PUBLIC HEALTH
OpenJournal
Openventio PUBLISHERS

Research

Two-Stage Zero Inflated Forecasting of Monthly Monsoon Rainfall: A Evaluation with Decision Focused Metrics

Sujith Arokia Swamy B[1,2], Raj Shekar M[1,2,*], Bhavana Kabber[3], Naveen Kumar P[3], Harish Nayak[4], B Samuel Naik[5], Karthik VC[6], Veershetty[5] and Ramya Jagadish[7]

[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]University of Agricultural Sciences, Dharwad, Karnataka, India – 580005
[5]Central Silk Board, Mysore, Karnataka, India – 570008
[6]Central Sericultural Research and Training Institute, Pampore, Jammu & Kashmir, India – 192121
[7]Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India – 641003

Corresponding Author
Raj Shekar M
Email: rajshekar4262@gmail.com

Article Information:
Received: 14 December 2025 | Revised: 11 January 2026 | Accepted: 13 January 2026 | Published: January 14, 2026

Cite this article:
Sujith Arokia Swamy B, Raj Shekar M, Bhavana Kabber, Naveen Kumar P, Harish Nayak, B Samuel Naik, Karthik VC, Veershetty and Ramya Jagadish(2026). Two-Stage Zero Inflated Forecasting of Monthly Monsoon Rainfall: A Evaluation with Decision Focused Metrics *Public Health Open Journal*. 11(S1):72–84. DOI: 10.17140/PHOJ.11.S1.72

Abstract
Monthly rainfall in north-west India is strongly seasonal, right-skewed and, in the dry season, zero-heavy, and it is usually evaluated by aggregate error measures that may hide failures in the months that matter for agriculture. This study proposes a decision-focused evaluation protocol and applies it to a 124-year series of monthly rainfall for the Haryana-Chandigarh-Delhi region (1,488 months, 1901-2024). A two-stage forecaster, which multiplies the probability of a wet month (rainfall above 5 mm) from a gradient-boosted classifier by the expected amount in wet months from a gradient-boosted regressor, is compared with nine benchmarks (persistence, seasonal naive, monthly climatology, ridge regression, a Tweedie generalised linear model, a random forest, single-stage boosting with squared and Poisson loss, and a two-stage linear model) in an expanding-window, one-month-ahead evaluation over 1926-2024 (1,188 forecasts) with annual re-estimation. Accuracy is assessed by MAE and RMSE on all, monsoon and wet months, by the Brier skill score and calibration for occurrence, by drought-month recall, flood-month precision and onset detection, and by stability across 25-year blocks, with Diebold-Mariano tests and a block bootstrap. The main finding is negative and informative: no model beats the climatological benchmark. The two-stage boosted model has an MAE 7.3% above climatology (bootstrap interval 1.031–1.110) and a Jun-Sept MSE skill of -0.152, and is statistically indistinguishable from single-stage boosting (Diebold-Mariano p = 0.83 for absolute error). The best-calibrated probabilistic forecast is climatology itself (Brier skill scores of all other forecasters are negative). Decision-focused metrics rank models very similarly to aggregate metrics, drought recall (0.32-0.46 at a 20% alarm rate) and flood precision (0.15-0.25 at a 15% alarm rate) differ among models by amounts well inside bootstrap uncertainty, and onset predictions do no better than always predicting July. Monthly anomalies have almost no autocorrelation (0.035 at lag 1), and the calendar month alone explains 63% of the variance; the remainder is not predictable from past rainfall alone. Skill relative to climatology does not change systematically across 25-year blocks, and there is no significant trend in monsoon rainfall (Kendall τ = -0.02, p = 0.74). The results show that the benefit of two-stage structure depends on the presence of predictable occurrence and amount signals, which are absent in rainfall-history-only forecasting at monthly resolution, and they provide a reproducible benchmark and protocol for evaluating richer predictors such as sea-surface temperature or other states' rainfall.

Keywords: *zero-inflated time series; monsoon rainfall; two-stage forecasting; decision-focused evaluation; climatology benchmark; Brier skill score; gradient boosting*

1. Introduction

Agriculture in Haryana and the National Capital Region depends on the summer monsoon. Whether the rains arrive on time, whether they are adequate, and whether late-season rainfall is sufficient for crop completion drive sowing dates, irrigation scheduling and risk management for the kharif season (Gadgil and Gadgil, 2006). Operational forecasts of the Indian monsoon are mostly seasonal and probabilistic (Gadgil et al., 2005; Rajeevan et al., 2007), but decisions are taken month by month, and a forecaster who can say something about the next month's rainfall from the record of past months would be valuable if such skill existed.

Monthly rainfall is awkward for standard forecasting methods. It is non-negative and right-skewed, it has a strong seasonal cycle, and in dry months a large share of observations are close to zero. A model trained by minimising squared error on all months estimates a conditional mean, which for a skewed, seasonal series is dominated by the seasonal cycle and can understate peaks and overstate dry months. A natural response in hydrology is the occurrence-and-amount model, in which a first stage predicts whether a wet period occurs and a second stage predicts its amount (Stern and Coe, 1984; Woolhiser and Roldán, 1982). Evaluation, however, usually relies on MAE and RMSE over all months, which are dominated by the seasonal cycle and say little about the questions a farmer asks: will the monsoon be late, will a drought month occur, and will a flood-scale month occur?

This paper makes three contributions. It specifies a decision-focused evaluation protocol for zero-inflated monthly rainfall that combines level accuracy, probabilistic occurrence skill against climatology (the Brier skill score and calibration), event-based metrics for droughts, floods and monsoon onset, and stability across 25-year blocks. It applies the protocol to a 124-year series with a transparent, reproducible, expanding-window design, and compares a two-stage boosted model with nine benchmarks whose hyperparameters are listed in full. And it reports the result honestly, which is that on this series nothing beats monthly climatology, for reasons that the data explain. The negative result is informative because it identifies what a rainfall forecaster must add to a history-only model, and because the protocol and benchmarks provide a standard against which richer predictors can be evaluated. The study uses the Haryana-Chandigarh-Delhi series; data for other states are available to the author and are the natural extension.

The remainder of the paper reviews the literature (Section 2), describes the data (Section 3), sets out the methodology including the hyperparameters of all models (Section 4), reports results (Section 5), discusses them (Section 6) and concludes (Section 7).

