25.9	7.5	7.52	41.01	-0.22
Precipitation	1080	2.3	7.0	0.00	71.41	5.83

Note.Price units as supplied. The raw price is observed in 919 of 1,080 weeks; the Kalman price, temperature and precipitation are available in all 1,080 weeks.

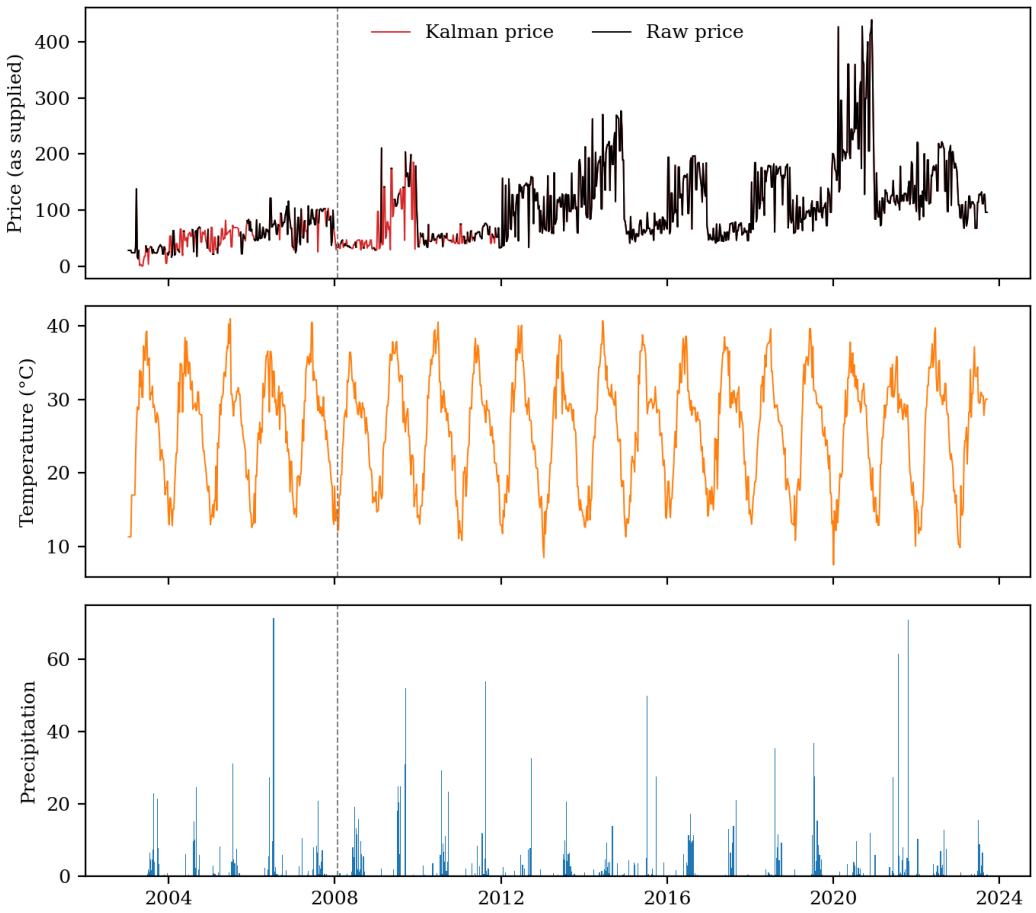


Figure 1. Weekly raw and Kalman-filled potato price, temperature and precipitation, 2003-2023. The dashed vertical line marks the start of the out-of-sample evaluation period (January 2008).

in 56% of weeks, as expected for a monsoon climate with a long dry season, while temperature has a strong seasonal cycle. Annual rainfall varies from 47.8 mm in 2020 to 244.1 mm in 2009. Defining drought years ex post as full calendar years in 2003-2022 with annual rainfall at or below the first quartile (93.8 mm) identifies 2005, 2007, 2014, 2017 and 2020; only 2014, 2017 and 2020 fall in the evaluation period that begins in 2008. Contemporaneous correlations between weekly log price changes and weather are close to zero (0.03 with temperature, 0.03 with precipitation and -0.01 with 13 week rainfall), a first indication that any weather signal must be subtle.

4. Methodology

The forecast target is the log price change over horizon h weeks, ln y(t+h) - ln y(t), and the level forecast is (t+h) = y(t) · exp(). Horizons of one and four weeks are studied. Forecasts are produced by expanding-window walk-forward evaluation: at each origin, models are trained using only samples whose targets were realised by that origin, and are re-estimated every 13 origins. The first origin is week 208 (about four years of training), and the evaluation window starts once a further 52 weeks plus h have elapsed, so that gated and ungated forecasts are scored on identical targets (713 origins at h = 1, 703 at h = 4, covering targets from January 2008 and March 2008 to September 2023, respectively). Because the raw price has gaps, only origins at which both the current and the target raw prices were observed are scored, and the models are trained on the same type of sample. Imputed Kalman values are therefore used only as inputs and never as targets, so all models are evaluated against the same observed prices.

