International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

Base-Learner Correlation Constrains The Benefit Of Consensus Machine Learning For Radiosonde-Based Convective Forecasting In The Tropics

Dima Soraya[1], Aulia Khamas Heikmakhtiar[1], Leli Setyaningrum[1]

[1]Cyber Defense Engineering Study Program, Faculty of Defense Engineering and Technology

Republic of Indonesia Defense University (RIDU), Bogor 16810, Indonesia

Corresponding author: Dima Soraya, dima.soraya@tp.idu.ac.id

Abstract— Ensemble and consensus methods are widely assumed to improve convective forecasting skill, yet the conditions under which that assumption holds are rarely tested. This study evaluates four consensus schemes built on three algorithmically diverse base learners for 0–12 h rain and thunderstorm prediction from radiosonde-derived instability indices at Soekarno-Hatta International Airport (WMO 96749), Indonesia. In total, 5,839 soundings (2016–2025) were converted into twelve instability indices and labelled against 173,865 METAR reports using an exclusive-lower-bound forecast window. Random Forest, Support Vector Machine, and Gradient Boosting were combined through hard voting, soft voting, skill-weighted voting, and regularised-weight optimisation on a chronologically partitioned, leakage-controlled pipeline. Peak critical success index (CSI) reached 0.460 for rain and 0.428 for thunderstorm. Consensus clearly outperformed a climatological random predictor (permutation test, $p < 0.0001$) but showed no significant advantage over the best single model (McNemar $p = 1.0000$ and $p = 0.2807$; paired-bootstrap ΔCSI intervals spanning zero). We attribute this plateau to high inter-model prediction correlation (Pearson $r = 0.753$–0.971), which leaves little independent error for voting to exploit and which follows from training all learners on an identical feature space. Precipitable water was the dominant predictor for both targets. Ensemble diversity in this problem is therefore limited by information source rather than algorithm choice, and multimodal inputs are more promising than additional combination rules.

Keywords: consensus machine learning; ensemble diversity; radiosonde; thunderstorm nowcasting; forecast verification; tropical convection.

I. INTRODUCTION

Convective weather remains one of the principal hazards to aviation. Cumulonimbus development produces turbulence, wind shear, microbursts, lightning, and visibility reduction, and these effects concentrate in the take-off and landing phases where operational margins are smallest [10]. Forecasting convective initiation in the tropics is particularly difficult. The Indonesian Maritime Continent sustains frequent deep convection with strong spatial and temporal variability, and the relationship between the pre-convective environment and subsequent occurrence is markedly nonlinear [6], [14].

Radiosonde soundings occupy a distinctive position among observing systems for this problem. Radar and satellite imagery characterise convection that has already organised, whereas soundings measure the thermodynamic state *in situ* before initiation, which is precisely the information required for anticipatory guidance [20]. Instability indices derived from soundings, including CAPE, Lifted Index, K Index, and precipitable water, have long served as convective indicators. Their discriminating power is, however, regionally variable. Haklander and Van Delden [9] documented substantial differences in index skill for the

Netherlands. Over Indonesia, Sagita and Takemi [18] found that surface-based CAPE separates thunderstorm from non-thunderstorm days poorly (AUC = 0.69), while deep-layer moisture indices perform considerably better, with precipitable water reaching AUC = 0.82. The more general difficulty of relying on individual indices and parameters in severe-storm forecasting has been discussed at length by Doswell and Schultz [7].

The limitations of single-threshold approaches motivated the application of machine learning, which can represent multivariate nonlinear structure without prescribing its functional form [15], [17]. A recurring finding across this literature is that no single algorithm dominates across datasets and regimes. That observation is the standard justification for ensemble learning, in which the errors of one learner are expected to be compensated by others [5], [21]. Consensus schemes, which aggregate predictions through voting or weighted probability averaging, are a direct expression of this idea and are commonly presented as a safe default.

The theoretical basis of that expectation is more restrictive than its practical usage suggests. Ensemble benefit scales with the diversity of member errors; when members agree, aggregation cannot recover information that no member possesses. This condition is frequently asserted in the meteorological machine learning literature but rarely quantified alongside the reported skill improvement, with the consequence that a null or marginal ensemble result is difficult to interpret. A related ambiguity concerns the level at which consensus is formed. Bondyopadhyay and Mohapatra [2] obtained clear benefit by combining heterogeneous thermodynamic indices, which is a consensus over information sources. Combining several algorithms trained on one shared feature vector is a different operation, and the two are not interchangeable.