2. Review of Literature

Occurrence-and-amount models have long been used for precipitation, usually at daily resolution, where a Markov chain governs wet and dry days and a skewed distribution, often gamma, governs amounts (Stern and Coe, 1984; Woolhiser and Roldán, 1982). The same logic carries over to machine-learning forecasters, which can model the occurrence probability with a classifier and the amount with a regressor trained on wet months only. A related family of models treats the zero-heavy response directly: the Tweedie compound Poisson–gamma distribution allows exact zeros with a continuous positive part (Jørgensen, 1987; Dunn and Smyth, 2005), and Poisson-loss gradient boosting handles non-negative, zero-heavy targets without a separate occurrence model. These approaches are natural benchmarks for a two-stage design, and the present study includes both.

Forecast verification has a mature methodology. Probabilistic occurrence forecasts are scored with the Brier score (Brier, 1950), which can be decomposed into reliability and resolution (Murphy, 1973), and with skill scores that measure improvement over a reference forecast such as climatology (Murphy, 1992; Murphy and Winkler, 1987; Wilks, 2011). Proper scoring rules and probabilistic forecasting are reviewed by Gneiting and Raftery (2007) and Gneiting and Katzfuss (2014), and practical guidance for weather and climate verification is given by Jolliffe and Stephenson (2012). A key lesson of that literature is that climatology is the minimum standard of reference: a forecast that cannot beat the long-term average for the calendar month adds no information. A second lesson is that skill measures should be tied to user decisions, so that event-based metrics such as hit rates and false-alarm ratios complement aggregate errors.

For the Indian monsoon, forecasting is a mature but difficult field. Seasonal monsoon rainfall is partly predictable from large-scale predictors such as sea-surface temperature, but operational models have had notable failures (Gadgil et al., 2005), and statistical long-range forecast models rely on external predictors (Rajeevan et al., 2007). Predictability at monthly lead times from the rainfall record alone is limited by the chaotic nature of the atmosphere (Lorenz, 1963). Studies of long-term monsoon change report mixed and sometimes contradictory results, with some attributing a weakening of the monsoon to anthropogenic aerosols (Bollasina

et al., 2011) and others emphasising internal variability and the difficulty of detecting trends (Turner and Annamalai, 2012). Trend tests for hydrological series commonly use the Mann–Kendall test and Sen's slope estimator (Mann, 1945; Sen, 1968).

For statistical comparison of forecasts, the Diebold–Mariano test is standard (Diebold and Mariano, 1995; Harvey et al., 1997) with heteroskedasticity and autocorrelation-consistent variance estimates (Newey and West, 1987), and uncertainty in verification measures is commonly assessed by block bootstrap (Künsch, 1989; Politis and Romano, 1992). Gradient-boosted trees (Friedman, 2001), random forests (Breiman, 2001) and regularised regression (Hoerl and Kennard, 1970) are standard learners for tabular features, implemented in widely used software (Pedregosa et al., 2011). The gap addressed here is a transparent evaluation of a two-stage rainfall forecaster against a full set of benchmarks, with decision-focused and stability analyses, over a long record.

3. Data Description

The data are average monthly rainfall (mm) for the Haryana-Chandigarh-Delhi region, supplied as a single column of 1,488 values with month labels of the form "Jan-01". The two-digit year labels repeat (the label "Jan-01" appears at the first row and at row 1,201), so the dates were reconstructed from the row order, beginning in January 1901 and ending in December 2024, and the reconstruction was checked against the labels: the month and the two-digit year of all 1,488 rows agree with the reconstructed date (labels run 01, 02, …, 99, 00, 01, …, 24). There are no missing values and no negative values. The data source and the method by which the regional average was constructed are not documented in the file and should be specified before submission.

Table 1 summarises the series. Mean monthly rainfall is 44.6 mm and mean annual rainfall 534.8 mm. The distribution is strongly right-skewed (skewness 2.10), with a median of 14.1 mm and a maximum of 404.7 mm (August 1995), followed by 398.2 mm (August 1908). The extent of zero inflation is more moderate than often assumed: 7.0% of months have exactly zero rainfall, 17.8% have less than 1 mm, and 32.5% have less than 5 mm, the wet–dry threshold used below. The monsoon dominates: June to September account for 82.3% of annual rainfall and July to September for 73.0%. Table 2 gives the monthly climatology. July and August are the wettest months (means of 153.5 and 148.7 mm) and almost never dry, November is the driest (mean 3.3 mm, below 5 mm in 81% of years), and June is the most variable transition month, with a mean of 49.7 mm, a standard deviation of 34.5 mm and rainfall below 5 mm in 1% of years.

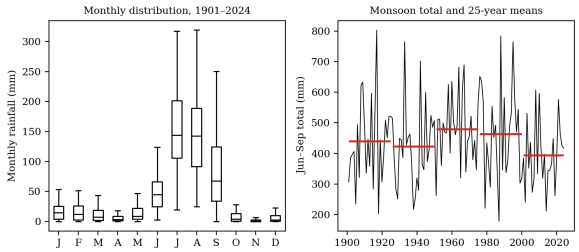


Figure 1. Left: distribution of monthly rainfall by calendar month. Right: June–September total rainfall by year (black) and means for 25-year blocks (red).

Two features of the series shape everything that follows. First, calendar month explains 63.3% of the variance of monthly rainfall, so a forecaster that knows only the month already accounts for most of the variation. Second, the anomalies that remain after removing the monthly mean are almost uncorrelated through time: the autocorrelation is 0.035 at lag 1, 0.007 at lag 2, 0.005 at lag 3 and -0.001 at lag 12. There is thus little persistence on which a history-only forecaster could build. The brief for this study anticipated that roughly half of all months would have near-zero rainfall and that June to September would account for 70-80% of annual rainfall; the first is not borne out by the data (about 18% of months are below 1 mm and 33% below 5 mm), and the second is slightly understated (82% for June to September).

4. Methodology

The task is to forecast rainfall one month ahead. At each forecast origin t (the end of month t), a model forecasts y(t+1). Evaluation uses an expanding window. The initial training sample is 1901–1925 (25 years), and forecasts are made for every month from January 1926 to December 2024 (1,188 forecasts over 99 years). Models are re-estimated once a year, at the end of each December, using only rows whose target month had been observed by that date, and the resulting model is used for the next 12 forecasts; this mimics a forecaster who updates the model annually. The 25-year burn-in means that the first 25-year block, 1901-1925, is used for training only and cannot be evaluated out of sample; stability is therefore assessed over the four later blocks (1926-1950, 1951-1975, 1976-2000, 2001-2024).

