Research Square

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

Rainfall timing and forecast verification for agricultural climate services in West Africa: a critical review

Segun Ezekiel Akorede

segun.akorede@st.futminna.edu.ng

Institute of Atmospheric Physics

Opeyemi Samuel Ibitoye

China Agricultural University

Chioma Suzan Dibani

College of Resources and Environment, University of Chinese Academy of Sciences

Chinyere Salome Ofordu

Forestry Research Institute of Nigeria

Systematic Review

Keywords: rainfall timing, rainfall onset, dry spells, forecast verification, climatological benchmarks, West Africa

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11264981/v1

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

Additional Declarations: No competing interests reported.

Rainfall timing and forecast verification for agricultural climate services in West Africa: a critical review

Segun Ezekiel AKOREDE[1,2]*; Opeyemi Samuel IBITOYE[3,4]; Chioma Suzan DIBANI[2,5]; Chinyere Salome OFORDU[6]

1. International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, 81 Beichen West Road, Chaoyang District, Beijing 100029, China.
2. University of Chinese Academy of Sciences, 19A Yuquan Road, Shijingshan District, Beijing 100049, China.
3. China Agricultural University, 2 Yuanmingyuan West Road, Haidian District, Beijing 100193, China.
4. Adekunle Ajasin University, P.M.B. 01, Akungba-Akoko, Ondo State, Nigeria.
5. College of Resources and Environment, University of Chinese Academy of Sciences, No. 1, Yanqihu East Road, Huairou District, Beijing, Xueyuan 4, Yanqihu Campus, 101408, China.
6. Environmental Modeling and Biometrics Department, Forestry Research Institute of Nigeria, P.M.B. 5054, Jericho Hills, Ibadan, Oyo State, Nigeria.

* Corresponding author

Abstract

Seasonal rainfall accumulation, planting onset, and dry-spell sequences are different verification targets for agricultural climate services. These targets have different seasonal frequencies, forecast horizons, and observational requirements. This focused critical review examines how those differences constrain agricultural interpretations of forecast skill in West Africa. It brings together rainfall-climatology studies, forecast evaluations, and reference-data assessments, and relates them to established verification and climate-service frameworks. The synthesis identifies four requirements: define the rainfall target and eligible calendar window; match forecast horizon and spatial support to that target; compare with independently fitted, seasonally appropriate benchmarks; and test sensitivity to rainfall references and justified event definitions. The Agricultural Rainfall Forecast Evaluation Framework (ARFEF) records these requirements alongside the action deadline and the evidence supporting an agricultural claim. A reproducible constructed example gives lower daily rainfall root mean square error for one forecast, 0.246 versus 1.279 mm, but a higher error fraction for a specified planting event, 0.50 versus 0.00. Threshold and reference changes also alter the ranking. The example isolates rainfall-threshold classification; identical subsequent dry-spell sequences do not test dry-spell prediction skill. The review provides a rainfall-specific reporting synthesis for applied climatology, while distinguishing meteorological skill from conditional loss estimates and observed agricultural outcomes. Operational archive comparisons are still required to test its practical benefit.

Keywords: rainfall timing; rainfall onset; dry spells; forecast verification; climatological benchmarks; West Africa

1 Introduction

Rainfall timing is central to the agricultural interpretation of climate forecasts. A seasonal accumulation summarises water supply over several months; onset and dry-spell indices describe when that supply becomes available and how it is interrupted. Forecast skill for one of these quantities is evidence about that quantity, calendar window and spatial support. It cannot be transferred directly to another rainfall target merely because both are relevant to the same crop. This review examines that transfer of evidence in agricultural climate services.

West Africa makes the problem concrete because rainfall seasonality and agricultural calendars vary within the region. The Sahelian monsoon season and the two wet seasons found in parts of southern coastal West Africa require different eligible periods and climatological comparisons (Nicholson 2013; Dunning et al. 2016). Regional monsoon onset, a local rainfall-season boundary and a rainfall threshold used for sowing are also different quantities. Crop responses depend on rainfall timing relative to development, rather than seasonal accumulation alone (Sultan et al. 2005). West Africa is therefore the principal application context for examining target and calendar dependence, rather than a homogeneous forecast-verification domain.

Published evaluations demonstrate why the distinction matters. Seasonal rainfall studies evaluate regional regimes or particular downscaling configurations (Rodrigues et al. 2014; Siegmund et al. 2015). Onset studies evaluate regional monsoon timing or locally defined rainfall-season indicators (Vellinga et al. 2013; Gbangou et al. 2019; Kumi et al. 2020). In Ghana, Gbangou et al. (2020) evaluate seasonal rainfall, maximum dry-spell length and dry-spell frequency separately and find performance that varies with the index, station and lead. These findings support target-specific interpretation; they do not establish that a skilful seasonal total predicts the dry spell following a particular sowing date.

Verification also depends on the reference used to label rainfall events. A gauge and a gridded rainfall product differ in support, coverage and measurement or estimation error. Agreement with one reference is therefore not agreement with error-free local rainfall. Forecast verification across African satellite rainfall products illustrates this dependence (Wang et al. 2023). For agricultural climate assessments, Romanovska et al. (2023) also show why general agreement

with a climate reference can coexist with discrepancies in agroclimatological indices. The reference and aggregation must be part of the interpretation of skill.

The distinction between forecast quality and value is established (Murphy 1993; Jolliffe and Stephenson 2012; Wilks 2019). Agricultural service reviews and recent evaluation frameworks already connect verification to users, actions and outcomes (Tall et al. 2018; Vaughan et al. 2019; Crochemore et al. 2024; Emerton et al. 2026). The unresolved task addressed here is more specific: to document how a rainfall performance statement depends jointly on the target, seasonal window, information available at issuance, spatial support, benchmark and rainfall reference. Agricultural consequences bound the claim that follows from this specification; they do not replace rainfall verification.

The primary audience is applied climatologists evaluating rainfall information for agricultural climate services. The review asks: under what climatological and forecast-verification conditions can a rainfall skill estimate support a specified agricultural application? Section 3 compares the rainfall targets and their seasonal and observational requirements. Section 4 synthesises these requirements into the Agricultural Rainfall Forecast Evaluation Framework (ARFEF). Section 5 defines planting-condition onset and gives a constructed arithmetic example. Sections 6-8 delimit agricultural claims and identify the independent archive tests still needed. The contribution is a focused critical synthesis and reporting sequence, not a new forecast model, score, or empirically validated evaluation method.

2 Review design and relationship to existing evaluation frameworks

2.1 Focused critical synthesis

The synthesis addresses the conditions needed to interpret rainfall forecast performance for agricultural applications. The initial synthesis covered literature identified through 24 August 2026, using targeted searches, citation tracing and institutional material. The original evidence streams cover verification theory, West African rainfall prediction, agricultural information use and forecast valuation. For the present synthesis, rainfall-climatology and forecast studies establish the target and evaluation conditions; service and valuation studies delimit the downstream claims. Appendix S1 (Online Resource 1) describes the reported approach and its documentation limits.

A targeted update on 5 October 2026 examined recent user-centred and operational-verification literature, including Emerton et al. (2026) and Pagano et al. (2024, 2025), and added directly relevant studies of rainfall indices and climate-reference uncertainty (Gbangou et al. 2020; Romanovska et al. 2023; Mendes et al. 2025). This update did not repeat the original retrieval and screening process. The review was not preregistered, did not use duplicate screening, and does not provide a complete exclusion log. The review is therefore a focused critical synthesis, with no claim to exhaustive retrieval, method prevalence or pooled regional effects.

Sources were compared by rainfall target, calendar definition, forecast horizon and spatial support, verifying reference, benchmark and supported inference. Where agricultural outcomes were discussed, their evidence type was recorded separately. A rainfall climatology, a hindcast comparison, a qualitative service-use study and a crop-model simulation answer different questions. Appendix S2 (Online Resource 1) states the role and limitation of the principal source groups; it does not imply that every study reports all evaluation fields.