This study takes up that question with a decade of soundings from a single well-instrumented tropical airport. We build four consensus schemes over three algorithmically distinct base learners, compare them with the single models under meteorological verification metrics, and test the comparison with permutation, McNemar, and paired-bootstrap procedures. We then measure inter-model prediction correlation directly, since that quantity determines how much voting can recover. The aim is diagnostic rather than competitive: we ask what limits consensus benefit in this setting and report the limit quantitatively, so that it can inform ensemble design in comparable problems.

II. DATA AND METHODS

A. Site and observations

Observations were taken at Soekarno-Hatta Meteorological Station (ICAO WIII, WMO 96749), serving Jakarta's principal international airport. The station performs routine radiosonde ascents at 00 and 12 UTC and issues METAR reports, providing collocated upper-air and surface records. The analysis period spans 1 January 2016 to 31 December 2025.

Sounding data comprising 5,839 profiles were obtained from the University of Wyoming Upper Air Archive, and 173,865 METAR reports at nominal 30-minute resolution were obtained from the Iowa Environmental Mesonet. Twelve instability indices were derived from each profile and used as the complete predictor set: CAPE, CIN, Lifted Index, Showalter Index, K Index, Cross Totals, Vertical Totals, Total Totals, SWEAT, level of free convection, equilibrium level, and precipitable water. Temporal predictors such as hour and month were deliberately excluded so that measured skill is attributable to thermodynamic information rather than to learned climatology. The complete processing pipeline is summarised in Fig. 1.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

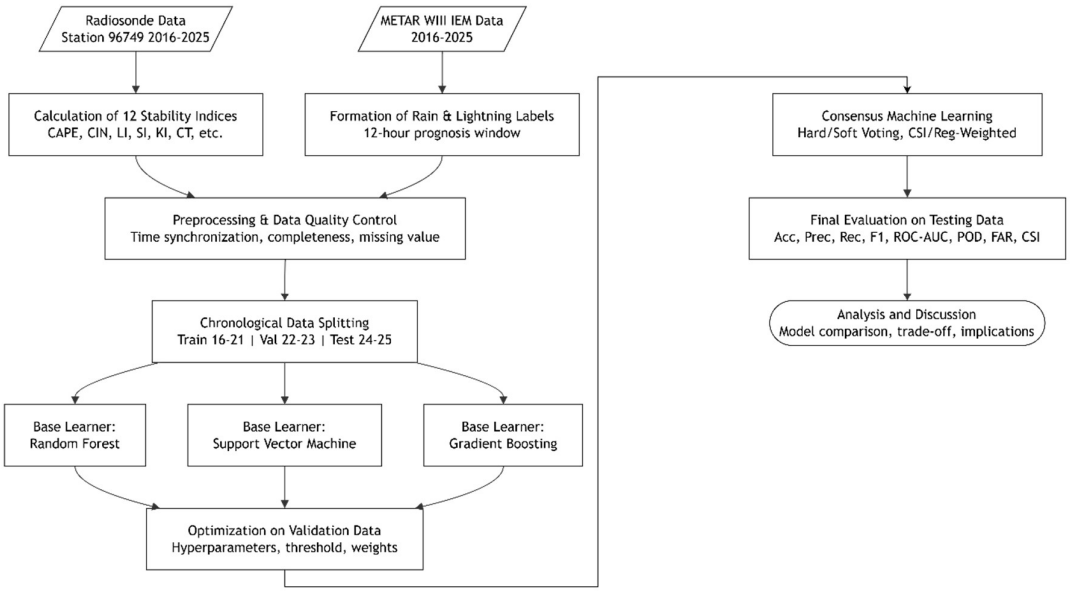

Fig. 1. Processing pipeline from radiosonde soundings and METAR reports to labelled data, base learners, and consensus schemes.

B. Target definition

Two independent binary targets were constructed from METAR present-weather groups. Labelling operated on tokenised weather groups rather than substring matching, which avoids two systematic errors: naive matching assigns VCRA to rain and, more damagingly, assigns TSNO, a sensor-status indicator denoting an inoperative lightning detector, to thunderstorm.

Treatment of the vicinity qualifier is deliberately asymmetric between targets and follows operational impact rather than notational consistency. For rain, VC-prefixed tokens are excluded because precipitation adjacent to the aerodrome does not affect runway state. For thunderstorm, VCTS is retained as positive because nearby electrical activity and associated wind shear constrain ground handling and departure operations. Drizzle (DZ) is excluded from the rain target because its stratiform microphysical origin is not represented by convective instability indices, and Cumulonimbus cloud reports without an accompanying TS group are not treated as thunderstorms, since cloud presence does not establish electrical activity at observation time. Both labelling functions were verified against 27 blocking unit tests, all of which passed.