The proposed two-stage forecaster decomposes the forecast into occurrence and amount. Stage 1 estimates the probability p(t+1) = P(y(t+1) > 5 mm) with

Table 1. Descriptive Statistics of Monthly Rainfall, Haryana-Chandigarh-Delhi (1901-2024)

Statistic	Value
Monthly observations	1,488
Mean monthly rainfall (mm)	44.6
Standard deviation (mm)	66.3
Median (mm)	14.1
Maximum (mm)	404.7
Skewness	2.10
Months with rainfall exactly 0 mm	7.0%
Months with rainfall < 1 mm	17.8%
Months with rainfall < 5 mm (dry by the 5 mm rule)	32.5%
Mean annual rainfall (mm)	534.8
Share of annual rainfall in Jun–Sep	82.3%
Share of annual rainfall in Jul–Sep	73.0%
Month-of-year share of variance (R^2)	0.634
Autocorrelation of monthly anomalies, lag 1 / lag 2 / lag 3	0.035 / 0.007 / 0.005

Note. Anomalies are deviations from the calendar-month mean. R^2 = share of the variance of monthly rainfall explained by calendar month.

Table 2. Monthly Climatology

Month	Mean (mm)	SD (mm)	Share < 5 mm	Share < 1 mm	Maximum (mm)
Jan	17.4	16.1	0.30	0.13	70.3
Feb	17.1	19.1	0.32	0.16	91.2
Mar	13.5	16.1	0.39	0.19	71.6
Apr	7.6	12.6	0.60	0.27	82.3
May	15.6	16.4	0.30	0.06	73.6
Jun	49.7	34.5	0.01	0.00	200.0
Jul	153.5	70.0	0.00	0.00	369.4
Aug	148.7	77.5	0.00	0.00	404.7
Sep	88.4	73.2	0.02	0.01	361.8
Oct	12.7	23.6	0.56	0.33	132.1
Nov	3.3	7.2	0.81	0.60	43.4
Dec	7.2	10.7	0.58	0.40	68.9

Note. Computed over 1901–2024 (124 values per month

a gradient-boosted classifier. Stage 2 estimates the conditional mean amount $m(t+1) = E[y(t+1) \mid y(t+1) > 5$ mm] with a gradient-boosted regressor trained only on months with rainfall above 5 mm. The point forecast is $\hat{y}(t+1) = p(t+1) \times m(t+1)$. Both stages use the same 16 predictors: the month of the target (an integer index treated as categorical by the tree models, and sine and cosine terms), rainfall of the current month and of the two preceding months, rainfall of the same calendar month a year earlier, rolling means and standard deviations over the latest 3, 6 and 12 months, the total rainfall of the most recent completed June-September season, the expanding-window climatological mean of the target month computed from data available at the origin, and a year index scaled to 0-1 for the century. The implementation uses the histogram-based gradient boosting estimators in scikit-learn (Pedregosa et al., 2011), which belong to the same family as Light GBM (Ke et al., 2017) but are not Light GBM itself, as the Light GBM library was not available in the computing environment.

Table 3 lists the benchmark models and, for every model, the hyperparameters. The benchmarks are of four kinds: naive rules (persistence, seasonal naive and monthly climatology, the last with an expanding window); single-stage learners trained on all months (ridge regression, a Tweedie generalised linear model, a random forest, and boosting with squared-error and with Poisson loss); a two-stage linear model (logistic regression for occurrence and a gamma generalised linear model for wet-month amounts); and the proposed two-stage boosted model. Hyperparameters were fixed in advance and were not tuned on the test data, so that the comparison cannot be affected by selection on the evaluation sample; the consequences of this choice are discussed in Section 6. Random seeds are fixed at 0. For the occurrence forecasts, the probabilistic competitors are the climatological occurrence rate for the month, the seasonal-naive indicator, logistic regression, a random-forest classifier and the boosted

classifier.

Accuracy is evaluated at four levels. Level accuracy is MAE and RMSE on all months, on monsoon months (June to September) and on wet months (rainfall above 5 mm), together with MAE relative to climatology and a mean-squared-error skill score, 1- MSE (model)/ MSE (climatology), computed on June to September; positive values mean the model beats climatology. Occurrence accuracy is the Brier score for P(y > 5 mm), the Brier skill score against the climatological occurrence rate (BSS, positive is better) for all months, for October to May and for June to September, and the expected calibration error (ECE) over ten equal-width probability bins, with reliability diagrams.

Decision metrics are defined on the 396 June to September target months. A drought month is a monsoon month with rainfall below 20 mm (37 events), a flood month one with rainfall above 200 mm (52 events). A forecast of the amount is turned into a warning by a fixed-alarm-rate rule, because forecasts that are conditional means rarely fall below 20 mm or exceed 200 mm in a monsoon month: a drought warning is issued for the 20% of monsoon months with the lowest forecasts, and a flood warning for the 15% with the highest. Drought recall is the share of drought months that were warned about, flood precision the share of flood warnings that were followed by a flood month, and the area under the ROC curve (AUC) of the forecast for each event is reported as a threshold-free summary. The literal thresholds (forecast above 200 mm, forecast below 20 mm) are also reported as the number of months in which they are triggered. Onset is defined for each of the 99 test years as the first month from June to September with rainfall above 50 mm; the predicted onset is the first month in which the one-month-ahead forecast exceeds 50 mm. Onset is evaluated by the exact month, by being within one month, and by classification of a late onset (June rainfall at or below 50 mm).

Stability is assessed by computing the main metrics separately for each 25-year block and by Mann-Kendall trend tests with Sen slopes for monsoon and monthly rainfall, together with Levene and Kruskal-Wallis tests across blocks. Statistical inference uses the Diebold-Mariano test of equal predictive accuracy with the small-sample correction of Harvey et al. (1997) and Bartlett-kernel variances (lag 6 for all months and lag 3 for the monsoon subset; Newey and West, 1987), applied to absolute and squared errors, with a positive statistic meaning that the first forecast in a pair has the larger loss. Uncertainty in the main metrics is assessed by a circular block bootstrap over the 99 test years with blocks of three years and 2,000 resamples (Kunsch, 1989; Politis and Romano, 1992), recomputing each metric from year-level sums; the intervals are 95% percentile intervals. No multiplicity correction is applied, and tests are descriptive.

5. Results

Table 4 and Figure 2 give the level-accuracy comparison. Monthly climatology is the best forecaster on every measure. Over all months its MAE is 23.65 mm and its RMSE 40.05 mm; for the monsoon months its RMSE is 65.78 mm. No other model improves on these values. The learners are within 1% to 7% of climatology in MAE and between 1% and 15% worse in monsoon-month mean squared error: ridge regression (MAE ratio 1.029), the Tweedie GLM (1.028), the random forest (1.056), Poisson boosting (1.051), single-stage boosted regression (1.070), the two-stage linear model (1.010) and the proposed two-stage boosted model (1.073). Persistence and seasonal naive forecasts are much worse (MAE ratios of 1.78 and 1.34), because they carry last months or last year's noise rather than an average. The two-stage boosted model has a June-September MSE skill of -0.152, the worst among the learners, and the best of them, ridge regression, has -0.025. These conclusions are the same for wet-month accuracy (Table 4, columns for wet months), so the zero-inflated structure of the data is not driving them.