Table 2 lists the forecasters. Model A is persistence, ŷ(t+h) = y(t). Model B is a price-only learner, either histogram gradient-boosted trees (HGB; 120 iterations, learning rate 0.05, depth 3, minimum leaf size 15, L2 penalty 1; Friedman, 2001; Ke et al., 2017; Pedregosa et al., 2011) or ridge regression (penalty 30; Hoerl & Kennard, 1970). Its features are log-price differences ln y(t-k) - ln y(t) for k = 1, 2, 3, 4, 8, 13, 26, 52, the standard deviation of recent log returns, the gaps between the current log price and its 13-week and 26-week means, and sine and cosine terms of the week of the year. Model C adds weather features: temperature at lags 1, 2 and 4 and its current value, 4- and 13-week mean temperature, precipitation at lags 1 and 2, 4-, 13- and 26-week precipitation totals, and a rainfall-anomaly index equal to the 26-week precipitation total divided by the mean of the same quantity in the same week of earlier years (computed from past data only, with at least three prior years). Each learner is run with two inputs: the raw price series (gaps carried forward in the feature construction) and the Kalman price series. A seasonal naive forecast (price 52 weeks before the target) serves as a turning-point benchmark. Hyperparameters were fixed beforehand and not tuned on the evaluation sample.

A turning point at an origin is a reversal in the direction of the h-step change: the realised change over the next h weeks has the opposite sign to the preceding h-week change. A forecast predicts a reversal when its implied change has the opposite sign to the preceding change, and turning-point skill is measured by the F1 score of predicted against realised reversals. Persistence never predicts a reversal, so its F1 is zero by construction and turning-point skill is benchmarked against seasonal naive for the comparison of B with A, and against the reference model itself for the comparison of C with B.

The confidence gate compares a candidate model with a reference using only forecast errors that would have been realised at the decision time. At each origin it takes the most recent W realised origins (W = 26 in the main analysis, with 13 and 52 as sensitivity checks) and requires, first, that the candidate's mean absolute error be lower than the reference's and, second, that the candidate's reversal F1 exceed the references. To require consistency, both conditions must hold in the full window and in its more recent half (minimum four origins). When the gate is open the candidate's forecast is used, and otherwise the references. Gate 1 applies this rule to B against A, and Gate 2 to C against B. A cascade uses C when Gate 2 is open, B when only Gate 1 is open and A otherwise.

Accuracy is measured by Mean Absolute Error (MAE) relative to the naive forecast, so that values below one favour the model, and by reversal F1. Pairwise comparisons use the Diebold-Mariano test of equal absolute-error loss with a Bartlett long-run variance of lag h - 1 and the Harvey et al. (1997) small-sample correction (Diebold & Mariano, 1995); a positive statistic means that the first forecast in a comparison has the larger loss. Because the comparison of C with B involves nested models, the tests should be read as descriptive (Clark & McCracken, 2001), and no multiplicity correction is applied.

Three analyses assess whether weather matters conditionally and whether the gate detects it. First, accuracy of C relative to B is computed separately in weather regimes defined from the past-only rainfall-anomaly index (dry when the 26-week rainfall total is below 70% of its same-season historical mean, wet when above 130%, normal otherwise) and by season (monsoon months June to September versus other months). Second, an ex-post comparison contrasts drought years (2014, 2017 and 2020 in the evaluation sample) with other years. Third, gate quality is assessed by comparing the gate state with a forward-looking benchmark: weather is said to help at an origin if C has lower mean absolute error than B over the next 13 scored origins. This benchmark uses future information and is used only to evaluate the gate, never to make forecasts. Gate quality is summarised by accuracy of the gate state against this benchmark (compared with the accuracy of an always-closed gate), precision, recall and the correlation between the gate state and the realised benefit of weather, and by the ratio of C's to B's MAE on origins at which the gate is open and closed. A useful gate gives a ratio below one when open and above one when closed.