2.2 What the rainfall-specific synthesis adds

Established verification texts explain format-matched scores, calibration, discrimination and comparisons with climatology (Jolliffe and Stephenson 2012; Wilks 2019). Murphy (1993) distinguishes quality from value. Agricultural service reviews examine information use and outcome evaluation (Tall et al. 2018; Vaughan et al. 2019). The present synthesis uses these established principles to specify a rainfall application: its season or sowing window, accumulation and dry-day definitions, future conditions needed to label onset, reference support, and independent benchmark fitting. ARFEF is the reporting sequence that assembles these conditions; its individual ingredients are not new to verification theory.

Crochemore et al. (2024) connect technical verification with evaluation of seasonal services across sectors. Emerton et al. (2026) provide a participatory framework covering decision context, user thresholds, historical event verification, and communication. ARFEF shares these premises and does not claim to originate action deadlines or user-centred evaluation. Its rainfall-specific emphasis is the combination of calendar-conditioned benchmarks, explicit future-window onset rules, member-wise event calculation, multi-reference sensitivity, and a stated boundary between rainfall skill and agricultural outcome evidence. Those requirements extend the application specification, rather than establishing superiority over the earlier frameworks.

Unlike the applied cases in Emerton et al. (2026), the present numerical illustration uses constructed rainfall.

Pagano et al. (2024, 2025) address operational verification planning, archives and institutional capability. ARFEF can specify the rainfall target within that wider programme, but cannot supply missing observations, resolve dissemination constraints or establish causal agricultural value. Its proposed practical contribution is an auditable statement of which rainfall result supports which agricultural inference. Whether using that statement changes product selection or advice is an empirical question, addressed in Section 8.2.

2.3 Boundary between rainfall verification and agricultural value

Let d identify a decision and x the user context, including resources and decision authority. Let $\mathcal{A}_d(x)$ be the feasible action set, S the environmental state, I the forecast information and U_d an explicitly specified utility or outcome function. A schematic forecast-informed choice is

$$a_d^*(I,x) = \underset{a\in\mathcal{A}_d(x)}{\arg\max}\mathbb{E}[U_d(a,S;x) \mid I,x]. \qquad (1)$$

This notation identifies the additional action and consequence assumptions needed after a rainfall target has been verified. It does not assume that farmers optimise a known utility function. Two users may select different actions from the same meteorological information because their resources or preferences differ (Katz and Murphy 1997; Meza et al. 2008). Equations (1)-(2) provide an interpretation boundary for value claims; they are not proposed as rainfall-verification scores.

For a common evaluation distribution of environmental states and information, expected incremental value relative to benchmark information B can be represented as

$$V_d(I,B;x) = \mathbb{E}[U_d(a_d^*(I,x),S;x) - U_d(a_d^*(B,x),S;x)] - [c_d(I,x) - c_d(B,x)]. \qquad (2)$$

Information costs c_d must be expressed in units consistent with U_d. The expectation must compare the two strategies under the same declared evaluation conditions. Equation (2) is an ex-ante representation, not an estimator of causal welfare effects. Sequential applications require a policy specifying what can be revised after each update and which earlier actions remain irreversible; their value cannot be obtained by adding unrelated single-forecast estimates.

2.4 Generative AI assistance and computational checks

ChatGPT (OpenAI) assisted language revision, manuscript organisation and targeted literature checking. Generative AI output is not treated as a source of literature evidence, and responsibility for the scientific interpretation, source verification, and submitted content rests with the human authors.

3 Rainfall climatology, forecast targets and verification conditions

3.1 Rainfall seasonality defines the evaluation window

The regional seasonal cycle determines when a rainfall event is relevant and how often it occurs. Dunning et al. (2016) distinguish regions with one or two rainy seasons using the annual rainfall cycle, including the southern coastal West African little dry season. Their cumulative-anomaly method characterises the season retrospectively. It does not define the same onset as a local accumulation threshold plus a future dry-spell condition. The eligible evaluation window must follow the chosen target and local calendar rather than an assumed common West African season.

Mendes et al. (2025) analyse rainy-season timing using the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) record and Guinea-Bissau observations for 1981-2020. Their agronomic onset criterion gives a south-to-north progression, with cessation progressing in the opposite direction. This is evidence of spatially varying season timing under a particular definition, not a validation of forecast issue dates or one sowing calendar for all West Africa. Taken with Dunning et al. (2016), it shows why regional seasonal boundaries and local planting-condition dates must be distinguished.

Gbangou et al. (2020) evaluate Ghanaian seasonal totals, maximum dry-spell length and dry-spell frequency against station records, using locally relevant seasonal windows. Skill differs by index, station, and forecast lead; dry-spell-frequency correlation is significant at fewer stations than the other indices. The review inference is that forecast performance must be reported for the actual rainfall index and calendar. A zone-average score cannot establish equivalent performance for every station or for the timing of a post-sowing dry spell.

3.2 Onset and dry-spell targets are not interchangeable

Regional monsoon onset, local rainfall-season onset, and planting-condition onset describe different temporal transitions. Vellinga et al. (2013) evaluate regional monsoon onset; Gbangou et al. (2019) tailor onset estimates to Ghanaian agricultural calendars; and Kumi et al. (2020) evaluate rainfall-onset prediction over West Africa. Their reported skill remains conditional on the onset definition, sample, and forecast configuration. Table 1 summarises why these targets and seasonal rainfall indices support different interpretations.

A planting-condition definition can combine initial rainfall with protection against a subsequent dry spell. It still labels a rainfall condition, not observed germination or establishment. A seasonal maximum dry-spell length is a different target: it does not identify the spell's occurrence relative to planting or flowering. Dry-spell frequency also depends on the dry-day and duration thresholds. The evaluation must preserve these distinctions when transforming daily trajectories into indices. For fertiliser use, rainfall occurrence may be insufficient without intensity and field-access information; for soil-water management, rainfall is an input to a water balance. Agronomic evidence and local practice justify the application target, while verification tests whether that target is predicted.

Table 1 Rainfall targets in selected West African evaluations and limits of agricultural interpretation

Target and source	Verification object	Limit of interpretation
Seasonal accumulation Rodrigues et al. (2014); Siegmund et al. (2015)	Seasonal amount at the evaluated region, aggregation and configuration	Amount skill does not establish local onset or the sequence of wet and dry days
Regional monsoon onset Vellinga et al. (2013)	Regional transition under the study's onset definition	A regional onset date is distinct from a local planting event or establishment condition
Local rainfall onset Gbangou et al. (2019); Kumi et al. (2020)	Onset date or category for a defined location, season and forecast	A planting claim additionally needs local justification, complete future-window coverage and a usable issue date
Dry-spell length or	Duration or number of spells	Seasonal index skill does not

frequency Gbangou et al. (2020)	in a specified season and dry-day definition	establish spell timing relative to sowing or a sensitive crop stage

3.3 Forecast horizon, availability and spatial support

A forecast must cover the complete temporal definition of the rainfall target. A season-ahead probability of above-normal total rainfall does not by itself provide the daily sequence needed for a sowing-window dry-spell rule. A subseasonal or short-range product may supply daily trajectories, but its lead and valid period still need to cover the relevant future window. Statistical translation of seasonal categories into a local onset or dry-spell probability creates a separate forecast that requires independent evaluation. Agricultural actionability then imposes an additional cut-off: delivery and preparation must finish before the action is fixed (Ingram et al. 2002; Hansen et al. 2011).

Forecast grid spacing, reference support, and advisory area must be distinguished. A fine grid does not establish field-level accuracy, and extracting a grid value at a station does not remove point-to-area representativeness differences. Published West African downscaling and rainfall-regime comparisons show why configuration and aggregation belong to the evaluated product (Rodrigues et al. 2014; Siegmund et al. 2015). Comparisons should use matched accumulation periods and explicit spatial operators. A regional average may diagnose the monsoon forecast while masking the local event error relevant to a farm; neither result should be substituted for the other.