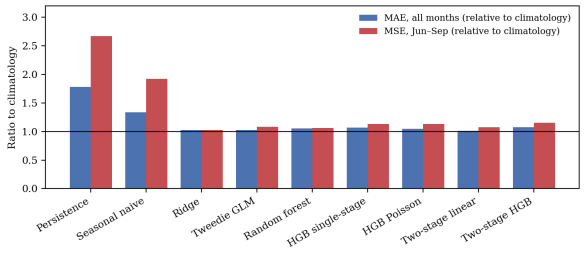


Figure 2. MAE over all months and mean squared error over June–September, relative to climatology (values above 1 are worse than climatology).

The block-bootstrap intervals in Table 5 show how decisive the ordering is. The MAE ratio of the two-stage boosted model has a 95% interval of 1.031 to 1.110, which excludes one; so do the intervals of the single-stage boosted model (1.034 to 1.106) and the random forest (1.030 to 1.081). The only learner whose MAE interval includes one is the two-stage linear model (0.989 to 1.031), which is therefore not distinguishable from climatology in absolute error, although its monsoon MSE skill is significantly negative (-0.076, interval -0.130 to -0.022). The Diebold–Mariano tests in Table 6 agree: the two-stage boosted model has a significantly larger

Table 3. Models, Roles and Hyperparameters

Model	Role	Specification and hyperparameters
Persistence	Benchmark	Forecast = rainfall of the current month, $\hat{y}(t+1) = y(t)$. No parameters.
Seasonal naive	Benchmark	Forecast = rainfall of the same calendar month one year earlier, $y(t+1-12)$. No parameters.
Climatology	Benchmark	Expanding-window mean of the target calendar month using only data up to the forecast origin; occurrence probability = past share of months above 5 mm.
Ridge regression	Benchmark (single-stage, linear)	L2 penalty α = 10; standardised features with 12 month indicators in place of the month index; predictions truncated at 0.
Tweedie GLM	Benchmark (single-stage, zero-inflated aware)	Power 1.5 (compound Poisson–gamma), log link, L2 penalty α = 0.1, up to 2,000 iterations, standardised features with month indicators.
Random forest	Benchmark (single-stage, trees)	200 trees, minimum leaf size 5, 50% of features per split, bootstrap sampling, seed 0; trained on all months with squared error.
HGB single-stage	Benchmark (single-stage, boosting)	Histogram gradient-boosted regression trees (scikit-learn): 200 iterations, learning rate 0.05, maximum depth 3, minimum leaf 15, L2 penalty 1.0, 255 bins, month index as categorical feature, squared-error loss, seed 0; trained on all months.
HGB Poisson	Benchmark (single-stage, boosting)	As HGB single-stage but with Poisson loss (non-negative, suited to zero-heavy data).
Two-stage linear	Benchmark (two-stage)	Occurrence: logistic regression, L2 penalty C = 1.0, up to 1,000 iterations. Amount: gamma GLM with log link, α = 0.1, trained on wet months only. Forecast = $P \times \hat{E}$.
Two-stage HGB	Proposed	Stage 1 occurrence: HGB classifier, log-loss, 200 iterations, learning rate 0.05, depth 4, minimum leaf 20, L2 1.0, month categorical. Stage 2 amount: HGB regressor trained only on months with rainfall above 5 mm, 200 iterations, learning rate 0.05, depth 3, minimum leaf 15, L2 1.0. Forecast = $P(y > 5) \times \hat{E}(y \mid y > 5)$.
Occurrence classifiers (Brier)	Probabilistic benchmarks	Climatological rate; seasonal-naive indicator; logistic regression (as above); random-forest classifier (200 trees, minimum leaf 5, 50% of features); HGB classifier (as Stage 1).

Note. All learners use the 16 predictors described in the text. Linear models use standardised predictors and 12 months indicators instead of the month index. HGB = histogram gradient boosting. All learners are re-estimated every December on data realised by then.

loss than climatology for all months and for the monsoon months under both absolute and squared error (statistics of 4.25, 3.50, 3.27 and 3.28, all $p < .001$), and a significantly smaller loss than seasonal naive (statistics from -5.77 to -6.65). Importantly, the two-stage boosted model is not different from single-stage boosting (statistics 0.22 for absolute and 1.00 for squared error over all months, $p = 0.83$ and .32; 0.32 and 0.92 for the monsoon, $p = 0.75$ and 0.36), so the two-stage decomposition neither helps nor hurts relative to a single-stage model of the same type. It does significantly worse than the two-stage linear model (statistic 3.73, $p < 0.001$, in absolute error) than the random forest in squared error (2.89, $p = 0.004$).

Occurrence forecasts tell the same story (Table 7, Figure 3). The Brier score of the climatological occurrence rate is 0.152 over all months and 0.224 for October to May, and no other forecaster beats it: the Brier skill

Table 4. Level Accuracy of One-Month-Ahead Forecasts, 1926-2024 (mm)

Model	MAE all	RMSE all	MAE Jun–Sep	RMSE Jun–Sep	MAE wet	RMSE wet	MAE rel. to clim.	Jun–Sep MSE skill
Persistence	42.14	70.57	82.82	107.56	52.17	79.79	1.782	-1.673
Seasonal naive	31.63	55.24	68.77	91.20	42.47	66.17	1.337	-0.922
Climatology	23.65	40.05	50.12	65.78	30.54	48.00	1.000	0.000
Ridge	24.34	40.69	50.97	66.60	31.27	48.60	1.029	-0.025
Tweedie GLM	24.31	41.50	52.11	68.36	31.61	49.75	1.028	-0.080
Random forest	24.98	41.15	52.55	67.79	31.80	49.10	1.056	-0.062
HGB single-stage	25.31	42.50	53.66	69.94	32.72	50.90	1.070	-0.130
HGB Poisson	24.84	42.43	53.20	70.06	32.37	50.89	1.051	-0.134
Two-stage linear	23.88	41.44	51.21	68.24	31.29	49.77	1.010	-0.076
Two-stage HGB	25.37	42.93	53.85	70.62	33.11	51.50	1.073	-0.152

Note. All months n = 1,188; Jun–Sep n = 396; wet months = months with rainfall above 5 mm. MAE rel. to clim. = MAE over all months divided by that of climatology. MSE skill = 1 - MSE/MSE of climatology over Jun–Sep (positive is better).