C. Forecast window and synchronisation

Each sounding at time t was paired with all METAR reports in the interval (t, t + 12 h], with the lower bound exclusive. The target takes the value one if at least one report within the window satisfies the event criterion. Exclusivity guarantees that labels describe conditions strictly after the profile was observed, preserving the temporal ordering required of a predictive relationship, following the pairing principle applied by Nayak and Mandal [16]. The window length matches the interval between successive ascents, so windows tile the record without overlap and no event contributes to more than one sample.

D. Quality control and partitioning

Two sequential quality-control rules were applied. The first discarded soundings whose forecast window contained fewer than six METAR reports, since sparsely reported windows cannot reliably establish a negative label. The second applied complete-case analysis to rows with any missing index, without imputation. In this dataset the first rule removed two soundings and the second removed none, reducing the sample from 5,839 to 5,837 with negligible change in prevalence.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

Data were partitioned chronologically by year to emulate operational deployment, in which a model trained on historical data is applied to future observations (Table 1). Because undetected leakage is a documented source of over-optimistic performance estimates in applied machine learning [12], its absence was verified computationally by confirming that the pairwise intersections of subset indices were empty and that the partition was exhaustive.

Table 1. Chronological partitioning of the dataset.

Subset	Period	N	Rain prevalence	Thunderstorm prevalence
Training	2016–2021	3,484	0.3542	0.3232
Validation	2022–2023	1,309	0.3316	0.3079
Testing	2024–2025	1,044	0.3592	0.3228

E. Base learners and threshold selection

Three base learners were selected to span distinct inductive biases: Random Forest, which reduces variance through bootstrap aggregation with random feature subsetting [3]; Support Vector Machine with a radial basis kernel, which optimises a maximum-margin decision boundary [4]; and Gradient Boosting, which reduces bias through sequential residual correction [8]. Diversity of learning paradigms is a precondition for ensemble benefit. Whether it translates into diversity of predictions is examined in Section III-D.

Standardisation was applied selectively to the SVM only, since kernel distances are scale-sensitive whereas tree-based learners are invariant under monotonic transformation. Scaling parameters were estimated on the training subset alone and applied unchanged to validation and testing.

Hyperparameters were selected by grid search, with each configuration trained on the training subset and scored on validation. For every configuration the decision threshold was swept from 0.01 to 0.99 in steps of 0.01 and set to maximise validation CSI. Threshold optimisation was the sole mechanism for handling class imbalance; no resampling, class weighting, or probability calibration was applied, so reported probabilities are the raw model outputs. All stochastic components used a fixed random seed.

F. Consensus schemes

Four consensus mechanisms were constructed, spanning the range from unweighted majority logic to constrained weight optimisation.

C1, hard voting. A positive prediction is issued when at least two of three base learners predict positive, each applying its own optimised threshold rather than 0.5. Being already binary, C1 requires no further threshold tuning.

C2, soft voting. Predicted probabilities are averaged with uniform weights.

$$p_{C2}(x) = \frac{1}{3}[p_{RF}(x) + p_{SVM}(x) + p_{GB}(x)] \quad (1)$$

C3, skill-weighted voting**.** Weights are proportional to each learner's validation CSI at its optimal threshold, so that better-validating models exert proportionally greater influence.

$$w_m = \frac{\mathrm{CSI}_m}{\sum_k \mathrm{CSI}_k}, \quad p_{C3}(x) = \sum_m w_m p_m(x) \quad (2)$$

C4, regularised-weight optimisation. Weights minimise validation log-loss subject to a penalty that shrinks them toward uniformity, preventing weight overfitting on a finite validation set. The problem is solved by sequential least-squares programming under non-negativity and sum-to-one constraints, with λ chosen from {0.05, 0.1, 0.5, 1.0} by validation CSI.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

$$\min_w \text{LogLoss}(\sum_m w_m p_m) + \lambda \sum_m \left(w_m - \frac{1}{3}\right)^2, \quad \text{subject to} \quad w_m \geq 0, \quad \sum_m w_m = 1 \tag{3}$$

The three probability-based schemes then underwent independent threshold optimisation on validation data.

G. Verification and inference

All seven methods were evaluated once on the held-out testing subset. Primary metrics were probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI), supplemented by accuracy, precision, F1, and ROC-AUC. These categorical verification measures are standard in operational forecast assessment [11], [19].

$$\text{POD} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FAR} = \frac{\text{FP}}{\text{TP} + \text{FP}}, \quad \text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

Internal consistency was enforced by two blocking checks: confusion-matrix totals were required to equal the test-set size for every method, and every reported metric was recomputed from the confusion matrix with zero tolerance for discrepancy.

