

Response to Reviewer – Major Comments

The responses below retain the authors' proposed methodological approach while clarifying the scientific rationale and ensuring that each response directly addresses the reviewer's concern.

Comment 1. Predictor selection and information leakage

Reviewer comment:

The manuscript should clearly explain whether correlation-based basin identification and SHAP feature selection were performed independently within each training fold. If the full dataset was used before validation, information leakage could inflate the reported forecast skill. A nested or strictly out-of-sample selection procedure is recommended.

Response:

We thank the reviewer for raising this important methodological concern. We agree that predictor-basin identification and feature selection should be independent of the data used for forecast verification to avoid information leakage.

In the revised analysis, the 86 historical MAM seasons (1940–2025) will be divided into approximately 75% for model development (64 seasons) and 25% for independent verification (22 seasons). The 64 development seasons will be used for predictor-basin identification, statistical screening, feature selection and machine learning model training. The remaining 22 seasons will be excluded from all model-development procedures and reserved exclusively for independent forecast verification.

To assess the robustness of the identified basins, we will adopt a subsampling approach inspired by the Complementary Pairs Stability Selection framework of Shah and Samworth (2013). The 64 development seasons will be randomly partitioned three times into complementary pairs of 32 seasons, producing six subsets. MCA-based basin identification and regression screening ($p \leq 0.05$) will be performed independently for each subset. Basin stability will be assessed based on geographical overlap and detection frequency, allowing for moderate spatial shifts. Basins identified in at least three of the six subsets ($\geq 50\%$) will be considered stable candidates, with those appearing in at least four subsets ($\geq 66.7\%$) classified as more robust.

The final predictor selection and model training will use only the development dataset, while forecast skill will be evaluated using the 22 withheld seasons. This approach will improve the reproducibility of predictor identification and provide a more rigorous, leakage-free assessment of forecast performance.

I think it is better to directly show the improved results and findings, and mention how the new results have ended up in the revised paper.

Comment 2. Physical drivers versus statistical predictors

Reviewer comment:

The terms “physical controls,” “drivers,” and “influence” appear stronger than supported by correlation and ML methods alone. The authors should distinguish predictive association from causality and moderate the wording unless stronger dynamical evidence is provided.

Response:

We agree with the reviewer that statistical and machine-learning relationships alone should not be interpreted as direct evidence of causality.

To strengthen the interpretation of the selected predictors, we will perform additionally regression analyses between each retained climate predictor and the corresponding zonal MAM rainfall variability. These analyses allow us to evaluate the statistical significance, sign, and spatial structure of the relationship between each selected predictor and rainfall within the individual rainfall zones.

The purpose of this additional analysis is not to establish causality, but rather to demonstrate that the predictors retained by the machine-learning framework have statistically identifiable relationships with the rainfall variability they are being used to predict.

We will therefore revise the terminology throughout the manuscript. Expressions such as “physical controls,” “influence,” or statements implying direct causality will be replaced, where appropriate, by more precise terms such as “predictors,” “associated climate signals,” “predictive relationships,” or “physically plausible teleconnections.”

The regression analysis will consequently be presented as additional evidence supporting the physical and statistical relevance of the selected predictors, rather than as proof of causal influence.

Comment 3. Justification of EOF1/PC1

Reviewer comment:

The choice of rainfall PC1 as the predictand needs stronger justification. The authors should report its explained variance and spatial pattern and explain why it is preferable to a regional-mean rainfall index. A sensitivity test using regional mean MAM rainfall would strengthen the analysis.

Response:

We thank the reviewer for this comment. After reconsidering the use of PC1, we agree that representing regional rainfall variability using a single leading mode may unnecessarily constrain the spatial rainfall signal used during predictor identification.

In the revised methodology, we will therefore no longer use EOF1/PC1 as the rainfall field for identifying predictor relationships. Instead, predictor-basin identification will be based directly on standardized grid-point MAM rainfall anomalies. <ADD A SENTENCE OR TWO THAT JUSTIFIES THIS REVISED APPROACH>.

This approach allows the analysis to retain the spatial variability of rainfall across the study domain without requiring the rainfall field to be represented by a single EOF mode. It is particularly appropriate for East Africa, where MAM rainfall variability is spatially heterogeneous and different rainfall zones may respond differently to the same large-scale climate signal.

The standardized grid-point rainfall anomaly will therefore be used to identify coherent relationships between antecedent predictor fields and rainfall variability, while model training and verification will continue to be conducted separately for the defined rainfall zones.

We believe this revised approach provides a more direct representation of the rainfall predictand and removes the need to justify the use of PC1 relative to a regional mean index.

The Methods and associated figures will be revised accordingly and the revised manuscript will have improved figures

Comment 4. Multiple testing

Reviewer comment:

The use of $p < 0.1$ for global grid-point correlation screening may produce false-positive regions because many correlations are tested. The authors should address multiple comparisons and provide sensitivity tests using alternative significance thresholds or field-significance methods.

Response:

We agree that the use of $p < 0.1$ requires clarification, particularly when examining correlations over a large spatial domain.

We will use $p < 0.05$ for both driver basin identification as well as for final driver screening before model training and prediction

Comment 5. Additional value of T2 predictors

Reviewer comment:

The conclusion that T2 provides information beyond SST should be demonstrated quantitatively. Comparisons of SST-only, T2-only, SLP-only, and combined models would show whether T2 predictors provide genuine incremental forecast skill.