Table 5. Block-Bootstrap 95% Intervals for Level-Accuracy Measures

Model	MAE all (95% CI)	MAE relative to climatology (95% CI)	Jun–Sep MSE skill (95% CI)
Climatology	23.65 [22.51, 24.76]	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
Seasonal naive	31.63 [29.34, 33.93]	1.337 [1.266, 1.414]	-0.922 [-1.189, -0.677]
Random forest	24.98 [23.61, 26.20]	1.056 [1.030, 1.081]	-0.062 [-0.116, -0.010]
HGB single-stage	25.31 [23.77, 26.74]	1.070 [1.034, 1.106]	-0.130 [-0.214, -0.037]
Two-stage linear	23.88 [22.56, 25.13]	1.010 [0.989, 1.031]	-0.076 [-0.130, -0.022]
Two-stage HGB	25.37 [23.86, 26.73]	1.073 [1.031, 1.110]	-0.152 [-0.243, -0.055]

Note. Circular block bootstrap over 99 test years (3-year blocks, 2,000 resamples). Values in brackets are 95% percentile intervals.

scores for October to May are -0.024 for logistic regression (interval -0.038 to -0.009), -0.056 for the random forest (-0.085 to -0.027) and -0.138 for the boosted classifier (-0.188 to -0.089), and -0.877 for the seasonal-naive indicator. The reliability diagram shows why. Climatology is well calibrated by construction (ECE 0.028), logistic regression and the random forest are slightly over-dispersed (ECE 0.046 and 0.034), and the boosted classifier, which fits the training data more closely, is the least well calibrated of the learned models (ECE 0.072), assigning probabilities that are more extreme than the observed frequencies justify. For June to September, rain above 5 mm occurs in about 99% of months, so the Brier scores are tiny (0.008) and skill scores there are not informative.

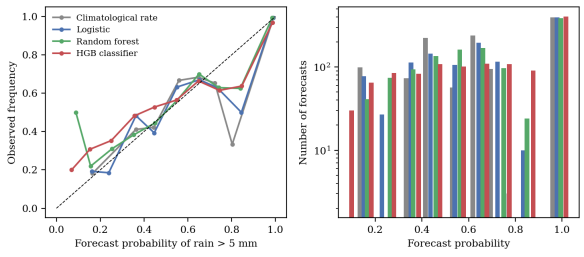


Figure 3. Reliability of occurrence forecasts (left) and distribution of forecast probabilities (right, log scale).

Decision-focused metrics (Table 8) do not reverse the ranking. Among the 396 monsoon months, 37 are drought months (rainfall below 20 mm), of which 19 fall in June, 16 in September and only two in July; the drought events are therefore late onsets and early withdrawals of the monsoon rather than mid-season failures. With a 20% alarm rate, drought recall is 0.405 for the two-stage boosted model, 0.405 for the single-stage boosted model, 0.459 for the random forest, 0.378 for

Table 6. Diebold–Mariano Statistics (p-values in parentheses)

Comparison (first vs second)	All months, abs. error	All months, sq. error	Jun–Sep, abs. error	Jun–Sep, sq. error
Two-stage HGB vs Climatology	4.25 (< .001)	3.50 (< .001)	3.27 (.001)	3.28 (.001)
Two-stage HGB vs HGB single-stage	0.22 (.827)	1.00 (.320)	0.32 (.750)	0.92 (.356)
Two-stage HGB vs Random forest	1.19 (.234)	2.89 (.004)	1.40 (.161)	2.56 (.011)
Two-stage HGB vs Seasonal naive	-6.65 (< .001)	-5.94 (< .001)	-5.81 (< .001)	-5.77 (< .001)
Two-stage HGB vs Two-stage linear	3.73 (< .001)	1.90 (.058)	2.26 (.024)	1.62 (.106)
HGB single-stage vs Climatology	4.18 (< .001)	3.27 (.001)	3.22 (.001)	3.10 (.002)
Random forest vs Climatology	4.81 (< .001)	2.30 (.021)	3.29 (.001)	2.36 (.019)
Two-stage linear vs Climatology	0.97 (.332)	3.05 (.002)	1.56 (.119)	2.96 (.003)
Ridge vs Climatology	4.72 (< .001)	3.27 (.001)	2.69 (.007)	2.37 (.018)

Note. Positive statistics indicate that the first forecast has the larger loss. Lag 6 for all months, lag 3 for June–September.

Table 7. Occurrence Forecasts: Brier Skill Score and Calibration

Occurrence forecaster	Brier, all months	BSS, all months	BSS, Oct–May (95% CI)	BSS, Jun–Sep	Calibration error (ECE)
Climatological rate	0.152	0.000	0 (reference)	0.000	0.028
Seasonal naive	0.285	-0.879	-0.877 [-1.007, -0.755]	-0.994	0.285
Logistic	0.156	-0.024	-0.024 [-0.038, -0.009]	-0.022	0.046
Random forest	0.160	-0.056	-0.056 [-0.085, -0.027]	-0.078	0.034
HGB classifier	0.173	-0.137	-0.138 [-0.188, -0.089]	-0.062	0.072

Note. BSS = Brier skill score relative to the climatological occurrence rate; values below zero mean worse than climatology. ECE = expected calibration error over 10 probability bins. Interval = block-bootstrap 95% interval.

the two-stage linear model and 0.324 for climatology. These differences are small relative to sampling uncertainty: the bootstrap interval for the two-stage boosted model is 0.237 to 0.576, and that of climatology is 0.179 to 0.484, so the data cannot distinguish them. The AUC for droughts tells a more coherent story, with the highest values for the two-stage linear model (0.735), ridge regression (0.727) and climatology (0.718), and the two-stage boosted model at 0.709. For floods (52 events: 28 in July, 20 in August, four in September), flood precision at a 15% alarm rate lies between 0.150 and 0.250 for all models (0.217 for the two-stage boosted model and 0.217 for climatology, against a base rate of 13.1%), so flood warnings from any of these models are only slightly better than random alarms; the AUCs are 0.757 (random forest), 0.747 (climatology) and 0.710 (two-stage boosted). The literal thresholds are informative on their own: climatology and the two-stage linear model never forecast more than 200 mm, and the boosted models do so in 5 to 16 of 396 months, while forecasts below 20 mm occur in only two months for the two-stage boosted model. A decision rule based on absolute thresholds applied to conditional-mean forecasts would therefore issue almost no warnings, which supports the use of rank-based or probabilistic warnings.