5. Results

Table 3 gives the central accuracy comparison. Persistence is a demanding level benchmark: seasonal naive has an MAE 2.96 times larger at one week and 2.41 times larger at four weeks (DM statistics 20.3 and 12.6, both $p < 0.001$), so repeating last year's price is not a viable alternative. The persistence MAE is 26.44 at h = 1 and 32.48 at h = 4 in the supplied units. Price-only machine learning improves on persistence, and the improvement grows with the horizon. At one week the relative MAE is 0.961 for HGB with raw prices and 0.934 with Kalman prices, and 0.981 and 0.982 for ridge, none of which is statistically distinguishable from persistence (Table 4). At four weeks the gains are larger and clearly significant for every learner and input: relative MAE of 0.866 (HGB, raw), 0.855 (HGB, Kalman), 0.898 (ridge, raw) and 0.893 (ridge, Kalman), with DM

Table 2. Forecasters and Gate Variants

Model	Specification
A. Naive	Persistence: ŷ(t+h) = y(t).
B. Price-only ML	Histogram gradient-boosted trees (HGB) or ridge regression on price-derived features only; raw-price or Kalman-price inputs.
C. Price + weather ML	Model B features plus lagged temperature, lagged precipitation, rolling precipitation and temperature summaries and a rainfall-anomaly index.
Gate 1 (B vs A)	Use B only when it has shown consistent level and turning-point improvement over naive; otherwise use A.
Gate 2 (C vs B)	Use C only when it has shown consistent level and turning-point improvement over B; otherwise use B.
Cascade	C if Gate 2 is open; otherwise B if Gate 1 is open; otherwise A.

Note.h = forecast horizon in weeks.

statistics between -3.79 and 4.31 (p < 0.001). Boosted trees are slightly more accurate than ridge regression at both horizons.

Adding weather does not improve accuracy on average. For boosted trees with raw prices, relative MAE moves from 0.961 (B) to 0.967 (C) at one week and from 0.866 to 0.862 at four weeks; with Kalman prices it moves from 0.934 to 0.924 and from 0.855 to 0.852. None of these differences is statistically significant (DM statistics for C versus B between -1.06 and 0.51, p 0.29; Table 4). For ridge regression the weather features worsen accuracy: relative MAE rises from 0.981 to 0.989 at one week and from 0.898 to 0.915 at four weeks with raw prices, and the deterioration at four weeks is statistically significant (DM = 2.39, p = 0.017; with Kalman prices, DM = 1.82, p = 0.069). The overall picture is that weather features are, on average, neutral for tree ensembles and mildly harmful for a linear model, in line with the view that the average marginal value of weather is small. Turning-point F1 is also unaffected: reversal F1 of the learners is 0.62-0.63 at one week and 0.70-0.74 at four weeks regardless of whether weather is included, which exceeds seasonal naive (0.54 and 0.60) but should be read against the high reversal frequency in these weekly data.

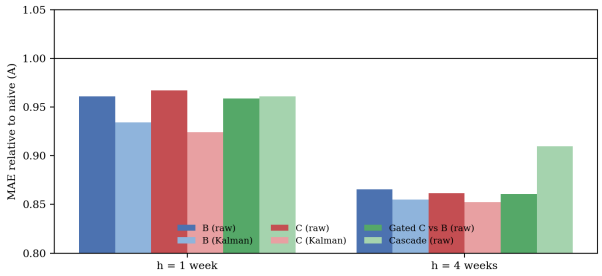


Figure 2. MAE of selected forecasters relative to the naive forecast (HGB learner), by horizon.

Table 5 shows what the gates do. Gate 2, which chooses between C and B, is open for only 5––21% of origins, and its forecasts are almost indistinguishable from those of B: for boosted trees with raw prices the gated relative MAE is 0.959 at one week and 0.861 at four weeks, compared with 0.961 and 0.866 for B and 0.967 and 0.862 for C. The gate therefore neither adds nor removes value, which is what one expects if weather has little to add and the gate, correctly or by chance, mostly stays closed. Gate 1, which chooses between B and A, behaves differently. It is open in 33––39% of origins at one week and 51––57% at four weeks, and it is costly: because B is reliably better than naive at four weeks, switching back to naive whenever the recent record looks unfavourable gives up much of the gain. For HGB with raw prices the four-week relative MAE rises from 0.866 (B always on) to 0.932 (Gate 1), and with Kalman prices from 0.855 to 0.897; the pattern is similar for ridge (0.898 to 0.964 and 0.893 to 0.958). The gated forecasts still beat persistence significantly at four weeks (for example, Gate 1 with raw HGB, DM = 2.57, p = .010), but they are worse than the ungated price-only model. The cascade, which includes both gates, is similar to Gate 1 (relative MAE 0.910 at four weeks for raw HGB). At one week none of the gated or cascaded forecasts is significantly different from persistence.