Three hypotheses were tested on the same test-set predictions without retraining. A permutation test with 10,000 permutations assessed CSI against a prevalence-matched random predictor. A two-sided McNemar test with continuity correction compared correct-classification proportions between consensus and the best single model. A paired bootstrap with 10,000 resamples assessed POD superiority jointly with FAR non-inferiority at an *a priori* margin of 0.02. The consensus scheme entering these tests was selected on validation CSI, never on test CSI, so that test data did not influence model selection.

III. RESULTS

A. Data characteristics and base learner configuration

Labelling yielded sounding-level prevalences of 0.3499 for rain and 0.3196 for thunderstorm. The median forecast window contained 24 METAR reports with a maximum of 29, consistent with the expected 24 routine reports per 12-hour window plus occasional SPECI reports. Prevalences across the three subsets remained within 0.31–0.36 for both targets, indicating no severe label-distribution shift between the training period and the 2024–2025 evaluation period.

Class-conditional statistics are physically coherent. For the thunderstorm target, mean CAPE rises from 579.9 to 854.7 J kg^{-1} between non-event and event soundings, CIN weakens from −83.2 to −56.1 J kg^{-1}, and precipitable water increases from 52.4 to 57.7 mm. Distributional overlap between classes nevertheless remains substantial, confirming that no single index threshold can resolve the classification.

Optimal thresholds for all six target–learner combinations fell between 0.26 and 0.37, well below 0.5. This is the expected consequence of operating at roughly one-third prevalence: the conventional 0.5 cut is overly conservative and suppresses detection. Validation CSI values were closely clustered, spanning 0.4202–0.4248 for rain and 0.3909–0.4101 for thunderstorm.

B. Consensus weights

Estimated consensus weights are given in Table 2. For rain, all schemes converged to near-uniform weights between 0.325 and 0.348, which follows directly from the near-identical validation skill of the three learners. For thunderstorm, C4 departed markedly from uniformity, downweighting the SVM to 0.157 and upweighting Gradient Boosting to 0.471, consistent with validation log-loss identifying the SVM as the weakest probabilistic contributor. The selected penalty $\lambda^* = 0.05$ was the weakest on the grid, so the optimiser was permitted to move away from uniform weights and did so only where the data supported it.

Table 2. Consensus weights by scheme and target.

Target	Scheme	w(RF)	w(SVM)	w(GB)	Note
Rain	C2 soft	0.3333	0.3333	0.3333	uniform
Rain	C3 skill-weighted	0.3351	0.3314	0.3334	$\propto$ validation CSI
Rain	C4 regularised	0.3480	0.3253	0.3267	$\lambda^* = 0.05$
Thunderstorm	C2 soft	0.3333	0.3333	0.3333	uniform
Thunderstorm	C3 skill-weighted	0.3285	0.3277	0.3438	$\propto$ validation CSI
Thunderstorm	C4 regularised	0.3722	0.1568	0.4710	$\lambda^* = 0.05$

C. Test-set performance

Test-set results for both targets are reported in Tables 3 and 4.

Table 3. Test-set verification for the rain target (N = 1,044).

Method	Thr	Acc	Prec	POD	F1	AUC	FAR	CSI
Random Forest	0.37	0.6772	0.5354	0.7653	0.6301	0.7666	0.4646	0.4599
SVM	0.29	0.6849	0.5449	0.7440	0.6291	0.7811	0.4551	0.4589
Gradient Boosting	0.26	0.6216	0.4854	0.8880	0.6277	0.7642	0.5146	0.4574
C1 hard	—	0.6638	0.5211	0.7920	0.6286	0.7331	0.4789	0.4583
C2 soft	0.34	0.6753	0.5345	0.7440	0.6221	0.7748	0.4655	0.4515
C3 skill-weighted	0.34	0.6762	0.5354	0.7467	0.6236	0.7747	0.4646	0.4531
C4 regularised	0.34	0.6753	0.5342	0.7493	0.6238	0.7746	0.4658	0.4532

AUC for C1 is computed from discrete vote fractions and is not strictly comparable with the continuous probabilistic methods.

Table 4. Test-set verification for the thunderstorm target (N = 1,044).

Method	Thr	Acc	Prec	POD	F1	AUC	FAR	CSI
Random Forest	0.36	0.6638	0.4870	0.7774	0.5989	0.7464	0.5130	0.4274
SVM	0.28	0.6571	0.4801	0.7507	0.5856	0.7228	0.5199	0.4141
Gradient Boosting	0.32	0.6331	0.4623	0.8368	0.5956	0.7345	0.5377	0.4241
C1 hard	—	0.6571	0.4811	0.7953	0.5996	0.7216	0.5189	0.4281
C2 soft	0.31	0.6427	0.4690	0.8071	0.5932	0.7476	0.5310	0.4217
C3 skill-weighted	0.31	0.6418	0.4680	0.8042	0.5917	0.7476	0.5320	0.4202
C4 regularised	0.32	0.6418	0.4683	0.8101	0.5935	0.7462	0.5317	0.4219

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

The dominant feature of both tables is compression. For rain, CSI spans 0.4515 to 0.4599, a total range of 0.0084 across seven methods; for thunderstorm the range is 0.4141 to 0.4281. The best rain CSI belongs to Random Forest, a single model. The best thunderstorm CSI belongs to hard voting, but its margin over Random Forest is 0.0007, which is operationally indistinguishable. Gradient Boosting attains the highest POD for both targets (0.8880 and 0.8368) at correspondingly elevated FAR, reflecting the aggressive operating point implied by its low threshold. Probability-averaging schemes raise thunderstorm POD relative to Random Forest, from 0.7774 to between 0.8042 and 0.8101, but incur proportionate FAR increases so that net CSI does not improve. Fig. 2 shows the corresponding ROC curves for both targets.

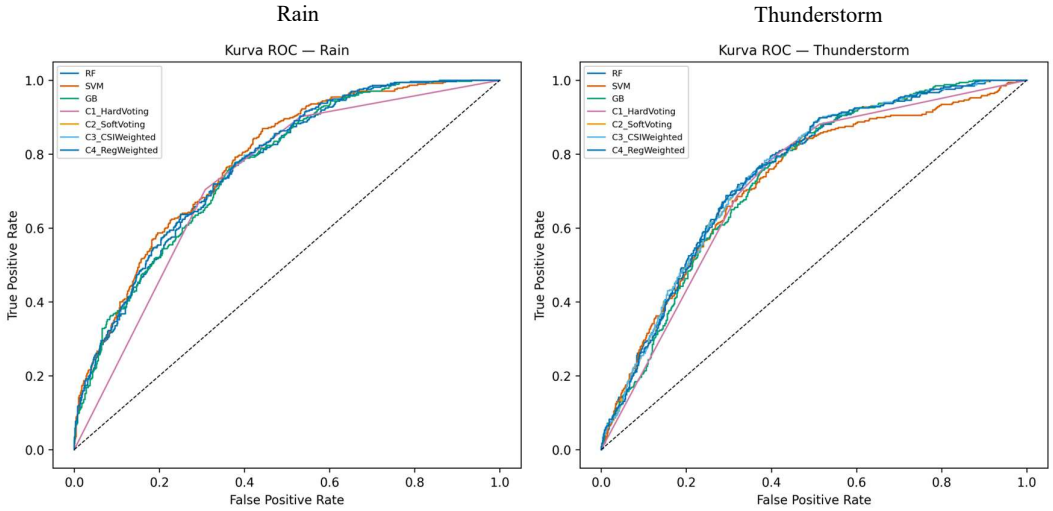

Fig. 2. ROC curves for all seven methods on the test set, for the rain and thunderstorm targets.

D. Inter-model prediction correlation

Pearson correlations between base-learner predicted probabilities on the test set are reported in Table 5.

Table 5. Pearson correlation between base-learner predicted probabilities (test set).

Target	RF–SVM	RF–GB	SVM–GB
Rain	0.9234	0.9711	0.9173
Thunderstorm	0.7656	0.9267	0.7526

Correlations of 0.92–0.97 for rain indicate that the three learners agree almost completely and therefore err on largely the same cases. Thunderstorm correlations involving the SVM are lower, near 0.75, indicating somewhat greater prediction diversity. The correspondence with the skill results is direct: the only instance in which a consensus scheme attained the highest CSI occurred for the target with the lower inter-model correlation.

E. Inferential tests

The permutation test rejected the null of random-predictor equivalence for both targets and all consensus mechanisms at $p < 0.0001$, with observed CSI lying approximately 13–14 standard deviations above the null mean of about 0.26. Radiosonde-derived indices therefore carry substantial predictive information, which is the necessary precondition for the subsequent comparison.

The comparison against the best single model did not reject the null. For rain, discordant pairs were almost exactly balanced, with 28 cases correct under consensus but not the single model and 29 in the reverse direction, giving $p = 1.0000$. For

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

thunderstorm the imbalance favoured consensus, 32 against 23, but did not reach significance ($p = 0.2807$). In the paired bootstrap, FAR non-inferiority was satisfied for rain, with the upper confidence bound on ΔFAR at 0.0143 against a margin of 0.02, but POD superiority failed because the ΔPOD interval included zero. Bootstrap ΔCSI intervals included zero for both targets. The paired-bootstrap distributions are shown in Fig. 3.

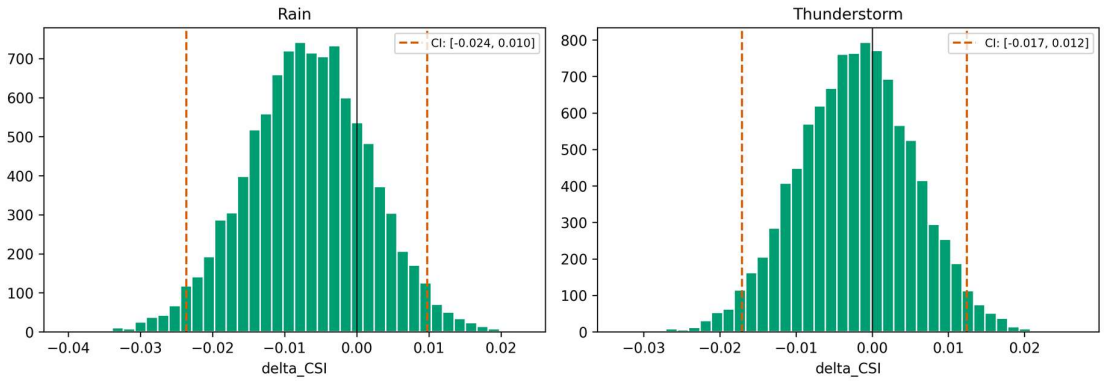

Fig. 3. Paired-bootstrap distributions of ΔCSI between the best consensus scheme and the best single model.

F. Feature attribution

Global attribution based on mean absolute SHAP values [13] aggregated across the three learners is summarised in Table 6.

Table 6. Five highest-ranked predictors by mean |SHAP| across three models.

Rank	Rain	mean |SHAP|	Thunderstorm	mean |SHAP|
1	Precipitable water	0.3203	Precipitable water	0.2158
2	K Index	0.0731	CAPE	0.1211
3	CIN	0.0696	SWEAT	0.0954
4	Level of free convection	0.0648	Cross Totals	0.0847
5	SWEAT	0.0628	K Index	0.0632

Precipitable water ranks first for both targets by a wide margin, exceeding the second-ranked predictor by nearly a factor of five for rain. This is physically expected, since column-integrated moisture is a direct precondition for precipitation. The thunderstorm ranking places CAPE second, followed by SWEAT and Cross Totals, so that the moisture-dominated rain structure gives way to a greater role for convective energy and shear-related quantities. These results independently reproduce the finding of Sagita and Takemi [18] that deep-layer moisture rather than CAPE is the strongest univariate discriminator over Indonesia. They extend it by placing precipitable water inside a multivariate nonlinear model, where its interactions with eleven other indices are also represented.

Variance inflation analysis showed that Total Totals is the exact sum of Cross Totals and Vertical Totals, producing exact collinearity. Attribution is therefore interpreted at the level of concept groups, namely moisture, convective energy, and instability, rather than at the level of individual indices. Collinearity of this kind degrades the interpretability of individual attributions but not the predictive performance of tree ensembles or kernel machines. Fig. 4 ranks the predictors by aggregated SHAP importance.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 1016-1027

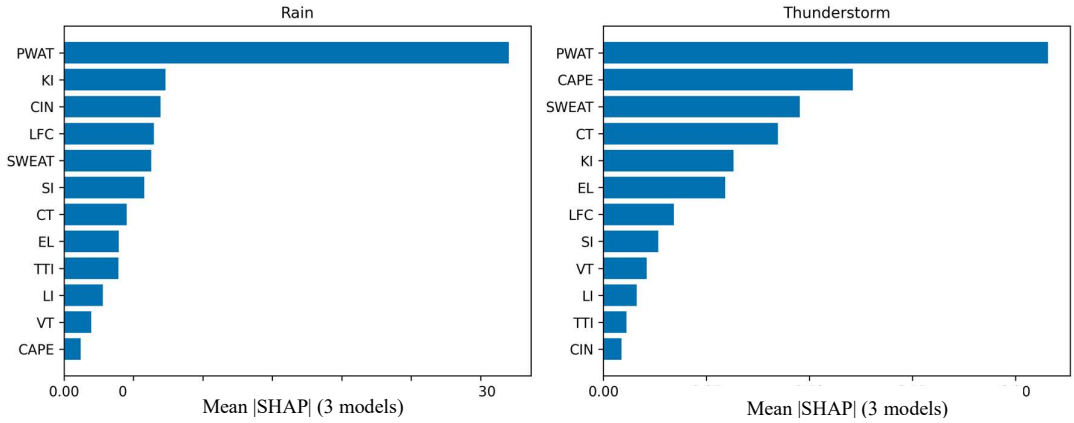

Fig. 4. Predictors ranked by mean |SHAP| value aggregated over the three base learners, for both targets.

IV. DISCUSSION

A. Why consensus does not improve on its best member

The central result is a null with a mechanism. Consensus schemes matched but did not exceed the best single learner, and the reason is visible in the correlation structure rather than in the combination rules. Weighted voting can only recover cases on which members disagree; when pairwise prediction correlations reach 0.92–0.97, the disagreement set is small and its composition is close to random with respect to the label. The near-exact balance of McNemar discordant pairs for rain, 28 against 29, is precisely what that condition predicts.

This also explains the one partial exception. Thunderstorm prediction exhibits lower correlations involving the SVM, near 0.75, and it is for that target that hard voting attains the highest CSI and that C4 finds genuinely non-uniform weights. The magnitude remains small, but the direction is consistent: consensus benefit tracks prediction diversity, not the sophistication of the aggregation rule. The four schemes tested here span uniform averaging, skill weighting, and constrained optimisation, and they differ from one another by less than 0.005 CSI. Effort invested in more elaborate combination rules is therefore poorly directed when members are this concordant.

B. Consensus over algorithms versus consensus over information

The contrast with Bondyopadhyay and Mohapatra [2], who reported clear benefit from consensus over thermodynamic indices, is instructive and, we argue, not contradictory. Their consensus operates over heterogeneous information: distinct indices encode different aspects of the atmospheric state and therefore fail in different situations. The consensus evaluated here operates over algorithms trained on one identical twelve-dimensional feature vector. Whatever their inductive biases, all three learners extract signal from the same information, and the ceiling imposed by that information is shared. Algorithmic diversity is thus a weaker guarantee of prediction diversity than is commonly assumed, and the level at which consensus is formed matters more than the number of members.

The practical implication for ensemble design is specific. In problems where a single observing system supplies all predictors, adding learners or refining weights should not be expected to yield appreciable gains. Progress requires enlarging the information base, for example by combining soundings with radar reflectivity, satellite brightness temperature, or numerical weather prediction output, which would introduce members whose failure modes are genuinely distinct.

C. What consensus does provide

The null on peak skill should not be read as an argument against consensus. Its demonstrated value here is in error composition rather than in aggregate accuracy. For thunderstorm, hard voting raises POD to 0.7953 relative to 0.7774 for Random Forest and 0.7507 for the SVM, while holding FAR at 0.5189 against 0.5377 for Gradient Boosting. The majority rule thus filters false alarms originating from the aggressive member while recovering events missed by conservative ones. Equally relevant operationally, it removes the requirement to identify the best single model *a priori*, a choice that is only verifiable in hindsight and that the present results show to be target-dependent, with Random Forest best for rain and hard voting marginally best for thunderstorm.

Consistently sub-0.5 optimal thresholds are a further design finding. At approximately one-third prevalence, the default cut is not a neutral choice but a conservative one that suppresses detection. For hazard warning applications in which the cost of a missed convective event exceeds that of a false alarm, explicit threshold optimisation against a detection-sensitive score is preferable to reporting metrics at 0.5.

D. Limitations

Five limitations bound these conclusions. Soundings are available twice daily, so afternoon convection initiating well after the 00 UTC ascent may be poorly constrained by the available profile. All learners share one feature space, which is the source of the correlation constraint identified here and which the present design can diagnose but not remedy. Training and evaluation are specific to one tropical station over 2016–2025, so transfer to other stations or regimes requires retraining and independent validation. Seasonal behaviour lies outside the present scope: the wet and dry monsoon differ markedly in convective frequency at this station, and the consequences of that difference for threshold selection and skill are treated in a separate study. Finally, evaluation rests on a single point estimate from one held-out period. Consistency of skill across successive periods was not measured and would require rolling-origin evaluation. That consistency is what corresponds to reliability in the sense of continuity of correct service [1], and it is the property that matters for deployment.

With FAR in the range 0.46–0.54, the models are not suitable for autonomous decision-making. Their appropriate role is as a supplementary early-awareness layer within a forecasting workflow that retains human interpretation and other observing systems.

V. CONCLUSIONS

Four consensus schemes over three algorithmically diverse base learners were evaluated for 0–12 h rain and thunderstorm prediction from radiosonde instability indices at a tropical airport, using a chronologically partitioned and leakage-controlled pipeline. Peak CSI reached 0.460 for rain and 0.428 for thunderstorm. Consensus was decisively better than a climatological random predictor ($p < 0.0001$) but not significantly better than the best single model under either McNemar or paired-bootstrap testing.

The explanation is quantitative and, we believe, generalisable. Inter-model prediction correlations of 0.753–0.971 leave little independent error for voting to exploit, and those correlations arise because all learners are trained on an identical feature space. Consensus benefit in this problem is limited by information source rather than by algorithm selection or combination rule.

Precipitable water was identified as the dominant predictor for both targets, reproducing published univariate findings for Indonesia within a multivariate nonlinear framework. For subsequent work, these results indicate that multimodal inputs combining soundings with radar, satellite, and numerical model fields are a more promising direction than further refinement of combination rules, and that rolling-origin evaluation is needed to characterise skill consistency over time.

ACKNOWLEDGEMENTS

This article is developed from the first author's master's thesis at the Republic of Indonesia Defense University, prepared under the supervision of the co-authors. The authors thank the university for institutional support, and the University of Wyoming Upper Air Archive and the Iowa Environmental Mesonet for providing the observational data used in this study.

DATA AVAILABILITY

Radiosonde profiles are available from the University of Wyoming Upper Air Archive and METAR reports from the Iowa Environmental Mesonet. Derived datasets and analysis code are available from the corresponding author on reasonable request.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

[1] A. Avižienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Trans. Dependable Secure Comput.*, vol. 1, pp. 11–33, 2004, doi: 10.1109/TDSC.2004.2.

[2] S. Bondyopadhyay and M. Mohapatra, "Determination of suitable thermodynamic indices and prediction of thunderstorm events for Eastern India," *Meteorol. Atmos. Phys.*, vol. 135, art. 4, 2023, doi: 10.1007/s00703-022-00942-1.

[3] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[5] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, LNCS 1857. Berlin, Germany: Springer, 2000, pp. 1–15, doi: 10.1007/3-540-45014-9_1.

[6] C. A. Doswell III, "Severe convective storms—An overview," in *Severe Convective Storms*, Meteorol. Monogr. 28. Boston, MA, USA: American Meteorological Society, 2001, pp. 1–26, doi: 10.1175/0065-9401-28.50.1.

[7] C. A. Doswell III and D. M. Schultz, "On the use of indices and parameters in forecasting severe storms," *Electron. J. Severe Storms Meteorol.*, vol. 1, pp. 1–22, 2006.

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[9] A. J. Haklander and A. Van Delden, "Thunderstorm predictors and their forecast skill for the Netherlands," *Atmos. Res.*, vol. 67–68, pp. 273–299, 2003, doi: 10.1016/S0169-8095(03)00056-5.

[10] International Civil Aviation Organization, *Annex 3: Meteorological Service for International Air Navigation*. Montreal, QC, Canada: ICAO, 2022.

[11] I. T. Jolliffe and D. B. Stephenson, Eds., *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed. Chichester, U.K.: Wiley-Blackwell, 2012.

[12] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, art. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.

[14] P. Markowski and Y. Richardson, *Mesoscale Meteorology in Midlatitudes*. Chichester, U.K.: Wiley-Blackwell, 2010.

[15] A. McGovern, R. Lagerquist, D. J. Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, "Making the black box more transparent: Understanding the physical implications of machine learning," *Bull. Amer. Meteorol. Soc.*, vol. 100, pp. 2175–2199, 2019, doi: 10.1175/BAMS-D-18-0195.1.

[16] H. P. Nayak and M. Mandal, "Analysis of stability parameters in relation to precipitation associated with pre-monsoon thunderstorms over Kolkata, India," *J. Earth Syst. Sci.*, vol. 123, pp. 689–703, 2014.

[17] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, pp. 195–204, 2019, doi: 10.1038/s41586-019-0912-1.

[18] N. Sagita and T. Takemi, "Exploring the thunderstorm predictors in Indonesia," *SOLA*, vol. 21, pp. 51–60, 2025, doi: 10.2151/sola.2025-007.

[19] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Oxford, U.K.: Academic Press, 2011.

[20] World Meteorological Organization, *Guide to Aeronautical Meteorological Practices*, WMO-No. 732. Geneva, Switzerland: WMO, 2023.

[21] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2012.