Onset results (Table 9) show that the test has little discriminating power in this series. In 42 of the 99 test years the monsoon started in June (June rainfall above 50 mm), in 54 in July and in three in August; there are no years without onset. Being within one month is almost guaranteed, with every model at 96-100%, because the onset is always June, July or August. The exact onset month is predicted correctly in 47.5% of years by the two-stage boosted model, 51.5% by climatology and 53.5% by the two-stage linear model, but always predicting July would be correct in 54.5% of years. For the classification of a late onset (June rainfall at or below 50 mm, 57.6% of years), accuracy is 48.5% for the two-stage boosted model, 53.5% for climatology and 55.6% for the two-stage linear model, against 57.6% for always predicting a late onset. Persistence, which predicts a late onset in most years, attains the highest late-onset F1 (0.720) only by predicting lateness almost always. No model can anticipate the timing of the onset from the preceding months' rainfall.

As it is seen the models fail to beat climatology, Table 10 gives diagnostics. The anomaly correlation between forecast and observed departures from climatology is between -0.08 and 0.05 for all models over June to September, with the two-stage boosted model at -0.006, so forecasts contain essentially no information

Table 8. Drought and Flood Warnings in Monsoon Months

Model	Drought recall	Drought precision	Drought AUC	Flood precision	Flood recall	Flood AUC	Months forecast > 200 mm	Months forecast < 20 mm
Persistence	0.405	0.188	0.614	0.150	0.173	0.538	48	89
Seasonal naive	0.378	0.175	0.648	0.250	0.288	0.686	53	38
Climatology	0.324	0.150	0.718	0.217	0.250	0.747	0	0
Ridge	0.324	0.150	0.727	0.250	0.288	0.739	0	0
Random forest	0.459	0.212	0.722	0.233	0.269	0.757	2	0
HGB single-stage	0.405	0.188	0.705	0.217	0.250	0.722	5	2
HGB Poisson	0.351	0.163	0.708	0.250	0.288	0.736	16	0
Two-stage linear	0.378	0.175	0.735	0.200	0.231	0.730	0	0
Two-stage HGB	0.405	0.188	0.709	0.217	0.250	0.710	8	2

Note. Drought = Jun–Sep rainfall below 20 mm (37 events); flood = above 200 mm (52 events). Warnings: lowest 20% of forecasts (drought) and highest 15% (flood) among 396 Jun–Sep months. AUC = area under the ROC curve for the forecast as a score. The last two columns count months in which the literal thresholds are triggered.

Table 8B. Bootstrap Intervals for Drought Recall and Flood Precision

Model	**Drought recall (95% CI)**	**Flood precision (95% CI)**
Climatology	0.324 [0.179, 0.484]	0.217 [0.125, 0.299]
Seasonal naive	0.378 [0.222, 0.544]	0.250 [0.148, 0.348]
Random forest	0.459 [0.289, 0.628]	0.233 [0.123, 0.339]
HGB single-stage	0.405 [0.243, 0.571]	0.217 [0.122, 0.302]
Two-stage linear	0.378 [0.226, 0.526]	0.200 [0.097, 0.292]
Two-stage HGB	0.405 [0.237, 0.576]	0.217 [0.120, 0.316]

Note. Block-bootstrap 95% intervals (3-year blocks, 2,000 resamples).

beyond the monthly mean. The regression slope of observed on forecast anomalies is near zero or negative for most models, and in flood months (observed mean 261.3 mm) the forecasts average 134-157 mm for the learners, so the peaks are not anticipated by any model. The underlying reason is that the within-month variation is not predictable from past rainfall: the lagged autocorrelations of anomalies are close to zero (Table 1). A flexible learner trained on such data fits noise, and the resulting forecasts are worse than the stable climatological mean by the amount of the estimation error.

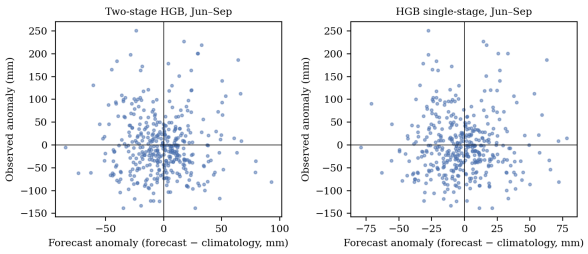


Figure 4. Observed versus forecast anomalies in June–September for two-stage and single-stage boosting. The cloud is flat, so forecast anomalies carry no information about observed anomalies.

Stability across the record is shown in Tables 11-14 and Figure 5. The monsoon (June to September) total has no significant trend (Kendall τ = -0.02, p = 0.74; Sen slope -0.12 mm per year), nor does annual rainfall (τ = -0.01, p = 0.83), July, August or September rainfall (all p > 0.80), or June rainfall (τ = 0.09, p = 0.16). The mean June-September total was 440 mm in 1901-1925, 423 mm in 1926-1950, 479 mm in 1951-1975, 464 mm in 1976-2000 and 394 mm in 2001-2024; the last block is the lowest and has the smallest dispersion, but the variation among blocks is not significant (Kruskal-Wallis p = 0.12; Levene test for equal variances p = 0.26). Model skill does not change systematically across blocks. The two-stage boosted model has a June-September MSE skill of -0.139, -0.131, -0.266 and -0.062 in the four blocks, and single-stage boosting -0.037, -0.153, -0.246 and -0.079; the two models do not differ significantly in any block (Diebold-Mariano p from 0.28 to 0.57). Both are significantly worse than climatology in absolute error in the middle blocks ($p \leq 0.05$ for 1926-1950 and $p < 0.001$ for 1951-1975 and 1976-2000), and indistinguishable from it in 2001-2024 (p = 0.70 and 0.65), the block that has the least training data relative to variance. Brier skill scores for the boosted classifier are negative in every block (-0.251, -0.098, -0.101, -0.103). Drought and flood metrics by block rest on very few events (four to 15 drought months and five to 17 flood months per block) and should be read as illustrations rather than evidence of change.

Table 9. Monsoon Onset Detection

Model	Exact onset month	Within one month	Late-onset accuracy	Late-onset F1	Late-onset recall
Persistence	0.030	0.980	0.576	0.720	0.947
Seasonal naive	0.444	0.960	0.465	0.531	0.526
Climatology	0.515	0.990	0.535	0.685	0.877
Ridge	0.434	1.000	0.465	0.576	0.632
Random forest	0.465	0.980	0.475	0.500	0.456
HGB single-stage	0.485	0.970	0.485	0.541	0.526
HGB Poisson	0.525	0.970	0.525	0.584	0.579
Two-stage linear	0.535	0.990	0.556	0.699	0.895
Two-stage HGB	0.475	0.980	0.485	0.532	0.509

Note. Onset = first Jun–Sep month with rainfall above 50 mm; predicted onset = first month with forecast above 50 mm. Late onset = June rainfall at or below 50 mm (57.6% of 99 test years). Always predicting July gives an exact-month accuracy of 54.5%; always predicting a late onset gives an accuracy of 57.6%.