Table 6 addresses the central question, whether Gate 2 opens when weather helps. It does not do so reliably. For every learner and input, the correlation between the gate state and the realised benefit of weather over the next 13 origins lies between -0.03 and 0.09, which is essentially zero. The accuracy of the gate state against the forward-looking benchmark ranges from 0.45 to 0.65, but this must be compared with the accuracy of simply leaving the gate closed, which is 1 minus the share of origins at which weather in fact helps, and the gate rarely beats that baseline by a meaningful margin (for example, 0.552 against 0.501 for raw HGB at one week, and 0.575 against 0.576 for raw ridge). Precision is moderate in a few cases (0.70 for raw HGB at one week) but recall is low (0.06-0.19), so the gate misses most of the occasions on which weather later helps. The conditional MAE ratios tell a similar story: for raw HGB at one week, C's MAE relative to B is 0.983 when the gate is open and 1.010 when closed, in the direction

Table 3. Out-of-Sample Accuracy Relative to the Naive Forecast and Turning-Point F1

Forecaster	h = 1	h = 1	h = 4	h = 4
	rel. MAE	F1	rel. MAE	F1
A. Naive (persistence)	1.000	-	1.000	-
Seasonal naive	2.958	0.543	2.407	0.602
B. HGB, raw	0.961	0.631	0.866	0.715
B. HGB, Kalman	0.934	0.628	0.855	0.728
B. Ridge, raw	0.981	0.629	0.898	0.709
B. Ridge, Kalman	0.982	0.624	0.893	0.725
C. HGB, raw	0.967	0.624	0.862	0.725
C. HGB, Kalman	0.924	0.622	0.852	0.735
C. Ridge, raw	0.989	0.628	0.915	0.701
C. Ridge, Kalman	0.988	0.618	0.906	0.711

Note: Relative MAE = MAE of the forecaster divided by MAE of persistence (A); values below 1.000 favour the forecaster. Persistence MAE = 26.44 (h = 1) and 32.48 (h = 4) in supplied units. F1 = F1 score of predicted versus realised direction reversals; persistence predicts none, so its F1 is zero by construction. B = price-only; C = price + weather.

Table 4. Diebold-Mariano Statistics for Absolute-Error Loss (p-values in parentheses)

Comparison	Learner / input	h = 1	h = 4
B vs A	HGB, raw	-0.92 (.357)	-4.23 (< .001)
B vs A	HGB, Kalman	-1.49 (.138)	-4.31 (< .001)
B vs A	Ridge, raw	-0.50 (.617)	-3.79 (< .001)
B vs A	Ridge, Kalman	-0.44 (.660)	-3.92 (< .001)
C vs B	HGB, raw	0.51 (.607)	-0.38 (.701)
C vs B	HGB, Kalman	-1.06 (.290)	-0.29 (.775)
C vs B	Ridge, raw	1.27 (.205)	2.39 (.017)
C vs B	Ridge, Kalman	1.17 (.244)	1.82 (.069)
C vs A	HGB, raw	-0.78 (.438)	-4.02 (< .001)
C vs A	HGB, Kalman	-1.77 (.078)	-4.31 (< .001)
C vs A	Ridge, raw	-0.28 (.776)	-3.06 (.002)
C vs A	Ridge, Kalman	-0.28 (.777)	-3.30 (.001)

Note.Positive statistics indicate that the first forecast in the comparison has the larger loss. A = naive; B = price-only; C = price + weather. "Raw" and "Kalman" refer to the price input.

a useful gate would produce, but the difference is small, and with Kalman prices the ordering reverses (1.036 open, 0.987 closed). In sum, the gate is a weak and unstable detector of conditional value in these data.

Figure 3 shows that Gate 2 opens in short, scattered episodes that do not line up systematically with low-rainfall periods, while Gate 1 is open for long stretches interrupted by closures. Table 7 asks directly whether weather helps more in dry regimes. For boosted trees with raw prices, the ratio of C's to B's MAE at four weeks is 0.980 in dry regimes, 1.005 in normal regimes and 1.025 in wet regimes, and at one week 0.991, 0.992 and 1.084, respectively. With Kalman prices at one week the ratios are 0.967, 1.004 and 1.024. The ordering, with weather helping most in dry regimes and hurting in wet regimes, is in line with the conditional-value hypothesis, but the magnitudes are small (typically 1-3%), they are not uniform across learners (ridge regression shows ratios above one in nearly every regime), and they are not tested here for significance. The season split gives no consistent pattern either: weather features help slightly in the monsoon at four weeks for raw HGB (0.971) but hurt at one week (1.019). Gate 2 does behave sensibly in the wet regime, where weather was most harmful: its open rate for raw HGB at one week is 2% in wet regimes against 15-16% in dry and normal regimes (Table 7b). That correct closing in wet periods is the clearest sign of any regime sensitivity in the gate; it does not extend to dry regimes, where the gate opens no more often than in normal ones.