3.4 Climatological and operational forecast benchmarks

A skill statement requires a benchmark for the same rainfall quantity or event, location, calendar window, and lead-time task. For a sowing-window event, an annual wet-day probability is not an equivalent climatological comparator. An onset-date climatology needs the same eligible search period and no-onset convention as the evaluated forecast. Estimate benchmark quantities from a declared training record, separated from the test cases. Report absolute and benchmark scores together with event counts: a difference in skill between locations can reflect differences in event frequency as well as forecast performance (Jolliffe and Stephenson 2012; Wilks 2019).

Persistence, a simple statistical forecast, and the current operational product can answer additional comparison questions. Vogel et al. (2021) demonstrate the relevance of statistical rainfall-occurrence benchmarks for northern tropical Africa; dynamical complexity alone does

not establish added performance. A decision baseline is a separate comparator describing the action taken with usual information. Local observations, experience, and extension advice can contribute to that baseline only when their availability and response rule can be reconstructed (Roncoli et al. 2002, 2009). A technical forecast benchmark measures predictive improvement; a decision baseline measures change under a specified action strategy.

3.5 Rainfall-reference uncertainty and event labels

Rainfall reference selection defines both continuous verification values and the event labels derived from them. CHIRPS blends satellite and station information, while Tropical Applications of Meteorology using Satellite data and ground-based observations (TAMSAT) supplies satellite-based rainfall estimates for Africa (Funk et al. 2015; Maidment et al. 2017). Their provenance does not establish universal accuracy for planting thresholds. Gauge quality control, missingness, and representativeness remain important. Compare references on common available cases and recompute the onset or dry-spell labels for each. Shared observations or algorithms limit the independence of apparently separate reference checks (Wang et al. 2023).

Romanovska et al. (2023) compare climate inputs for West African agricultural impact assessment and identify reference-dependent differences and residual regional deviations in agroclimatological indices after bias adjustment. Their study concerns climate data and simulations, not operational forecast skill. Its relevance here is the need to assess reference and processing choices for the index used in an application. Reproducing a rainfall mean does not establish that onset or dry-spell labels are equally reliable.

3.6 Calendar conditioning and the period of inference

Verification samples should represent the rainfall conditions to which the application refers. Pooling all calendar days can make an error score primarily a measure of off-season dry-day agreement. Pooling a single northern rainy season with two southern wet seasons mixes different event opportunities and climatologies. Stratify or otherwise account for eligible calendar windows, rainfall regimes, and issue dates, and report the number of events in each comparison. Small strata may have wide uncertainty; this is a limit to report, rather than a reason to imply that a pooled result establishes uniform local performance.

The climatological fitting period also sets the comparison. If event frequency changes between training and application periods, a fixed historical benchmark and a benchmark updated from

data available before each issuance answer different questions. State whether the goal is retrospective hindcast diagnosis or a comparison with information available operationally. Moving-window or trend-conditioned benchmarks need an explicit fitting rule and enough independent information; they should not be fitted to the test outcomes. This review does not estimate rainfall trends. It identifies the period and information restrictions needed to interpret the forecast comparison.

3.7 Transferable inference and local specifications

What transfers across locations is the verification logic, not a universal rainfall threshold or sowing calendar. A published skill estimate applies first to its own target, reference, cases, issue dates and processing. Transferring it to another location or management window requires checking those conditions. A changed rainfall reference, finer advisory scale or locally retuned threshold defines a changed evaluation. West African evidence is used here to expose these dependencies; it does not establish one performance estimate for every country, rainfall regime or farming system.

A reproducible performance statement can therefore take a specific form: the product improves a declared score over a named benchmark for a defined rainfall event, at specified leads, in a stated calendar window, against a stated reference and common sample. An agricultural statement adds the action deadline and the evidence connecting the event to the operation. An outcome statement adds observed or modelled responses and a defensible comparison. These are progressively different claims, rather than interchangeable descriptions of a single skill score.

Sections 4-6 organise these conditions without treating a checklist as an overall quality score. Failure to document a condition restricts the inference that can be made. It does not demonstrate that the forecast is universally poor, nor does documenting every field validate the agricultural service. This distinction keeps the review focused on defensible applied-climatological interpretation.

4 Reporting synthesis for agricultural rainfall verification

ARFEF assembles the climatological and verification conditions reviewed in Section 3 into six reporting stages. The sequence starts with an application because that determines the relevant rainfall target; its central technical tasks are defining the event, matching horizon and support, verifying forecast attributes, and specifying benchmarks and references. The final stage limits the

agricultural claim to the available evidence. Figure 1 summarises the sequence, and Appendix S3 (Online Resource 1) supplies the reporting record. Iteration is allowed when a target or application changes; each revised configuration needs its own documented evaluation.

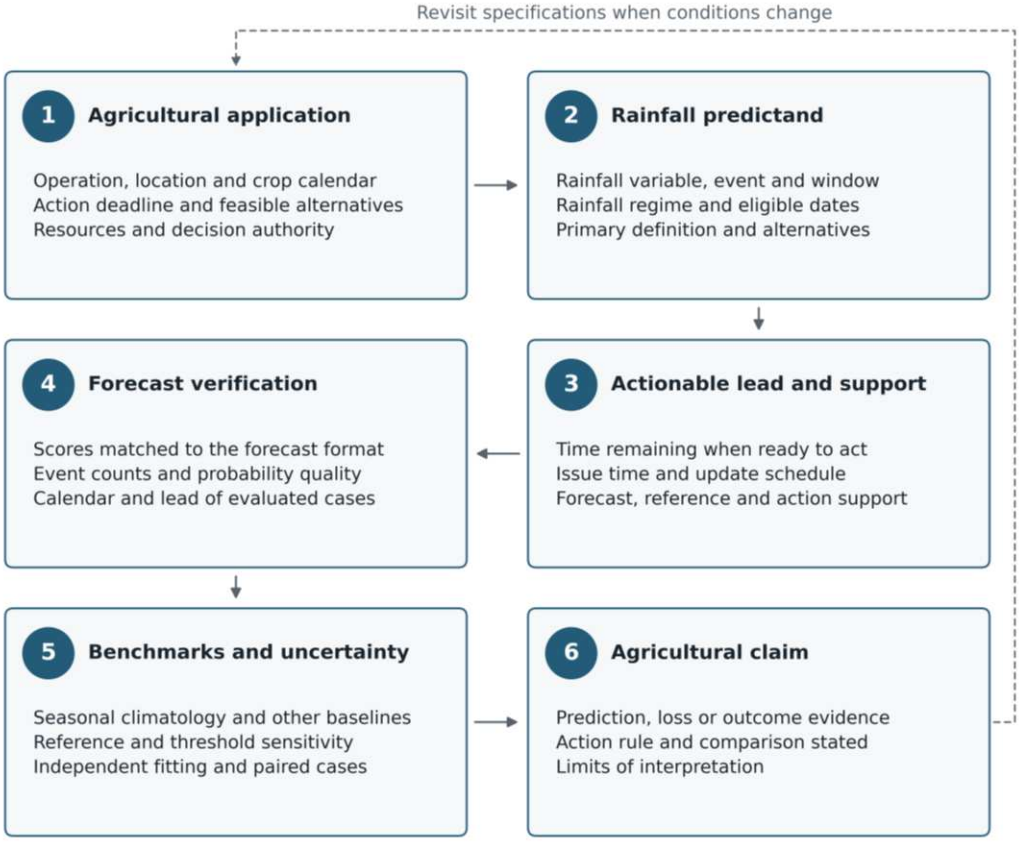


Figure 1. Reporting sequence for agricultural rainfall forecast evaluation. The six stages connect the application and rainfall target to available lead and spatial support, format-matched verification, seasonal benchmarks and uncertainty, and the supported agricultural claim. Rainfall regime, calendar window, and event frequency enter the target and benchmark specifications. The dashed path indicates revision when conditions change. ARFEF is a reporting synthesis; completing the sequence does not validate a forecast or demonstrate agricultural benefit

4.1 Stage 1: Specify the agricultural application

State the crop or production system, location, management operation, and action deadline. Identify feasible alternatives and the person with authority and resources to act. Define the outcome being considered, such as establishment risk, delayed field access, or irrigation demand. This stage fixes the application to which rainfall verification will refer; it does not require reducing all objectives to profit.