Table 10. Forecast Anomaly Diagnostics

Model	Anomaly correlation, all months	Anomaly correlation, Jun–Sep	Regression slope of observed on forecast anomaly, Jun–Sep	Mean forecast in flood months (mm; observed mean 261.3)
Persistence	-0.025	-0.032	-0.024	95.8
Seasonal naive	0.040	0.031	0.031	156.6
Ridge	-0.053	-0.073	-0.729	144.7
Tweedie GLM	-0.051	-0.055	-0.230	141.3
Random forest	0.006	-0.016	-0.069	148.4
HGB single-stage	0.012	-0.003	-0.006	141.7
HGB Poisson	0.047	0.042	0.103	140.5
Two-stage linear	-0.066	-0.079	-0.311	134.5
Two-stage HGB	0.004	-0.006	-0.014	140.1

Note. Anomaly = deviation from the expanding-window climatological mean of the target month. Slope = coefficient from regressing observed anomalies on forecast anomalies in Jun–Sep (1 would indicate calibrated anomaly forecasts).

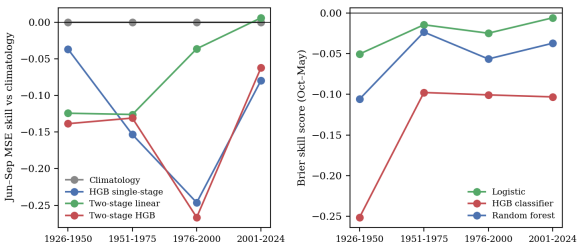


Figure 5. Stability across 25-year blocks. Left: Jun–Sep MSE skill against climatology. Right: Brier skill score for occurrence in October–May. All values are at or below zero, with no systematic trend across blocks.

6. Discussion

The central finding is that on this series nothing beats climatology at a one-month horizon, neither in the accuracy of amounts nor in the probability of occurrence, and that the two-stage structure does not change this. This result contradicts the working expectation that a two-stage model would capture occurrence information that a single-stage model discards, and it also contradicts the assumption that single-stage models fail because they collapse toward zero. In the present data the single-stage models do not collapse: they track the seasonal cycle well, because the calendar month explains 63% of the variance and the climatological mean of the target month is included among the predictors. What they cannot do is predict departures from the seasonal cycle, since the departures are almost serially uncorrelated. The failure is therefore not of the loss function but of the information set.

This changes how the evaluation protocol should be read. Decision-focused metrics were expected to rank models differently from aggregate metrics, and the question was whether the model with the best RMSE is the model with the best drought recall. In the data, the ranking is nearly identical, because all models have little skill, and differences in recall or precision are within sampling noise. This is not evidence against decision-focused evaluation. It is a reminder that such metrics rest on few events (37 drought and 52 flood months in 99 years), have wide bootstrap intervals, and cannot rescue a forecaster without signal. The protocol did show something useful: it identified that drought months in this region are late onsets (June) and early withdrawals (September) rather than mid-season fail-

Table 11. June-September Rainfall by 25-Year Block

Block	Years	Jun–Sep mean (mm)	SD (mm)	CV
1901-1925	25	440.1	138.3	0.31
1926-1950	25	422.8	134.0	0.32
1951-1975	25	479.5	112.7	0.24
1976-2000	25	463.9	158.0	0.34
2001-2024	25	393.6	107.0	0.27

Note. Kruskal-Wallis test across the five blocks, p = .12; Levene test for equal variances, p = 0.26. CV = coefficient of variation.

Table 12. Mann-Kendall Trend Tests, 1901-2024

Series	Kendall τ	p-value	Sen slope (mm per year)
Annual total	-0.013	.829	-0.087
JJAS total	-0.020	.736	-0.117
June	0.086	.157	0.107
July	-0.011	.859	-0.030
August	-0.015	.809	-0.043
September	0.004	.948	0.006
Occurrence (>5 mm) share, Oct–May	0.020	.761	0.000

Note. Annual and Jun–Sep totals are yearly series (124 values); monthly series use the 124 values for that month; the last row is the yearly share of Oct–May months above 5 mm.

ures, that conditional-mean forecasts almost never cross absolute event thresholds, so that rank-based or probabilistic warnings are needed, and that onset timing has a base rate that no model beats.

The benchmark comparison offers further lessons. The best learners are the most heavily regularised ones (ridge regression and the two-stage linear model), and the proposed boosted models, which have more flexibility, perform worse, in line with the statistical-learning view that flexible models need signal to justify their variance (and consistent with the sample-size analysis in the companion study). The Poisson-loss and Tweedie models, designed for zero-heavy data, perform similarly to squared-error models, indicating that the loss function matters little when there is no signal. The two-stage boosted classifier is the worst calibrated of the occurrence forecasters, suggesting that probability calibration (for example by Platt scaling or isotonic regression) would be a necessary addition to a practical system, although calibration cannot create skill that is not there.

Several limitations apply. The series is a regional aggregate for Haryana, Chandigarh and Delhi, not station-level data, and the data source and construction are not documented in the file. Monthly resolution hides the daily structure, such as active and break spells, that matters for crops. The horizon is one month, and seasonal or two-month horizons may behave differently, particularly with predictors such as sea-surface temperature, which are known to carry monsoon information (Gadgil et al., 2005; Rajeevan et al., 2007) but were not available here. Hyperparameters were fixed rather than tuned, and a tuned model, particularly one with stronger regularisation, might approach climatology more closely, though the evidence suggests it would be unlikely to exceed it. The test excludes 1901-1925, and the number of drought and flood events per block is small. Data quality in the early record is uncertain, and the wet-dry threshold (5 mm) and event thresholds (20 mm, 200 mm, 50 mm) follow the study design rather than being estimated. Finally, because the study used only this region, the results cannot be generalised to other states, whose rainfall regimes may be more persistent or more predictable; a multi-state panel would be a natural extension, and it would also provide more events for drought and flood evaluation.

The practical implication for agricultural planning is a cautionary one. A farmer or planner should not expect a model that uses only past monthly rainfall for the region to say anything useful about the next month beyond the climatological expectation, and any forecast that claims skill from such inputs alone should be evaluated against climatology with the metrics used here. Improvements will have to come from additional information, such as ocean–atmosphere indices, soil moisture, daily data or upstream stations, evaluated with the same protocol.