The ex-post classification of years gives a sharper, though still tentative, contrast. In the three drought years inside the evaluation sample (2014, 2017 and 2020; 157 scored weeks) the weather model's MAE is 3.3% lower than the price-only model's at one week (ratio 0.967), whereas in the remaining years it is 2.6% higher (ratio 1.026). The mean absolute-error difference is -1.25 in drought years (t = -1.36, p = 0.176) and +0.57 in other years (t = 1.75, p = 0.081), and the difference between the two groups has a Welch p of 0.064. At four weeks the contrast vanishes (ratios 0.992 and 0.997). This is the single result most supportive of the title's hypothesis. It is based on only three drought years, the classification is ex-post and based on annual rainfall, and the difference is not significant at conventional levels, so it should be treated as a hypothesis for further testing rather than as a finding.

Table 8 compares raw and Kalman-filled inputs. For

Table 5. Relative MAE of Ungated and Gated Forecasts (relative to naive)

h	Learner / input	B	Gate 1	C	Gate 2	Cascade	Gate 1 open	Gate 2 open
1	HGB, raw	0.961	0.991	0.967	0.959	0.961	0.34	0.13
	HGB, Kalman	0.934	0.967	0.924	0.936	0.965	0.39	0.07
	Ridge, raw	0.981	0.992	0.989	0.982	0.995	0.33	0.05
	Ridge, Kalman	0.982	0.999	0.988	0.979	0.999	0.33	0.09
4	HGB, raw	0.866	0.932	0.862	0.861	0.910	0.51	0.21
	HGB, Kalman	0.855	0.897	0.852	0.856	0.891	0.57	0.19
	Ridge, raw	0.898	0.964	0.915	0.896	0.961	0.52	0.07
	Ridge, Kalman	0.893	0.958	0.906	0.892	0.948	0.53	0.08

Note.B = price-only; C = price + weather; Gate 1 = B gated against A; Gate 2 = C gated against B; Cascade = C if Gate 2 open, else B if Gate 1 open, else A. W = 26 origins. Open rates are shares of origins at which the gate is open.

Table 6. Gate 2 Diagnostics: Does the Gate Open When Weather Helps?

h	Learner / input	Open rate	MAE ratio (open)	MAE ratio (closed)	Weather helps next 13 wks	Gate accuracy	Always-closed accuracy	Correlation
1	HGB, raw	0.13	0.983	1.010	0.50	0.55	0.50	0.03
	HGB, Kalman	0.07	1.036	0.987	0.55	0.45	0.45	-0.03
	Ridge, raw	0.05	1.023	1.007	0.42	0.57	0.58	-0.02
	Ridge, Kalman	0.09	0.965	1.010	0.39	0.61	0.61	0.08
4	HGB, raw	0.21	0.976	1.001	0.46	0.50	0.54	0.04
	HGB, Kalman	0.19	1.003	0.995	0.53	0.48	0.47	-0.01
	Ridge, raw	0.07	0.966	1.024	0.32	0.65	0.68	0.09
	Ridge, Kalman	0.08	0.985	1.018	0.35	0.62	0.65	0.05

Note.Open rate = share of origins at which Gate 2 (C against B) is open. MAE ratio = MAE of C divided by MAE of B on origins at which the gate is open or closed. "Weather helps next 13 wks" = share of origins at which C has lower MAE than B over the next 13 scored origins (forward-looking, used only to evaluate the gate). Gate accuracy = share of origins at which the gate state agrees with this benchmark; always-closed accuracy = 1 minus the benchmark share. Correlation = correlation between the gate state and the realised benefit of weather.

boosted trees at one week, Kalman-filled inputs improve accuracy modestly: relative MAE for B falls from 0.961 to 0.934 and for C from 0.967 to 0.924, and the latter difference is significant (DM = -2.69, p = 0.007), while the former is not (p = 0.119). At four weeks and for ridge regression, differences are small and not significant. Because the raw price is complete from 2012 and the gaps lie in 2003-2011, the filled series affects the training data and the early part of the evaluation window rather than recent forecasts, and the gain at one week for trees is consistent with better lag features during gap periods. Kalman filling does not substitute for gating: it improves the level forecasts of C and B similarly but leaves unchanged the weak relationship between gate state and the benefit of weather (Table 6), and Gate 1 remains costly with Kalman inputs.

Table 9 reports robustness to the gate window. Gate 2 remains open for only 5-26% of origins whatever the window, and its forecasts stay within about 0.6% of B in all cases (relative MAE to B between 0.994 and 1.004). Gate 1 stays costly at four weeks (relative MAE between 0.879 and 0.941 against 0.855-0.866 for always using B), and longer windows help it only slightly because B is genuinely better than naive. The conclusions do not depend on the window.