Record whether the action is one-off, staged or updated. For staggered sowing, specify the land fraction or planting opportunity that remains adjustable at each issue date. Record constraints on

labour, inputs or water only where they affect whether the forecast can alter the specified operation. Broader service governance may be evaluated separately.

4.2 Stage 2: Specify the rainfall predictand

Define the rainfall quantity or event, units, accumulation period, dry-day threshold, duration rule and location. Record the rainfall regime, crop-calendar window and eligible dates. Identify the physical link to the operation and the agronomic evidence supporting the definition. For a rainfall-to-soil-water translation, specify the additional model or observation process so that its performance can be evaluated separately.

Prespecify the primary definition and justified alternatives. A threshold chosen to maximise skill on the test sample requires independent evaluation. User consultation establishes relevance and agronomic evidence tests the physical interpretation; neither establishes predictive performance. Record whether a target can be calculated directly from the product's temporal resolution and horizon. A seasonal category translated into a within-season event probability is a separately evaluated product.

4.3 Stage 3: Establish actionable lead and matching support

Record issue time, the time information reaches the user, any further preparation needed, valid period, and update schedule. If τ_{ready} is the elapsed time from issuance to readiness to act, the remaining action margin is

$$L_{\text{act}} = t_{\text{deadline}} - \left(t_{\text{issue}} + \tau_{\text{ready}}\right). \qquad (3)$$

The readiness interval includes delivery, interpretation, and preparation, with overlapping activities recorded appropriately. A non-positive margin means that the forecast cannot change this action under the stated timing. A positive margin establishes temporal feasibility only. Separately check that the valid period covers the event's full future window. Report performance by issue date and lead rather than pooling forecasts that differ in calendar eligibility or arrive on opposite sides of the deadline.

Record forecast grid, reference support, advisory area and temporal aggregation. Interpolation, area averaging, bias correction and downscaling are part of the evaluated configuration. Archive the forecast as issued and distinguish it from a later advisory. Area or neighbourhood diagnostics

may assess displacement where area information is relevant; a broader-support result still needs separate evidence before supporting a field-specific claim.

4.4 Stage 4: Verify the attributes required by the target

Match diagnostics to the rainfall target and forecast format. Continuous amounts require magnitude error and bias, using mean absolute error (MAE) or root mean square error (RMSE) with mean error. Onset dates require a date convention and explicit handling of absent or incomplete onsets. Binary events require hits, misses, false alarms, and correct negatives, with the trigger rule declared. Report event and trigger counts with derived rates. The false alarm ratio uses forecast events as its denominator; the false alarm rate uses observed non-events. Their interpretation differs, especially when event frequency varies across seasons or locations.

For event probabilities p_i and binary reference outcomes y_i, the Brier score is

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 . \qquad (4)$$

Reliability, resolution, or discrimination diagnostics explain aspects of the score that a single mean can conceal. A probabilistic trigger threshold is a decision rule and should be reported separately from the meteorological threshold defining the event. Searching many probability triggers and reporting only the best result requires an independent test. Ensemble distributions can be evaluated with the continuous ranked probability score (CRPS), accompanied by calibration diagnostics where the application uses distributional information (Hersbach 2000; Gneiting and Raftery 2007).

Event verification complements aggregate rainfall diagnostics. Proper probability scores remain necessary when probabilities are issued; contingency counts at one trigger cannot establish calibration. A thresholded action assessment should be reported alongside the probability-quality assessment that supports it. Communication and comprehension are separate service evaluations and should not be inferred from meteorological reliability.

4.5 Stage 5: Compare against defensible benchmarks and assess uncertainty

Use the same cases, target definition, and information cut-off for the forecast and benchmark. Estimate climatological probabilities from training data or an appropriate independent record. A Brier skill score relative to a benchmark score $\mathrm{BS}_B > 0$ is

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_B}. \qquad (5)$$

Positive skill identifies improvement over this benchmark for this event, calendar window, sample, and reference. It does not establish agricultural value. If the benchmark score is zero, the ratio is undefined. Report both absolute scores, the benchmark specification, and the event frequency. This record enables differences in relative skill to be interpreted in light of event frequency and benchmark difficulty.

Separate training and testing when fitting climatology, post-processing, or event thresholds. Independent periods or blocked cross-validation should respect dependence between overlapping windows and forecast cycles. Specify evaluation years, issue dates, calendar eligibility, available cases, and missingness for each stratum. Use paired common-case differences for product comparisons, with uncertainty procedures that preserve relevant dependence. Sparse event counts restrict local conclusions even when the pooled score is precise.

Recalculate scores and event labels for justified alternative thresholds, references, and spatial supports. Report how rankings and absolute errors change, including cases lost when reference coverage differs. Treat a ranking that depends on the chosen reference as conditional evidence; averaging incompatible rankings does not remove the dependence. Record shared inputs across references and any overlap with calibration data. If no locally representative reference exists, state which application claim remains untested.

4.6 Stage 6: Bound the agricultural claim

State whether the evidence supports rainfall prediction, conditional decision loss, simulated crop or water outcomes, stated preferences, observed management response, or attributed outcome effects. An assumed action rule and loss function allow a potential-value calculation. A crop or soil-water model introduces additional structural and parameter assumptions. Neither calculation demonstrates receipt or adoption of the forecast.

An observed benefit claim requires information-delivery and response records and a credible comparison for the outcome. Report whether the design supports association or attribution. Subgroup differences matter when the proposed action is feasible for some users but unavailable to others. These conditions bound the climate-service claim; they should not be folded into an unexplained composite meteorological score.

5 Planting-condition onset: event definition and constructed example

5.1 Temporal definition and information available at issuance

For a planting application, a candidate date can be labelled using initial rainfall and a subsequent dry-spell constraint. Let P_τ denote daily rainfall and $D_{\max}(u, v; q)$ the longest run of days with rainfall below the dry-day threshold q between u and v, inclusive. One explicit implementation is

$$Y_t = \mathbf{1}\{\sum_{j=0}^{k-1} P_{t+j} \geq r^*\}\mathbf{1}\{D_{\max}(t+k, t+k+m-1; q) \leq d^*\}. \qquad (6)$$

Here k is the initial accumulation window, r^* its rainfall threshold, m the subsequent checking window and d^* the maximum allowed dry-spell duration. The two windows do not overlap in this implementation. Its thresholds must be justified for the crop and location. Sivakumar (1992) illustrates the use of an initial rainfall requirement with a subsequent dry-spell condition; Gbangou et al. (2019) illustrate why onset definitions need to follow local calendars and forecast configurations. Equation (6) specifies a rainfall event, not a guarantee of crop establishment.

If the target is an onset date rather than a candidate-window event, an additional rule is needed:

$$O = \min\{t \in \mathcal{T} : Y_t = 1\}. \qquad (7)$$

The eligible search dates $\mathcal{T}$, date convention and treatment of no qualifying date must be specified. An incomplete future window is missing information, not automatically a non-event. Date-error statistics restricted to cases with onset in both forecast and reference should report that conditioning and separately describe false or absent onsets.

A future dry-spell condition can label onset retrospectively. It cannot be observed at the initial planting decision. An operational forecast must predict the remaining window using information

available at issuance, while reference rainfall can subsequently verify the label. For ensemble members that supply complete rainfall trajectories, the uncalibrated event frequency is

$$\hat{p}_t = \frac{1}{M}\sum_{j=1}^{M} Y_t^{(j)}. \qquad (8)$$

Compute the event for each member before averaging its occurrence. Applying a nonlinear onset rule to ensemble-mean rainfall defines a different forecast and can conceal dry-spell risk. The resulting frequency still requires evaluation and, where appropriate, calibration. A seasonal total or tercile probability does not supply the daily trajectories needed for this operation.