7. Conclusion

This paper specified a decision-focused evaluation protocol for zero-inflated monthly rainfall and applied it to a 124-year series for the Haryana-Chandigarh-Delhi region, comparing a two-stage boosted forecaster with nine benchmarks in a one-month-ahead expanding-

Table 13. Level and Event Metrics by 25-Year Block

Block	Model	Jun–Sep RMSE	Jun–Sep skill	MAE all	Drought recall (events)	Flood precision (alarms)
1926-1950	Climatology	67.47	0.000	23.24	0.33 (15)	0.00 (0)
	HGB single-stage	68.69	-0.037	24.43	0.47 (15)	0.13 (8)
	Two-stage linear	71.54	-0.124	23.25	0.53 (15)	0.50 (2)
	Two-stage HGB	71.99	-0.139	25.24	0.53 (15)	0.09 (11)
1951-1975	Climatology	69.23	0.000	23.91	0.45 (11)	0.00 (2)
	HGB single-stage	74.35	-0.153	26.61	0.36 (11)	0.21 (19)
	Two-stage linear	73.46	-0.126	24.77	0.45 (11)	0.00 (7)
	Two-stage HGB	73.62	-0.131	26.05	0.27 (11)	0.25 (16)
1976-2000	Climatology	65.25	0.000	23.80	0.25 (4)	0.28 (40)
	HGB single-stage	72.85	-0.246	26.16	0.25 (4)	0.29 (24)
	Two-stage linear	66.42	-0.036	23.88	0.25 (4)	0.29 (35)
	Two-stage HGB	73.43	-0.266	26.19	0.25 (4)	0.29 (24)
2001-2024	Climatology	60.68	0.000	23.63	0.14 (7)	0.11 (18)
	HGB single-stage	63.04	-0.079	24.00	0.43 (7)	0.11 (9)
	Two-stage linear	60.50	0.006	23.62	0.00 (7)	0.06 (16)
	Two-stage HGB	62.53	-0.062	23.93	0.43 (7)	0.11 (9)

Note. Skill = 1 - MSE/MSE of climatology over Jun–Sep. Drought recall: share of drought months warned about at a 20% alarm rate (number of drought months in brackets). Flood precision: share of flood warnings followed by a flood month (number of warnings in brackets); warning thresholds are fixed over the full test period.

Table 14. Occurrence Skill and Pairwise Tests by Block

Block	BSS: logistic	BSS: random forest	BSS: HGB classifier	DM: two-stage vs single (Jun–Sep)	DM: two-stage vs climatology (all months)	DM: single vs climatology (all months)
1926-1950	-0.050	-0.106	-0.251	1.08 (.281)	1.94 (.053)	1.36 (.176)
1951-1975	-0.015	-0.024	-0.098	−0.85 (.395)	3.09 (.002)	3.75 (< .001)
1976-2000	-0.025	-0.057	-0.101	0.80 (.427)	3.59 (< .001)	3.21 (.001)
2001-2024	-0.006	-0.037	-0.103	−0.58 (.566)	0.39 (.696)	0.46 (.649)

Note. BSS for Oct–May relative to climatological occurrence rate. DM statistics for absolute error (p-values in parentheses); positive means the first model is worse.

window evaluation over 1926-2024. No model beats monthly climatology on level accuracy, occurrence skill, drought recall, flood precision or onset detection, and the two-stage structure is statistically indistinguishable from single-stage boosting. Calendar month explains 63% of the variance, the remaining anomalies are almost uncorrelated through time, and the result is stable across 25-year blocks. The protocol and benchmark suite give a reproducible standard: a rainfall forecasting method for this region should be required to beat climatology on the Brier skill score and on event metrics before it is claimed to be useful.

Future work should extend the analysis to the other states for which rainfall series are available, which would broaden the range of rainfall regimes and increase the number of drought and flood events, add physically based predictors, evaluate longer horizons and daily data, tune and calibrate the learners, and test whether the two-stage structure helps when real signal is available. If it does, the protocol here will be the tool for demonstrating it; if it does not, the negative result will itself be a useful guide to where forecasting effort should be directed.

References

Bollasina, M. A., Ming, Y., and Ramaswamy, V. (2011). Anthropogenic aerosols and the weakening of the South Asian summer monsoon. Science, 334(6055), 502-505.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3.

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

Dunn, P. K., and Smyth, G. K. (2005). Series evaluation of Tweedie exponential dispersion model densities. Statistics and Computing, 15(4), 267-280.

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

Gadgil, S., and Gadgil, S. (2006). The Indian monsoon, GDP and agriculture. Economic and Political Weekly, 41(47), 4887-4895.

Gadgil, S., Rajeevan, M., and Nanjundiah, R. (2005). Monsoon prediction: Why yet another failure? Current Science, 88(9), 1389-1400.

Gneiting, T., and Katzfuss, M. (2014). Probabilistic forecasting. Annual Review of Statistics and Its Application, 1, 125-151.

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378.

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

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

Jolliffe, I. T., and Stephenson, D. B. (Eds.). (2012). Forecast verification: A practitioner's guide in atmospheric science.

Jørgensen, B. (1987). Exponential dispersion models. Journal of the Royal Statistical Society: Series B, 49(2), 127-162.

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

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. The Annals of Statistics, 17(3), 1217-1241.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20(2), 130-141.

Mann, H. B. (1945). Nonparametric tests against trend. Econometrica, 13(3), 245-259.

Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595-600.

Murphy, A. H. (1992). Climatology, persistence, and their linear combination as standards of reference in skill scores. Weather and Forecasting, 7(4), 692-698.

Murphy, A. H., and Winkler, R. L. (1987). A general framework for forecast verification. Monthly Weather Review, 115(7), 1330-1338.

Newey, W. K., and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3), 703-708.

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., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Politis, D. N., and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In R. LePage and L. Billard (Eds.), Exploring the limits of bootstrap, 263-270.

Rajeevan, M., Pai, D. S., Kumar, R. A., and Lal, B. (2007). New statistical models for long-range forecasting of southwest monsoon rainfall over India. Climate Dynamics, 28(7-8), 813-828.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63(324), 1379-1389.

Stern, R. D., and Coe, R. (1984). A model fitting analysis of daily rainfall data. Journal of the Royal Statistical Society: Series A, 147(1), 1-34.

Turner, A. G., and Annamalai, H. (2012). Climate change and the South Asian summer monsoon. Nature Climate Change, 2(8), 587-595.

Wilks, D. S. (2011). Statistical methods in the atmospheric sciences (3rd ed.). Academic Press.

Woolhiser, D. A., and Roldán, J. (1982). Stochastic daily precipitation models: 2. A comparison of distributions of amounts. Water Resources Research, 18(5), 1461-1468.