6. Discussion

Three conclusions emerge. First, price-only machine learning is genuinely useful for four-week-ahead potato price forecasting, reducing mean absolute error by 10-15% relative to a no-change forecast with high statistical confidence, while at one week the gain is small and not significant. This differs from the near-parity of complex models and persistence often seen at very short horizons in volatile markets and suggests that potato prices carry exploitable momentum and seasonal structure at the one-month horizon that gradient-boosted

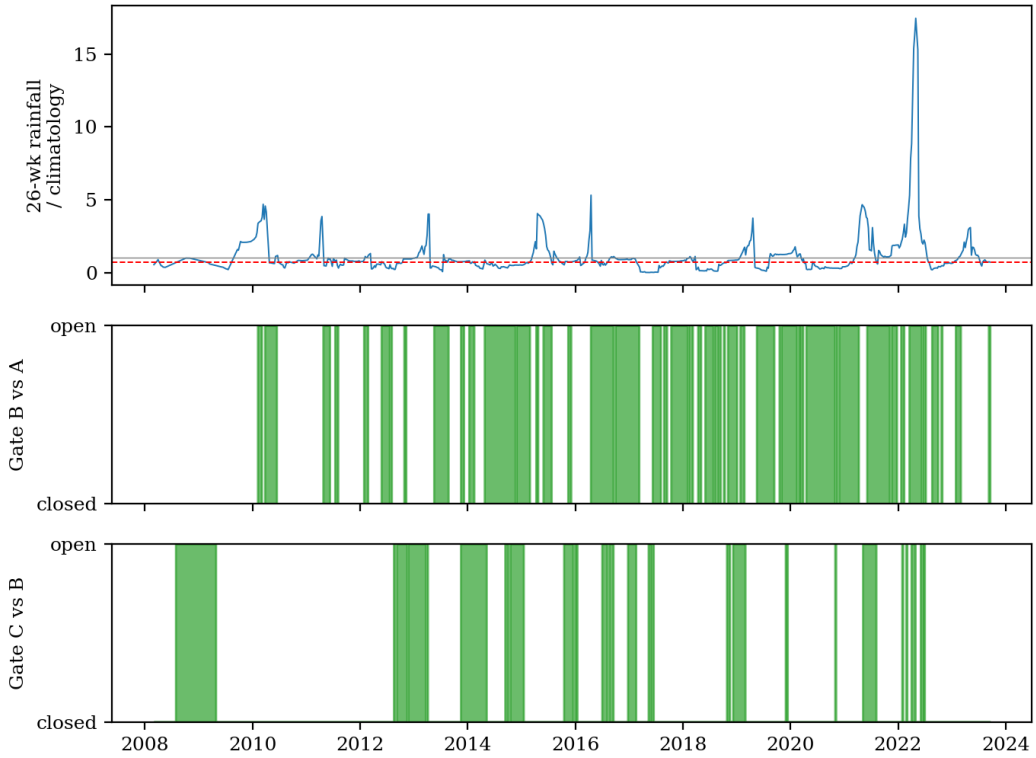


Figure 3. Four-week-ahead gate states for HGB with raw prices. Top: 26-week rainfall relative to same-season history (dashed line = 0.7 drought threshold). Middle: Gate 1 (B against A). Bottom: Gate 2 (C against B).

Table 7. MAE of Weather Model Relative to Price-Only Model, by Regime (C / B)

h	Learner / input	Dry	Normal	Wet	Monsoon	Non-monsoon
1	HGB, raw	0.991	0.992	1.084	1.019	1.001
	HGB, Kalman	0.967	1.004	1.024	0.978	0.995
	Ridge, raw	1.001	1.017	1.011	1.019	1.003
	Ridge, Kalman	1.003	1.011	1.009	1.014	1.004
4	HGB, raw	0.980	1.005	1.025	0.971	1.006
	HGB, Kalman	0.995	1.002	0.993	0.987	1.002
	Ridge, raw	1.026	1.019	1.000	1.018	1.020
	Ridge, Kalman	1.025	1.014	0.989	1.016	1.015

Note.Dry = 26-week rainfall below 70% of its same-season historical mean; wet = above 130%; normal otherwise. Monsoon = June–September. Values below 1 mean the weather model is more accurate. Regimes use past data only.

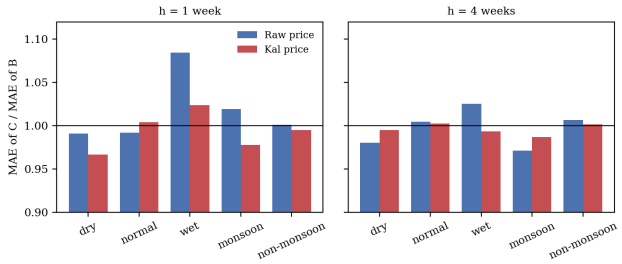


Figure 4. Figure 4. MAE of the weather model relative to the price-only model (HGB) by rainfall regime and season. Values below 1 favour the weather model.

trees can extract. Second, weather features do not improve average accuracy. This is consistent with findings that predictors with a sound economic rationale often fail to deliver out-of-sample gains (Goyal & Welch, 2008; Stock & Watson, 2003), and with a world in which storage and trade absorb much weather variation, so that weather matters only in unusual episodes. The deterioration for ridge regression is a reminder that additional regressors have an estimation cost.

Third, and most relevant to the title, the evidence for conditional value is suggestive but weak, and the gate does not detect it. Weather features were most useful in dry regimes and in the three drought years in the evaluation sample, and least useful or harmful in wet regimes and in other years. The gate between C and B closed nearly completely in wet regimes, where weather did harm, but opened no more often in dry than normal regimes, and its state was uncorrelated with the realised benefit of weather over the following 13 origins. The reasons are the ones the forecast-combination literature anticipates. A gate is a forecast of relative skill, and if the benefit of weather is small

Table 7b. Gate 2 Open Rates by Rainfall Regime (h = 1)

Regime	Weeks (h = 1)	Gate 2 open rate (HGB, raw)	Gate 2 open rate (HGB, Kalman)
dry	298	0.15	0.03
normal	274	0.16	0.12
wet	141	0.02	0.06
monsoon	245	0.14	0.07
non-monsoon	468	0.12	0.07

Note.Open rate = share of origins in the regime at which Gate 2 is open. Monsoon and non-monsoon rows classify by season rather than by rainfall.

Table 8. Raw versus Kalman-Filled Price Inputs

Comparison	Learner	Raw rel. MAE	Kalman rel. MAE	DM, Kalman vs raw (h = 1)	Raw rel. MAE	Kalman rel. MAE	DM, Kalman vs raw (h = 4)
B	HGB	0.961	0.934	-1.56 (.119)	0.866	0.855	-0.94 (.350)
C	HGB	0.967	0.924	-2.69 (.007)	0.862	0.852	-0.74 (.457)
B	Ridge	0.981	0.982	0.06 (.949)	0.898	0.893	-0.87 (.383)
C	Ridge	0.989	0.988	-0.10 (.923)	0.915	0.906	-1.43 (.154)

Note.Relative MAE is relative to the naive forecast. DM statistic compares the Kalman-input forecast with the raw-input forecast (negative values favour Kalman); p-values in parentheses.

(1-3%) relative to forecast-error noise, a window of 26 weekly errors cannot distinguish it reliably (Smith & Wallis, 2009; Claeskens et al., 2016). Drought regimes are also rare and slow-moving, so a 26-week window of realised forecast errors, which by construction lags the regime, may be uninformative about the next weeks. The substantive implication is that if weather has conditional value in potato prices, extracting it probably requires a regime variable based on the weather itself, such as the rainfall-anomaly index, rather than on realised forecast performance.

The B-versus-A gate illustrates the cost of a gate when the candidate model is genuinely better. Because B beats naive at four weeks by a large margin in nearly all sub-periods, a rule that opens only after recent proof of superiority forgoes accuracy at the times it is closed, and its closures are triggered by short noisy runs in which B happened to lose. This is a plain demonstration that switching rules based on short windows inherit estimation noise, and that for this problem the fixed choice (always use B at four weeks) dominates the gate.

The Kalman comparison shows that gap-filled inputs can improve forecasts for tree ensembles at one week, but they do not replace gating, because they change the level of the forecasts rather than the conditional value of weather. Two caveats attach to this result. The early part of the sample contains imperfect imputed values, and if the Kalman series was produced by a two-sided smoother over the full sample, filled values may embed future information, which could favour Kalman inputs. The modest gains should therefore be read as an upper bound pending verification of how the series was constructed.

Several limitations apply. The study covers a single potato series with a single weather station of unspecified location, and prices in a scaled unit; results may differ across markets. Only two learners and one set of fixed hyperparameters were used, and no other exogenous variables, such as arrivals, cold-storage stocks, area sown, import-export policy or fertiliser prices, were included. Weather was measured by simple lags and rolling summaries rather than by crop-stage-specific or agro-meteorological indices (for example, heat stress during tuber bulking). The drought-year contrast rests on three years, the regime thresholds (0.7 and 1.3) are conventional rather than estimated, and the tests are not adjusted for multiple comparisons or for the nested structure of the C-versus-B comparison. Finally, a gate defined by a dual criterion on short windows is one of many possible designs; smoother or probabilistic gating, or gating on the weather regime itself, might do better.