Where root-zone water is the decision state, a schematic extension is

$$Y_t^{\mathrm{soil}} = Y_t \mathbf{1}\{\min_{\tau \in \mathcal{W}_t} \theta_z(\tau) \geq \theta^*\}. \qquad (9)$$

The establishment window $\mathcal{W}_t$, soil layer z, moisture threshold θ^*, initial moisture and management assumptions need independent justification. If moisture is modelled, verification evaluates the combined rainfall-to-soil-water system; its errors should not be assigned wholly to the rainfall forecast.

5.2 Reproducible numerical construction

A constructed example makes the distinction between rainfall error and planting-event performance explicit. It contains eight separate candidate-window cases, each with 33 days of deliberately specified rainfall. They are not stations, seasons, or samples from a forecast archive. The event uses $k = 3$, $m = 30$, $r^* = 20$ mm, $q = 1$ mm and $d^* = 7$ days. The reference has 21 mm on day 1 in cases 1-4 and 19 mm on day 1 in cases 5-8, with zero rainfall on days 2-3. All cases have 6 mm on days 4, 10, 16, 22, and 28, and zero on the other days. The longest subsequent dry spell is five days, so four cases satisfy the event.

Forecast A changes day-1 rainfall from 21 to 19 mm in cases 1-2 and from 19 to 21 mm in cases 5-6. Everything else matches the reference. Forecast B adds 3 mm on day 1 in cases 1-4, subtracts 3 mm on day 1 in cases 5-8, and adds 3 mm on each of the five subsequent wet days in every case. Its larger rainfall errors preserve the primary event labels. Both forecasts are deterministic constructed trajectories, so no probability-reliability result is inferred.

Aggregate daily rainfall error is computed over all $N = 8$ cases and $T = 33$ days:

$$\mathrm{RMSE} = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{s=1}^{T}\left(F_{i,s} - P_{i,s}\right)^2}. \qquad (10)$$

5.3 Target, threshold and reference dependence

Forecast A has RMSE 0.246 mm, compared with 1.279 mm for B. Aggregate rainfall error therefore favours A. For the specified planting event, A has two hits, two misses, two false alarms and two correct negatives; B has four hits and four correct negatives. The event-error fractions are 0.50 and 0.00, respectively. B performs better for this constructed event despite its higher rainfall RMSE (Figure 2).

To illustrate asymmetric consequences, let action $a = 1$ mean plant and $a = 0$ mean wait, and let y be the reference event. A deliberately simplified loss is

$$\ell(a, y) = c_{\mathrm{FA}}\mathbf{1}\{a = 1, y = 0\} + c_{\mathrm{M}}\mathbf{1}\{a = 0, y = 1\}. \qquad (11)$$

With assumed false-alarm loss $c_{\mathrm{FA}} = 2$ and missed-opportunity loss $c_{\mathrm{M}} = 1$, mean loss is 0.75 for A and 0.00 for B. An always-wait decision has a mean loss of 0.50. A therefore increases this assumed loss by 0.25 relative to that decision baseline, while B reduces it by 0.50. These are arbitrary loss units, not measured crop losses or monetary returns. The example assumes action follows the event forecast and does not represent adoption, constrained responses, or the full costs and benefits of planting.

The preference also depends on the event definition (Figure 2e). At an 18 mm threshold, all primary-reference cases satisfy the event; A classifies all correctly, and B misses four. At 22 mm, no primary-reference case satisfies it; A again classifies all correctly, and B produces four false alarms. These alternatives demonstrate threshold dependence, not agronomic justification for changing the primary threshold.

Reference dependence can be demonstrated separately. A second constructed reference changes day-1 rainfall by 2 mm in cases 1-2 and 5-6 so that its event labels match A. At the unchanged 20 mm threshold, A's event-error fraction becomes 0.00, and B's becomes 0.50. This is a designed boundary perturbation, not a second observational product or an estimate of measurement uncertainty. It shows algebraically why event labels must be recomputed when a

reference changes. Appendix S4 (Online Resource 1) supplies the complete construction, results, and reproduction procedure.

The example isolates classification near the initial rainfall threshold. All reference and forecast trajectories have the same wet-day timing after day 3 and a longest subsequent dry spell of five days. It therefore does not test skill in predicting whether or when the dry-spell condition fails. It establishes a possible mismatch between aggregate amount error and one specified planting-event classification. Its reference perturbation is designed, and it supplies no observed event-frequency estimate or operational benchmark comparison. Archive testing is required to establish how often these dependencies occur and whether they change a usable forecast choice.

5.4 Boundaries of published performance and rainfall-to-impact translation

Published studies provide empirical evidence for their evaluated targets, not for ARFEF as a reporting sequence. Gbangou et al. (2019) and Kumi et al. (2020) evaluate onset configurations; Gbangou et al. (2020) compare seasonal rainfall and dry-spell indices. Bliefernicht et al. (2019) evaluate regional seasonal forecast quality and potential value. Reading them together exposes the additional evidence needed before a seasonal or onset score supports local planting advice: the specified calendar, complete future window, usable issue date, and establishment interpretation. No new scores are calculated from these archives in the present review.

Rainfall-to-impact translation introduces a further boundary. For supplemental irrigation, the relevant target may be root-zone water or accumulated deficit, with rainfall as only one input alongside initial moisture, evapotranspiration, and soil parameters. Equation (9) makes the same distinction for establishment conditions. Verification then addresses the coupled rainfall-to-water system, and its errors cannot be assigned solely to rainfall prediction. These extensions define the limit of the review's rainfall-specific conclusions; they are not evidence from an irrigation or crop trial.