Response:

We thank the reviewer for this observation and agree that the wording in the original manuscript may have unintentionally implied that T2 predictors provide greater or independent forecast skill relative to SST.

The primary objective of our study was not to isolate the individual predictive contribution of T2 relative to SST, SLP, or the other predictor families. Rather, our intention was to highlight the relatively frequent occurrence and prominence of T2-related predictors among the predictors identified by the feature-selection procedure.

Thus, our use of the term “dominance” refers to the frequency and relative importance with which T2 predictors were selected within the machine-learning framework, rather than demonstrating that T2 alone provides greater forecast skill or statistically significant incremental skill beyond SST.

We agree that demonstrating incremental predictive value would require a dedicated predictor-ablation experiment comparing, for example, SST-only, T2-only, SLP-only, and combined models. Such an attribution analysis is beyond the primary objective of the present study.

We will therefore revise the relevant statements throughout the manuscript to avoid suggesting that T2 provides independent or superior predictive skill relative to SST. Instead, we will state that T2 predictors emerged frequently among the selected predictors and therefore appear to contain potentially useful predictive information within the combined predictor framework.

This revision more accurately reflects the evidence provided by the present analysis.

Comment 6. Probabilistic skill assessment

Reviewer comment:

The reported BSS and AUC values are promising, but the reference forecast, forecast categories, sample size, validation procedure, and uncertainty estimates should be clearly described. The meaning of “spatially averaged” BSS should also be clarified given that PC1 is described as the target.

Response:

We agree with the reviewer that the description of the forecast-verification procedure requires greater clarity.

In the revised manuscript, we will provide a more detailed description of the forecast categories, reference forecast, validation sample, skill-score calculation, and spatial averaging procedure.

The reported forecast skill was calculated using 22 seasons selected as independent test seasons from the available historical record. For each test experiment, these seasons were excluded from model training and subsequently used for forecast verification. The probabilistic verification metrics, including BSS and AUC, were first calculated at individual rainfall grid points and were subsequently averaged spatially over the relevant rainfall zone.

We will revise the Methods section to clearly state how the 22 test seasons were selected, how the remaining seasons were used for model training, the probability thresholds used to define below-normal, near-normal, and above-normal rainfall categories, the climatological/reference forecast used in calculating BSS, how AUC was calculated for each rainfall category, and how grid-point skill scores were spatially aggregated to produce the reported zonal skill.

Because the revised predictor-identification procedure no longer uses PC1 as the rainfall predictand, the apparent inconsistency between PC1 and spatially averaged grid-point skill will also be removed.

We agree that these additions will make the probabilistic verification framework substantially clearer and more reproducible.

Comment 7. Predictor robustness

Reviewer comment:

The stability of the selected predictors across different training samples should be assessed. Reporting predictor-selection frequency or bootstrap stability would help establish whether the identified climate signals are robust rather than sample-dependent.

Response:

We will consider driver basin stability. See our response on comment 1 above

Comment 8. Physical mechanism

Reviewer comment:

The proposed SST–T2–SLP–circulation pathway is plausible, but the current analysis appears mainly associative. Additional lagged relationships, circulation composites, moisture convergence, or moisture-budget diagnostics would strengthen the physical interpretation.

Response:

We thank the reviewer and agree that the present analysis primarily identifies statistical associations rather than demonstrating a complete causal SST–T2–SLP–circulation pathway.

The discussion of the SST–T2–SLP relationship in the manuscript is based on established theoretical understanding of interactions among ocean-surface temperature anomalies, atmospheric temperature, pressure gradients, large-scale circulation, and rainfall variability. Our objective was to determine whether the climate predictors identified by the statistical and machine-learning framework are physically consistent with these established processes, rather than to demonstrate the complete causal chain experimentally.

A rigorous attribution of the SST–T2–SLP–circulation pathway would require a substantially different analysis, potentially involving detailed moisture-budget diagnostics, dynamical sensitivity experiments, atmospheric model perturbation experiments, or dedicated process-based analyses. Such an investigation is beyond the scope of the present forecasting-focused study.

Nevertheless, we agree that the original wording may have conveyed greater mechanistic certainty than is supported by the analysis. We will therefore revise the discussion to clearly distinguish between the statistical associations identified in this study, their consistency with established physical theory, and direct demonstration of the underlying causal mechanism.

The SST–T2–SLP relationships will consequently be described as physically plausible interpretations of the identified predictive associations rather than as a demonstrated causal pathway.

We will also explicitly identify detailed process-based attribution of these relationships as an important direction for future research.

Comment 9. Operational relevance

Reviewer comment:

The potential application to ICPAC and other climate services is valuable, but operational readiness should not be implied without independent hindcasts or real-time testing. The wording should be moderated to indicate potential operational usefulness.

Response:

We agree with the reviewer. The intention of the manuscript is not to claim that the proposed framework is already an operational seasonal forecasting system, but rather to demonstrate its potential relevance for future operational seasonal forecasting applications.

We will therefore revise the manuscript to clearly distinguish between demonstrated retrospective forecast skill and operational readiness.

The framework will be described as having potential operational utility, subject to further independent evaluation. In particular, we acknowledge that operational implementation would require additional testing using independent hindcasts, prospective real-time forecasts, and evaluation across seasons not involved in model development.

The discussion and conclusions will therefore be moderated accordingly, and statements implying current operational readiness will be removed.