7. Conclusion

This paper recast the question of whether weather helps potato price forecasting as a question about conditional value and its detectability. On 1,080 weekly observations and about 15 years of walk-forward forecasts, price-only machine learning clearly beat a no-change forecast at four weeks (relative MAE 0.85-0.90) and weakly at one week, whereas adding lagged temperature and precipitation did not improve accuracy on average and worsened a linear model. Weather appeared to help in dry regimes and in the three drought years in the sample and to hurt in wet regimes, in the direction predicted by the hypothesis that weather matters mainly in drought, but the evidence was small in magnitude and statistically weak. A confidence gate built on recent level and turning-point performance did not detect these regimes: the C-versus-B gate

Table 9. Sensitivity of the Gates to the Look-Back Window (HGB)

Design	Gate 1 rel. to A (raw)	Gate 1 open	Gate 2 rel. to B (raw)	Gate 2 open	Gate 1 rel. to A (Kalman)	Gate 2 rel. to B (Kalman)	Gate 2 open (Kalman)
h = 1, W = 13	1.005	0.27	1.003	0.10	0.997	1.001	0.05
h = 1, W = 26	0.991	0.34	0.998	0.13	0.967	1.001	0.07
h = 1, W = 52	0.993	0.41	0.996	0.15	0.959	1.004	0.10
h = 4, W = 13	0.939	0.34	0.996	0.12	0.941	1.001	0.11
h = 4, W = 26	0.932	0.51	0.994	0.21	0.897	1.001	0.19
h = 4, W = 52	0.904	0.69	1.002	0.26	0.879	0.995	0.22

Note.Gate 1 relative MAE is relative to naive; Gate 2 relative MAE is relative to the price-only model B. W = number of realised origins in the look-back window.

stayed mostly closed and was no better than chance at identifying when weather would help, and the B-versus-A gate reduced accuracy relative to simply using the price-only model. Kalman-filled prices improved boosted-tree forecasts at one week modestly but did not substitute for gating.

For practice, the results suggest that analysts should first establish the gain from price-only machine learning over a naive benchmark, treat weather features as optional and test them against a price-only model within regimes rather than on average, and, if conditional value is suspected, build the regime indicator from weather data directly rather than infer it from a short record of forecast errors. A null result for the gate is not a failure of the question: it shows that conditional value, if present, is too small or too slow-moving to be picked up from realised forecast performance on a 26-week window. Future work should repeat the experiment across markets and crops, include crop-stage weather indices and market covariates, test regime effects formally with conditional predictive ability tests, and verify whether the Kalman series was filtered or smoothed.

References

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. Operational Research Quarterly, 20(4), 451-468.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

Cashin, P., Mohaddes, K., & Raissi, M. (2017). Fair weather or foul? The macroeconomic effects of El Niño. Journal of International Economics, 106, 37-54.

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 109(11), 1997-2028.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting, 32(3), 754-762.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics, 105(1), 85-110.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253-263.

Durbin, J., & Koopman, S. J. (2012). Time series analysis by state space methods. Oxford University Press.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. Journal of Applied Econometrics, 25(4), 595-620.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74(6), 1545-1578.

Gilbert, C. L., & Morgan, C. W. (2010). Food price volatility. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1554), 3023-3034.

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. The Review of Financial Studies, 21(4), 1455-1508.

Granger, C. W. J., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. Journal of Forecasting, 19(7), 537-560.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Advances in Neural Information Processing Systems, 35.

Harvey, A. C. (1989). Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13(2), 281-291.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

Jha, G. K., & Sinha, K. (2014). Time-delay neural networks for time series prediction: An application to the monthly wholesale price of oilseeds in India. Neural Computing and Applications, 24(3-4), 563-571.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35-45.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 30, 3146–3154.

Lobell, D. B., Schlenker, W., & Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. Science, 333(6042), 616-620.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3),

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54-74.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting, 38(4), 1346-1364.

Meese, R. A., & Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? Journal of International Economics, 14(1-2), 3-24.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. The R Journal, 9(1), 207-218.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Pesaran, M. H., & Timmermann, A. G. (1992). A simple nonparametric test of predictive performance. Journal of Business & Economic Statistics, 10(4), 461-465.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., Ziel, F. (2022). Forecasting: Theory and practice. International Journal of Forecasting, 38(3), 705-871.

Roberts, M. J., & Schlenker, W. (2013). Identifying supply and demand elasticities of agricultural commodities: Implications for the US ethanol mandate. American Economic Review, 103(6), 2265-2295.

Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott & A. Timmermann (Eds.), Handbook of economic forecasting, 2, 1203-1324.

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), 15594-15598.

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics, 71(3), 331-355.

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. Journal of Economic Literature, 41(3), 788-829.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437-450.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), Handbook of economic forecasting, 1, 135–196.

Tomek, W. G., & Peterson, H. H. (2001). Risk management in agricultural markets: A review. Journal of Futures Markets, 21(10), 953-985.

Wright, B. D. (2011). The economics of grain price volatility. Applied Economic Perspectives and Policy, 33(1), 32-58.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35-62.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.