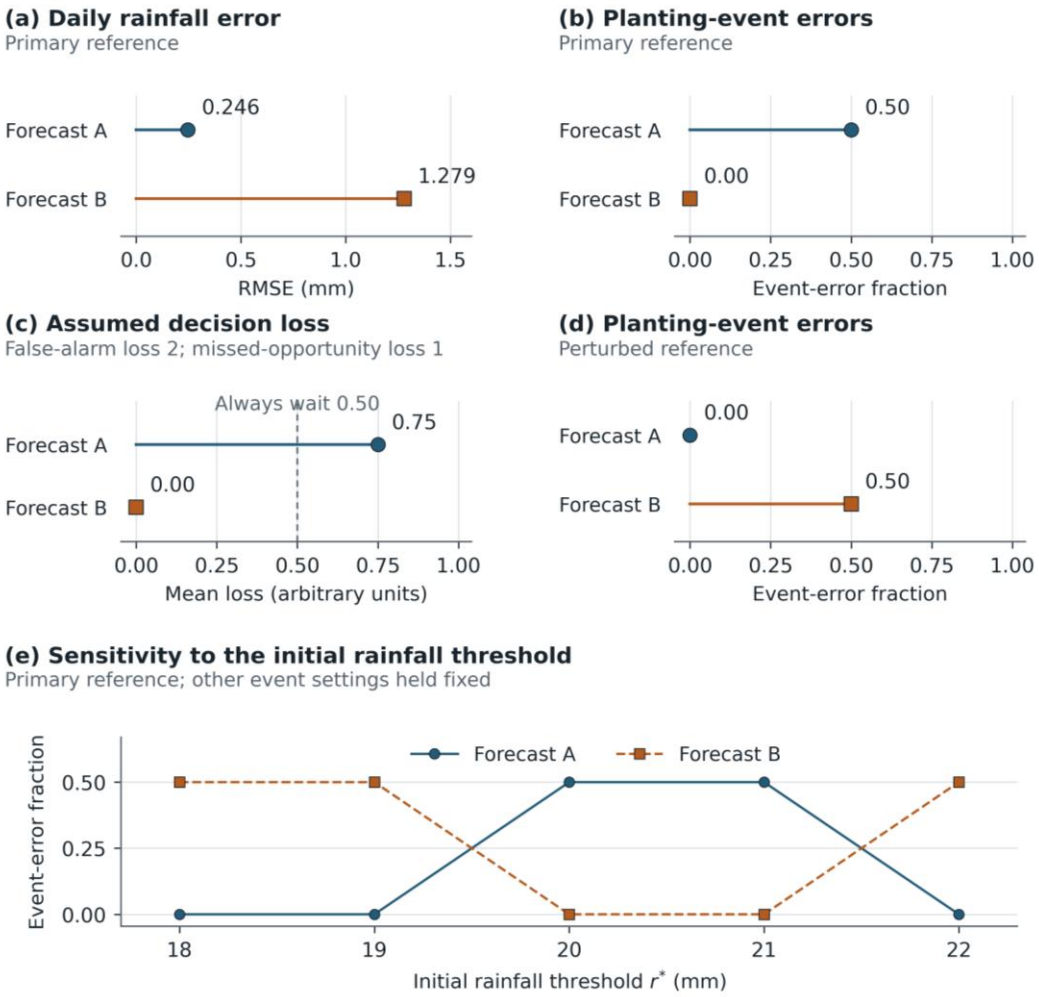


Figure 2. Forecast comparisons for eight constructed 33-day rainfall windows. Panels show (a) daily rainfall RMSE, (b) planting-event error fraction under the primary reference, (c) mean decision loss, (d) event-error fraction under the perturbed reference, and (e) threshold sensitivity. Event-error fraction equals misses plus false alarms divided by eight. The primary event requires at least 20 mm over days 1-3 and no dry spell longer than seven days during days 4-33, with rainfall below 1 mm defining a dry day. Assumed false-alarm and missed-opportunity losses are 2 and 1, in arbitrary units; always waiting gives mean loss 0.50. Panel (e) evaluates thresholds of 18, 19, 20, 21 and 22 mm. All rainfall arrays are constructed. The figure demonstrates possible ranking changes, without estimating empirical performance or observational uncertainty. Online Resources 1 and 2 supply the specifications and code

6 From verified rainfall to agricultural claims

6.1 Conditional loss and simulated outcomes

Conditional loss estimates require a declared action rule, costs, consequences, and decision baseline. Binary cost-loss methods illustrate this link, while multiple actions and resource constraints can alter it (Richardson 2000; Palmer 2002). The loss calculation in Section 5 uses

arbitrary assumptions to illustrate this dependence. It is separate from rainfall quality and is not measured in farm value.

Crop or economic simulations propagate forecast-informed actions under specified weather, prices, and resource assumptions. Examples from Senegal and Niger show this ex-ante approach (Sultan et al. 2010; Roudier et al. 2012, 2016). Comparisons require the same evaluation conditions and sensitivity to crop and behavioural assumptions. Perfect information supplies an upper-bound scenario, not an operational forecast. Simulated improvement remains conditional on the model and response rule.

6.2 Preference evidence, responses and attributed outcomes

Stated willingness to pay, as examined among maize farmers in Benin (Amegnaglo et al. 2017), measures perceived benefit under survey conditions. It does not measure realised yield or causal welfare gains. Observed choices also depend on access and available alternatives.

Participatory studies identify management responses (Roncoli 2006; Roudier et al. 2014), while outcome studies examine service-associated productivity (Sanfo et al. 2022). Selection into services, input access, spillovers, and accompanying interventions can prevent a difference in outcomes from isolating a forecast effect. State whether the design supports a behavioural observation, an association or an attributed effect.

The evidence categories answer different questions, rather than forming a simple quality ladder (Tall et al. 2018; Vaughan et al. 2019). A rainfall verification result supports a prediction claim. Linking that result to an agricultural outcome requires additional response and comparison evidence. Service delivery can fail even when rainfall skill is useful, and observed participation alone does not establish benefit.

7 Minimum reporting for regional rainfall climate services

A regional verification record should identify the product and version, forecast issue dates, valid periods, rainfall regime, eligible crop-calendar windows, reference and spatial operators. Preserve the forecast and any advisory as issued. Forecast inputs and action triggers must use only information available at historical issuance. Subsequent verifying rainfall may be quality controlled, with reference versions and revisions documented.

For a planting application, report aggregate rainfall diagnostics alongside the exact onset or candidate-window event, event counts, and probability quality where probabilities are issued. Specify the calendar-conditioned benchmark, fitting period, common sample, and actionable leads. Show how justified threshold and reference alternatives affect the conclusion. A separate record should state the rainfall-to-agronomic translation and which outcome claims remain unsupported. Appendix S3 (Online Resource 1) supplies these fields.

Regional coordination can provide compatible archives and calculation scripts while retaining local targets and calendars. Product changes, retuned thresholds, and updated references need versioned comparisons on independent cases. Operational verification frameworks provide the planning and data-management context (Pagano et al. 2024, 2025). Their institutional requirements should be used to maintain the technical record, rather than counted as additional evidence of rainfall skill.

Local feedback can identify a misleading target, delivery deadline, or presentation. It does not substitute for independent forecast evaluation. Likewise, statistical verification does not establish whether recipients understand the advisory or have resources to act (Lemos et al. 2012; Vedeld et al. 2020). Reporting these boundaries keeps regional forecast performance, local usability and agricultural outcomes connected without collapsing them into one service score.

8 Discussion: comparable rainfall evidence and research needs

8.1 What a comparable skill statement requires

The synthesis changes the specification of a rainfall performance statement, not the definition of forecast skill. A claim must identify its target, season or sowing window, lead, support, benchmark and rainfall reference. Amount error, onset timing and dry-spell probability may legitimately give different product rankings because they evaluate different quantities. A comparison is interpretable only when those quantities and the common sample are explicit. This reasoning applies to statistical, dynamical and machine-learning forecasts alike.

Three distinctions are central. First, a rainfall-season boundary inferred retrospectively is not an observed planting trigger available on that date. Second, a skill estimate relative to climatology is not comparable across regimes without its benchmark score and event frequency. Third, agreement with a rainfall reference does not establish accuracy for a locally thresholded event when the support or reference changes. The reporting sequence makes these distinctions visible

in an application, while Section 5 demonstrates only the arithmetic possibility of target and reference dependence.

West African evidence supports local specification rather than one regional performance claim. Gbangou et al. (2020) show index- and station-dependent forecast performance, Mendes et al. (2025) show why onset calendars vary, and Romanovska et al. (2023) show why agroclimatological interpretation depends on the input reference. These studies address different products and questions. Their combined implication is to document the conditions of each evaluation, rather than infer a uniform agricultural benefit or pool unlike skill measures.

The reporting logic can be used outside West Africa, but its portability remains a proposition for comparative application. Local thresholds, calendar windows, observational support and crop-water mechanisms must be re-established. A requirement that exposes a meaningful limitation in one setting may add little in another. ARFEF should therefore be judged by whether it improves the transparency and reproducibility of an evaluation, not by whether it produces a preferred forecast ranking.

8.2 Independent archive tests needed before operational claims

An empirical comparison should first specify one rainfall application and independently justified onset or dry-spell target. Obtain an archive whose daily resolution and horizon cover the target, with issue times and product versions retained. Declare training and test periods, the eligible calendar and regime, a climatological or operational benchmark, the primary reference and any decision baseline before inspecting results. A sample containing real subsequent dry-spell failures is needed to test that component of the planting-condition definition.

On the same independent cases, compare aggregate amount diagnostics with target-specific event or probabilistic performance. Then restrict to usable issue dates and test justified reference, threshold, and support alternatives. Report paired differences and uncertainty using blocks that preserve dependence. A ranking reversal is a possible outcome, not a criterion of success. An unchanged ranking can still reveal an unsupported actionability claim or sensitivity that was hidden by an aggregate score.

Testing the reporting synthesis itself requires assessing whether its use exposes unsupported inferences, improves reproduction, or changes a documented product-selection or advisory decision. This is distinct from showing that two metrics differ. Agricultural benefit requires a

588 further evaluation of delivery, response and outcomes with a credible comparison. The present
589 contribution is a critical synthesis with a constructed example; its operational implications
590 remain to be tested.

591 **8.3 Limitations**

592 The review is focused and interpretive. Its original search history is not a complete reproducible
593 screening record, and the targeted update does not establish exhaustive coverage through
594 October 2026. Country, rainfall-regime and production-system coverage is uneven. The
595 synthesis cannot estimate the regional prevalence of verification failures, a pooled forecast-value
596 effect, or the performance of present operational services.

597 No operational forecast archive was analysed. The eight constructed windows repeat a
598 deliberately simple pattern and contain identical subsequent wet-day timing. They do not
599 represent interannual variability, variable dry-spell occurrence, real reference errors, forecast
600 probability calibration, delivery timing, or farmer response. Their threshold and reference
601 changes demonstrate logical possibilities, not operational effect sizes or uncertainty. They
602 provide no empirical validation of ARFEF.

603 The reporting requirements do not resolve sparse rainfall observations, competing agricultural
604 objectives, or causal outcome attribution. Threshold sensitivity is informative only for
605 scientifically justified alternatives; it is not a licence to tune the event until a product wins.
606 Comparative archive applications are needed to determine which requirements materially alter
607 interpretation and which add work without improving an evaluation. These limits apply to the
608 review's contribution as well as to the agricultural claims it examines.

609 **9 Conclusions**

610 Rainfall timing makes forecast verification for agricultural climate services a target-specific
611 applied-climatology task. Seasonal accumulation, onset and dry-spell indices differ in temporal
612 definition, event frequency and observational support. Their skill estimates should not be treated
613 as interchangeable evidence for planting or crop-water decisions.

614 The review identifies the conditions needed for a defensible comparison: a specified rainfall
615 target and calendar window, complete forecast coverage and usable lead, matching spatial
616 support, an independently fitted seasonal or operational benchmark, and explicit reference

sensitivity. ARFEF records these conditions alongside the agricultural action and the evidence supporting the claim. Its contribution is the rainfall-specific reporting synthesis, not new verification theory or demonstrated operational improvement.

The constructed example confirms that low daily amount error can coexist with worse classification of a defined planting event, and that threshold or reference changes can alter the ranking. Its identical dry-spell sequences do not establish dry-spell prediction skill. Real forecast archives and subsequent service evaluations are needed to determine whether applying the reporting sequence changes useful product choices or agricultural outcomes.

Statements and Declarations

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Data and code availability

No new observational or operational forecast archive was analysed. The deliberately constructed rainfall arrays, calculation results, numerical reproduction script, and figure code are supplied in Online Resource 2 (ESM_2.zip). Appendix S4 in Online Resource 1 (ESM_1.pdf) specifies the construction and calculations. These files document the numerical illustration and do not represent observed rainfall or empirical forecast performance.

Ethics approval and consent

Not applicable. This review and its constructed numerical illustration involved no new research with human participants or animals.

Supplementary information

Online Resource 1 (ESM_1.pdf) contains the focused review approach and targeted update (Appendix S1), the evidence supporting the climatological and verification synthesis (Appendix S2), a rainfall-application reporting record (Appendix S3), and complete specifications for the constructed example (Appendix S4). Online Resource 2 (ESM_2.zip) contains constructed rainfall arrays, calculated results, numerical and figure reproduction scripts, and a scientific README.

References

Amegnaglo, C. J., Anaman, K. A., Mensah-Bonsu, A., Onumah, E. E., & Amoussouga Gero, F. (2017). Contingent valuation study of the benefits of seasonal climate forecasts for maize farmers in the Republic of Benin, West Africa. *Climate Services, 6*, 1-11

Bliefernicht, J., Waongo, M., Salack, S., Seidel, J., Laux, P., & Kunstmann, H. (2019). Quality and value of seasonal precipitation forecasts issued by the West African Regional Climate Outlook Forum. *Journal of Applied Meteorology and Climatology, 58*(3), 621-642. https://doi.org/10.1175/JAMC-D-18-0066.1

Coremberg, L., Materia, S., Delpiazzo, E., et al. (2024). A framework for joint verification and evaluation of seasonal climate services across socioeconomic sectors. *Bulletin of the American Meteorological Society, 105*(7), E1218-E1236. https://doi.org/10.1175/BAMS-D-23-0026.1

Dunning, C. M., Black, E. C. L., & Allan, R. P. (2016). The onset and cessation of seasonal rainfall over Africa. *Journal of Geophysical Research: Atmospheres, 121*(19), 11405-11424. https://doi.org/10.1002/2016JD025428

Emerton, R., van Andel, S. J., Egan, K., et al. (2026). How good is your forecast? The importance of user-centred, decision-relevant evaluation of forecasts and climate services. *Meteorological Applications, 33*(5), e70234. https://doi.org/10.1002/met.70234

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. *Scientific Data, 2*, 150066. https://doi.org/10.1038/sdata.2015.66

Gbangou, T., Ludwig, F., van Slobbe, E., Hoang, L., & Kranjac-Berisavljevic, G. (2019). Seasonal variability and predictability of agro-meteorological indices: Tailoring onset of rainy season estimation to meet farmers' needs in Ghana. *Climate Services, 14*, 19-30. https://doi.org/10.1016/j.cliser.2019.04.002

Gbangou, T., Ludwig, F., van Slobbe, E., Greuell, W., & Kranjac-Berisavljevic, G. (2020). Rainfall and dry spell occurrence in Ghana: Trends and seasonal predictions with a dynamical and a statistical model. *Theoretical and Applied Climatology, 141*, 371-387. https://doi.org/10.1007/s00704-020-03212-5

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359-378. https://doi.org/10.1198/016214506000001437

Hansen, J. W., Mason, S. J., Sun, L., & Tall, A. (2011). Review of seasonal climate forecasting for agriculture in sub-Saharan Africa. *Experimental Agriculture, 47*(2), 205-240. https://doi.org/10.1017/S0014479710000876

685 Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble
686 Prediction Systems. Weather and Forecasting, 15(5), 559-570. https://doi.org/10.1175/1520-
687 0434(2000)015<0559:DOTCRP>2.0.CO;2

688 Ingram, K. T., Roncoli, M. C., & Kirshen, P. H. (2002). Opportunities and constraints for
689 farmers of West Africa to use seasonal precipitation forecasts with Burkina Faso as a case
690 study. *Agricultural Systems, 74*(3), 331-349. https://doi.org/10.1016/S0308-521X(02)00044-6

691 Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2012). *Forecast verification: A practitioner's guide in*
692 *atmospheric science* (2nd ed.). Wiley-Blackwell. https://doi.org/10.1002/9781119960003

693 Katz, R. W., & Murphy, A. H. (1997). *Economic value of weather and climate forecasts*.
694 Cambridge University Press. https://doi.org/10.1017/CBO9780511608278

695 Kumi, N., Abiodun, B. J., & Adefisan, E. A. (2020). Performance evaluation of a subseasonal to
696 seasonal model in predicting rainfall onset over West Africa. *Earth and Space Science, 7*(8),
697 e2019EA000928. https://doi.org/10.1029/2019EA000928

698 Lemos, M. C., Kirchhoff, C. J., & Ramprasad, V. (2012). Narrowing the climate information
699 usability gap. *Nature Climate Change, 2*(11), 789-794. https://doi.org/10.1038/nclimate1614

700 Maidment, R. I., Grimes, D., Black, E., Tarnavsky, E., Young, M., Greatrex, H., Allan, R. P.,
701 Stein, T., Nkonde, E., Senkunda, S., & Alcántara, E. M. U. (2017). A new, long-term daily
702 satellite-based rainfall dataset for operational monitoring in Africa. *Scientific Data, 4*, 170063.
703 https://doi.org/10.1038/sdata.2017.63

704 Mendes, O., Correia, E., & Fragoso, M. (2025). Variability and trends of the rainy season in
705 West Africa with a special focus on Guinea-Bissau. *Theoretical and Applied Climatology,*
706 *156*, 242. https://doi.org/10.1007/s00704-025-05471-6

707 Meza, F. J., Hansen, J. W., & Osgood, D. (2008). Economic value of seasonal climate forecasts
708 for agriculture: Review of ex-ante assessments and recommendations for future research.
709 *Journal of Applied Meteorology and Climatology, 47*(5), 1269-1286.
710 https://doi.org/10.1175/2007JAMC1540.1

711 Murphy, A. H. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in
712 Weather Forecasting. Weather and Forecasting, 8(2), 281-293. https://doi.org/10.1175/1520-
713 0434(1993)008<0281:WIAGFA>2.0.CO;2

714 Nicholson, S. E. (2013). The West African Sahel: A review of recent studies on the rainfall
715 regime and its interannual variability. *ISRN Meteorology, 2013*, 453521.
716 https://doi.org/10.1155/2013/453521

717 Pagano, T. C., Casati, B., Landman, S., et al. (2024). Challenges of operational weather forecast
718 verification and evaluation. *Bulletin of the American Meteorological Society, 105*(4), E789-
719 E802. https://doi.org/10.1175/BAMS-D-22-0257.1

720 Pagano, T. C., Ebert, E. E., & Khanarmuei, M. (2025). Enhancing forecast verification in
721 national meteorological and hydrological services. *Meteorological Applications, 32*(3),
722 e70051. https://doi.org/10.1002/met.70051

723 Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment:
724 From days to decades. *Quarterly Journal of the Royal Meteorological Society, 128*(581), 747-
725 774. https://doi.org/10.1256/0035900021643593

726 Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction
727 system. *Quarterly Journal of the Royal Meteorological Society, 126*(563), 649-667.
728 https://doi.org/10.1002/qj.49712656313

729 Rodrigues, L. R. L., García-Serrano, J., & Doblas-Reyes, F. (2014). Seasonal forecast quality of
730 the West African monsoon rainfall regimes by multiple forecast systems. *Journal of*
731 *Geophysical Research: Atmospheres, 119*(13), 7908-7930.
732 https://doi.org/10.1002/2013JD021316

733 Romanovska, P., Gleixner, S., & Gornott, C. (2023). Climate data uncertainty for agricultural
734 impact assessments in West Africa. *Theoretical and Applied Climatology, 152*, 933-950.
735 https://doi.org/10.1007/s00704-023-04430-3

736 Roncoli, C. (2006). Ethnographic and participatory approaches to research on farmers' responses
737 to climate predictions. *Climate Research, 33*(1), 81-99. https://doi.org/10.3354/CR033081

738 Roncoli, C., Ingram, K., & Kirshen, P. (2002). Reading the rains: Local knowledge and rainfall
739 forecasting in Burkina Faso. *Society and Natural Resources, 15*(5), 409-427.
740 https://doi.org/10.1080/08941920252866774

741 Roncoli, C., Jost, C., Kirshen, P., et al. (2009). From accessing to assessing forecasts: An end-to-
742 end study of participatory climate forecast dissemination in Burkina Faso (West Africa).
743 *Climatic Change, 92*, 433-460. https://doi.org/10.1007/s10584-008-9445-6

744 Roudier, P., Sultan, B., Quirion, P., Baron, C., Alhassane, A., Traoré, S. B., & Muller, B. (2012).
745 An ex-ante evaluation of the use of seasonal climate forecasts for millet growers in SW Niger.
746 *International Journal of Climatology, 32*(5), 759-771. https://doi.org/10.1002/joc.2308

747 Roudier, P., Muller, B., D'Aquino, P., Roncoli, C., Soumaré, M. A., Batté, L., & Sultan, B.
748 (2014). The role of climate forecasts in smallholder agriculture: Lessons from participatory
749 research in two communities in Senegal. *Climate Risk Management, 2*, 42-55.
750 https://doi.org/10.1016/j.crm.2014.02.001

751 Roudier, P., Alhassane, A., Baron, C., Louvet, S., & Sultan, B. (2016). Assessing the benefits of
752 weather and seasonal forecasts to millet growers in Niger. *Agricultural and Forest*
753 *Meteorology, 223*, 168-180. https://doi.org/10.1016/j.agrformet.2016.04.010

754 Sanfo, S., Salack, S., Saley, I. A., Daku, E. K., Worou, N. O., Savadogo, A., Barro, H., Guug, S.,
755 Koné, H., Ibrahim, B., Rojas, A., Raymond, C., & Ogunjobi, K. O. (2022). Effects of

customized climate services on land and labor productivity in Burkina Faso and Ghana. *Climate Services, 25*, 100280. https://doi.org/10.1016/j.cliser.2021.100280

Siegmund, J., Bliefernicht, J., Laux, P., & Kunstmann, H. (2015). Toward a seasonal precipitation prediction system for West Africa: Performance of CFSv2 and high-resolution dynamical downscaling. *Journal of Geophysical Research: Atmospheres, 120*(15), 7316-7339. https://doi.org/10.1002/2014JD022692

Sivakumar, M.V. (1992). Empirical analysis of dry spells for agricultural applications in West Africa. Journal of Climate, 5, 532-539.

Sultan, B., Baron, C., Dingkuhn, M., Sarr, B., & Janicot, S. (2005). Agricultural impacts of large-scale variability of the West African monsoon. *Agricultural and Forest Meteorology, 128*(1-2), 93-110. https://doi.org/10.1016/j.agrformet.2004.08.005

Sultan, B., Barbier, B., Fortilus, J., Mbaye, S. M., & Leclerc, G. (2010). Estimating the potential economic value of seasonal forecasts in West Africa: A long-term ex-ante assessment in Senegal. *Weather, Climate, and Society, 2*(1), 69-87. https://doi.org/10.1175/2009WCAS1022.1

Tall, A., Coulibaly, J. Y., & Diop, M. (2018). Do climate services make a difference? A review of evaluation methodologies and practices to assess the value of climate information services for farmers: Implications for Africa. *Climate Services, 11*, 1-12. https://doi.org/10.1016/j.cliser.2018.06.001

Vaughan, C., Hansen, J., Roudier, P., Watkiss, P., & Carr, E. (2019). Evaluating agricultural weather and climate services in Africa: Evidence, methods, and a learning agenda. *Wiley Interdisciplinary Reviews: Climate Change, 10*(4), e586. https://doi.org/10.1002/wcc.586

Vedeld, T., Hofstad, H., Mathur, M., Bueker, P., & Stordal, F. (2020). Reaching out? Governing weather and climate services (WCS) for farmers. *Environmental Science and Policy, 104*, 208-216. https://doi.org/10.1016/j.envsci.2019.11.010

Vellinga, M., Arribas, A., & Graham, R. (2013). Seasonal forecasts for regional onset of the West African monsoon. *Climate Dynamics, 40*(11), 3047-3070. https://doi.org/10.1007/s00382-012-1520-z

Vogel, P., Knippertz, P., Gneiting, T., Fink, A. H., Klar, M., & Schlueter, A. (2021). Statistical forecasts for the occurrence of precipitation outperform global models over northern tropical Africa. *Geophysical Research Letters, 48*(3), e2020GL091022. https://doi.org/10.1029/2020GL091022

Wang, Y., Gueye, M., Greybush, S. J., Greatrex, H., Whalen, A. J., Ssentongo, P., Zhang, F., Jenkins, G. S., & Schiff, S. J. (2023). Verification of operational numerical weather prediction model forecasts of precipitation using satellite rainfall estimates over Africa. *Meteorological Applications, 30*(1), e2112. https://doi.org/10.1002/met.2112

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Elsevier. https://doi.org/10.1016/C2017-0-03921-6

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf