Volume 2, 2026, Pages 122-138 *Journal of Smart and Sustainable Farming*
| Regular Article | Open Access

Uncertainty Aware Crop Yield Forecasting Under Extreme Weather Conditions Using Hybrid Temporal Models

Zhang Wang
Department of Tea Science, College of Horticulture, South China Agricultural University, Guangdong, China.
zhangwangliu@mail.hzau.edu.cn

Anandakumar Haldorai
Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, India.
anandakumar.psgtech@gmail.com

Article Info
Journal of Smart and Sustainable Farming
https://www.ansispublications.com/jssf/jssf.html

Received 01 March 2026
Revised from 18 April 2026
Accepted 30 May 2026
Available online 22 June 2026
Published by Ansis Publications.

© The Author(s), 2026.
https://doi.org/10.64026/JSSF/2026011

Corresponding author(s):
Anandakumar Haldorai, Department of Computer Science and Engineering, Sri Eshwar College of Engineering, India.
Email: anandakumar.psgtech@gmail.com

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract – The challenge of forecasting yields is increased and complicated by extreme weather events like heatwaves, droughts, floods and cold snaps, as a forecast that is true on average becomes highly misleading in the seasons when it is needed most by those who need it most: farmers, insurers and planners. A framework for uncertainty-aware forecasting that combines a one-dimensional convolutional front-end, a heteroscedastic Gaussian output, a bidirectional gated recurrent unit and temporal attention, as well as normalizing with split-conformal prediction is presented, followed by presenting five members in a deep ensemble, and calibrating the prediction intervals with normalized split-conformal prediction. The evaluation of the framework is done on a dataset of simulated extreme events over 30 years and 100 regions with a strict temporal split (3000 region-seasons, 24 weekly steps, the number of events increasing with the simulation period). The ensemble outperforms the six point-forecast baselines with a test RMSE of 0.557 t/ha ($R^2 = 0.890$) and error in extreme seasons of 0.587 t/ha (gradient boosting = 0.671 t/ha). The 90% intervals of its 90% intervals cover 92.7% of the seasons, compared to 83–84% for constant-width conformal gradient boosting for seasons corresponding to heatwave, drought and compound seasons. The calibrated ensemble maintains 90.2% coverage under the same type of intensified extremes it didn't see in training, compared to 87.5% for the uncalibrated ensemble and 79.7% for gradient boosting. The gains are found primarily in the heteroscedastic output and the ensemble but not the convolutional, bidirectional or attention components and conclusions are constrained by simulated data.

Keywords – Crop Yield Forecasting, Extreme Weather, Uncertainty Quantification, Deep Ensembles, Conformal Prediction, Climate Adaptation.

I. INTRODUCTION

Crop yield forecasting is essential for decisions on input procurement, timing of irrigation, crop storage policies, crop insurance and food-security planning and for agricultural trade. Global warming has been affecting the production of major crops since the early 1980s [1] and extreme weather events or droughts and extreme heat have been causing losses at the national level [2]. Crop production is known to be nonlinear in response to temperature with potential for damage to escalate rapidly above critical temperature thresholds [3, 4]. With increasing frequency of heatwaves, long periods of dry weather and heavy rainfall events, forecasting is becoming more critical. The years with the fewest observations, however, are the most important in agricultural terms: it is especially difficult to predict in extreme years.

Agricultural yield forecasting is an effective method using machine learning (ML) techniques [5]. Recurrent and convolutional neural networks are able to model the temporal dependencies and weather influence at various crop growth stages [6]. These capabilities can be extended with probabilistic forecasting methods that estimate the uncertainty of

prediction [7]. However, most current work focuses on point forecast assessments based on an overall error measure. These assessments can mask significant deficiencies at critical times of the year, when errors in prediction can be larger and their impacts more damaging. In addition, without good estimates of uncertainty point forecasts are of little value for risk sensitive decisions. For example, when it is predicted that 5 tonnes per hectare would be produced, it would be helpful to have a reliable range of production it could be within that range.

There are three interwoven challenges brought about by extreme weather. The first is that yield variability is heteroscedastic, as it is more difficult to predict the yield under stress and the size of the prediction interval may not be consistent between different seasons. Second, the impact of an extreme event on agriculture is affected by its timing, and heat stress during flowering can result in higher yield losses than stress during early vegetation. Third, forecasting models can be affected by conditions more extreme than those encountered in the data used for model training, which can reduce the reliability of forecasting.

To solve these problems, this study introduces an uncertainty-aware temporal forecasting framework based on convolutional feature extractor, bidirectional gated recurrent unit (BiGRU), temporal attention mechanism and heteroscedastic Gaussian output. The architecture not only estimates the expected yield, but also the variance of the yield that depends on the inputs, which allows the uncertainty to adjust to the seasonal conditions. A deep ensemble is a collection of independently trained models used to quantify the uncertainty of the model and observation. Thereafter, normalized split-conformal calibration is a technique which modifies the prediction intervals to better meet the calibration requirements empirically.

There are four major contributions to the principal. First, the proposed hybrid architecture combines spatially structured feature extraction, bidirectional temporal learning and attention-based weighting. Second, ensemble-based uncertainty estimation and conformal calibration offers a single, coherent probabilistic forecasting procedure. Third, it uses a strict separation of time in the evaluation protocol and evaluates the predictive errors and interval coverage in normal, single-event, and compound-extreme seasons, including the intensified seasons not included in the training. 4th, the analysis of ablation and computation costs examines the contribution, the limitations and the practical trade-offs of the proposed framework, on a component-wise basis.

This paper is organized as follows: In Section II, the literature on the related work is reviewed. The forecasting problem is formulated in Section III and the description of the dataset is based on simulation in Section IV. The proposed methodology is given in Section V, the experimental setup in Section VI and the results in Section VII. Section VIII provides an overview of the key findings and Section IX concludes the study.

II. RELATED WORKS

Weather Extremes and Crop Yield

Timing and the extent of unfavourable conditions, as well as temperature, precipitation and soil-water availability, both strongly affect yield of crops. Geographical statistical analyses show that temperature and agricultural productivity do not follow a linear relationship, suggesting that beyond the crop-specific thermal thresholds, losses in agricultural productivity can increase rapidly [3]. Empirical models can work well in determining past associations, but can fall short when the environmental conditions are beyond the range of the data distribution that has been observed. Process-based crop models overcome this drawback by including physiological development and soil-water balance, as well as the environmental stress mechanisms. Crop water requirement and drought impact estimation are still significant, using the reference evapotranspiration or water-balance method [8] [9]. Multi-model assessments also confirm that increases in temperature have detrimental effects on yields of some of the major crops, but that the impact differs by crop type and region and depends on climatic conditions [4]. Still, there are difficulties in predicting extreme weather due to the rarity of the events and the impact requirements of extreme weather in the growth phases. These constraints encourage models that are able to learn a nonlinear relationship and that explicitly consider temporal variation and uncertainty for extreme seasons.

Machine Learning for Yield Forecasting

The ability of machine learning to predict crop yields using complex relationships between meteorological variables and agricultural outcomes has taken crop yield prediction beyond traditional statistical regression. Random forests are made by using a bunch of decision trees that make good predictions, and gradient boosting sequentially builds up good predictions by correcting errors of previous learners [10, 11]. This is the case for sequential neural networks, which observe sequentially. LSTMs have gated memory mechanisms to retain long-range dependencies [12] while GRUs offer a more compact temporal representation architecture [13]. Bidirectional recurrent networks are designed to include information from both directions in a finished seasonal sequence, enabling the identification of relationships between weather conditions in the different growth stages [14]. Attention mechanisms also enhance the temporal representation, giving different weights to observations based on their predictive relevance [15]. Hybrid CNN–RNN architectures have the potential of sequential learning and local feature extraction, and have been shown to be effective in agricultural yield prediction [6]. This general trend has been observed in the wider literature on agricultural forecasting and a growing interest in deep learning [5]. But just because a model is accurate does not mean it is reliable in extreme weather, especially when it is given a combination of weather conditions that it has never seen before.

Predictive Uncertainty Quantification

Along with agricultural forecasts, estimates of uncertainty are needed for reliable forecasting. Unlike least-squares regression, quantile regression directly estimates conditional quantiles and it is possible to obtain asymmetric prediction intervals without assuming a particular distribution for the outcome [16]. Monte Carlo dropout approximates the Bayesian inference by keeping dropout in the inference phase and calculating the uncertainty in repeated predictions [17]. Heteroscedastic neural networks on the other hand estimate an input-dependent predictive mean and variance, often optimized using the Gaussian negative log-likelihood [18]. This method can accommodate varying yield variability from normal and stressed seasons. In many applications, deep ensembles have been developed to improve uncertainty estimation and capture differences in the predictions of independently trained networks [19].

The complementary calibration mechanism is based on conformal prediction, which builds finite-sample prediction intervals with margin of coverage of the model under exchangeability [20]. Conformal methods are more general than distribution-specific methods of uncertainty estimation and can be used with a wide variety of predictive models. However, it is possible for a neural network to be well calibrated despite being accurate [23]. Moreover, uncertainty estimates and predictive reliability may be affected by changes between the training and deployment data sets if there are differing deployment situations [24]. Therefore, calibration needs to be considered in the context of predictive error and in conditions that are reflective of the practical risks in agriculture.

Research Gap and Motivation

The research attempted in the past include temporal learning, nonlinear prediction, uncertainty estimation, and interval calibration; but these abilities are not necessarily combined into a single crop yield forecasting system. Aggregate evaluation metrics can mask unreliable predictions during the extremes of a season, and traditional calibration can fail to give information on performance in more extreme weather [21]. The proposed work is a fusion of convolutional feature extraction, bidirectional GRU learning, temporal attention, heteroscedastic Gaussian prediction, deep ensembling, and normalized split-conformal calibration. It is evaluated to separate normal from single-event and compound-extreme seasons, and examines the extremes that are intensified that are not part of the training. Component-wise ablation and computational cost analysis further confirm the reliability of the prediction and efficiency of the computation of the proposed framework, respectively [22].

III. PROBLEM FORMULATION

Let a season i be described by a weather sequence $\mathbf{X}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,T})$ with $T = 24$ weekly steps and $d = 5$ channels (maximum and minimum temperature, rainfall, solar radiation and soil moisture), a static vector $\mathbf{s}_i$ of four region attributes, and a yield y_i in tonnes per hectare. The goal is to learn a predictive distribution for y_i given $(\mathbf{X}_i, \mathbf{s}_i)$ from which two outputs are derived: a point forecast $\hat{y}_i$ that minimizes squared error, and a prediction interval $[l_i, u_i]$ that contains the true yield with probability at least $1 - \alpha$. The interval should be narrow when the season is predictable and wide when it is not, and it should keep its coverage in extreme seasons and under changing climate.

IV. DATA AND SIMULATION FRAMEWORK

Overview

A simulated data set of 3,000 regions-seasons (100 regions observed over 30 years) is used in the study. Each season has 24 weekly time steps. The selection of a simulation was made because real yield records with weekly weather and soil moisture data and labelled extreme events were not available for this study and simulation allows for complete control of the frequency, timing and intensity of extreme events. These data are therefore not representative of actual farms, but rather a controlled test field and the implications are discussed in Section VII-M. The variables are given in **Table 1**.

Table 1. Variables of the Simulated Dataset

Variable	Type	Unit / range
Maximum temperature (Tmax)	Weekly input	°C; weekly mean of daily maxima
Minimum temperature (Tmin)	Weekly input	°C
Rainfall	Weekly input	mm per week; 0 to above 300
Solar radiation	Weekly input	MJ m^{-2} day^{-1}; 6 to 28
Soil moisture	Weekly input	fraction of water-holding capacity, 0 to 1
Soil water capacity	Static	80 to 200 mm
Soil fertility index	Static	0.70 to 1.10
Irrigation share	Static	0 to 0.80
Climate offset	Static	°C relative to the mean region (std. 2.2)
Yield (target)	Season output	t/ha; mean 5.67, std. 1.62, range 0.30 to 9.56

Weather Generator and Extreme Events

The weekly weather cycles are seasonally modulated and summatively offset by a region, with a linear warming trend of 0.04 °C per year, and with first-order autoregressive anomalies. The probability of rainfall in a particular week is 0.65 and the amount of rainfall is gamma-distributed, scaled by an aridity factor that is specific to the region. Soil moisture is calculated using a simple bucket model and weekly evapotranspiration is estimated using a temperature and radiation-based model similar to the Hargreaves equation [8] and a crop coefficient that reaches a maximum at mid-season [9], with irrigation refilling part of the deficit. These four base classes of extreme events are added on top of this baseline. Each event class is made more likely to occur over the 30-year period, increasing by 2% of the base rate each year. The proportion of seasons with at least one event (35.3% of seasons 1-6 and 50.3% of seasons 25-30) and two or more events (compound seasons) (7.9% of seasons 1-6 and 25-30) increases over all seasons. **Table 2** shows extreme-event generator

Table 2. Extreme-Event Generator

Event	Perturbation	Duration	Base prob.
Heatwave	Tmax +4 to +9 °C (Tmin +60% of that)	1–3 weeks, onset weeks 7–21	0.14
Drought	Rain × 0.05–0.30; Tmax +1 to +3 °C	4–8 weeks, onset weeks 4–17	0.12
Flood	+120 to +250 mm rain per week; less radiation	1–2 weeks, onset weeks 4–21	0.08
Cold snap	Tmin −5 to −9 °C	1 week, early (1–6) or late (21–24)	0.05

Hidden Crop Process and Observed Inputs

Yield is produced by a hidden, crop-growth-style process that the models never see. A potential yield depends on soil fertility, mean radiation and a thermal-suitability factor. It is multiplied by a stress factor $exp(-S)$, where S adds four weekly stress terms: heat stress above 34 °C, water stress when soil moisture falls below 45% of capacity, waterlogging when weekly rain exceeds 100 mm, and cold stress below 3 °C. Each term is weighted by a growth-stage sensitivity that peaks around flowering for heat and water stress and early in the season for waterlogging and cold. An interaction term couples heat stress and water stress, so compound events are worse than the sum of their parts. Observation noise is multiplicative (4%) plus an additive noise whose standard deviation grows with stress, from about 0.2 t/ha in good seasons to about 0.8 t/ha in the most stressed ones. In 3% of seasons an unobserved shock such as pests or hail removes 10–30% of the yield. The inputs given to the models carry measurement noise (0.3 °C for temperature and radiation, 5% for rainfall, 0.01 for soil moisture).

The resulting data behave as intended. Extreme seasons yield on average 5.17 t/ha against 6.02 t/ha for normal seasons (14% lower), and mean yield by event class is 5.03 t/ha for heatwaves, 4.71 for droughts, 5.31 for floods and 5.75 for cold snaps. The mean yield falls from 5.95 t/ha in years 1–6 to 5.40 t/ha in years 25–30 as extremes become more frequent.

Temporal Split and Pre-Processing

To mimic real forecasting, seasons are split by year, never at random: years 1–18 for training (1800 seasons), 19–20 for validation and early stopping (200), 21–24 for conformal calibration (400) and 25–30 for testing (600). Because extremes become more frequent over time, the test period contains more extreme seasons (50.3%) than the training period, which is a realistic mild distribution shift. A separate shifted scenario of 600 seasons was generated for years 25–30 with a different random seed, event intensities multiplied by 1.5 and event probabilities multiplied by 1.5; here 63.3% of seasons contain events and the mean yield is 4.96 t/ha. Every channel is standardized with training statistics:

$$\widetilde{x_{t,c}} = \frac{x_{t,c} - \mu_c}{\sigma_c} \quad (1)$$

Static features and the target are standardized in the same way, and predictions are transformed back to t/ha before any metric is computed.

V. PROPOSED METHODOLOGY

Architecture Overview

Fig. 1 shows the framework. Each ensemble member reads the sequence of weather signals through a standardized front-end that involves a convolutional layer and a bidirectional GRU layer, then generates a summary for the season by using temporal attention, embeds the static features and generates a mean and a variance. The intervals for the deep ensemble are calibrated using held-out seasons, via conformal prediction, from 5 members trained from different random initializations.

Convolutional Front-End

A one-dimensional convolution with kernel size three and 32 filters extract short local patterns, such as a sudden rise in temperature over three consecutive weeks, before the recurrent layer reads the sequence:

$$f_t = \mathrm{ReLU}\left(\sum_{j=1}^{1} bW_j \widetilde{x_{t+j}} + b\right) \quad (2)$$

where $\mathbf{W}_j$ are the filter weights for offset *j*, **b** is a bias, and zero padding keeps the length at *T*.

Bidirectional GRU

The GRU [13] summarizes the sequence with gated recurrence. With update gate $\mathbf{z}_t$ and reset gate $\mathbf{r}_t$, the hidden state evolves as

$$z_t = \sigma\left(W_{zf_t} + U_{zh_{t-1}} + b_z\right), r_t = \sigma\left(W_{rf_t} + U_{rh_{t-1}} + b_r\right) \quad (3)$$

$$\widetilde{h_t} = \tanh\left(W_{hf_t} + r_t \odot U_{hh_{t-1}}\right), h_t = (1 - z_t) \odot \widetilde{h_t} + z_t \odot h_{t-1} \quad (4)$$

where σ is the logistic function and $\odot$ is element-wise multiplication. A second GRU reads the sequence in reverse, and the two hidden states are concatenated at every week, which is allowed because the whole season is observed before the yield is forecast [14].

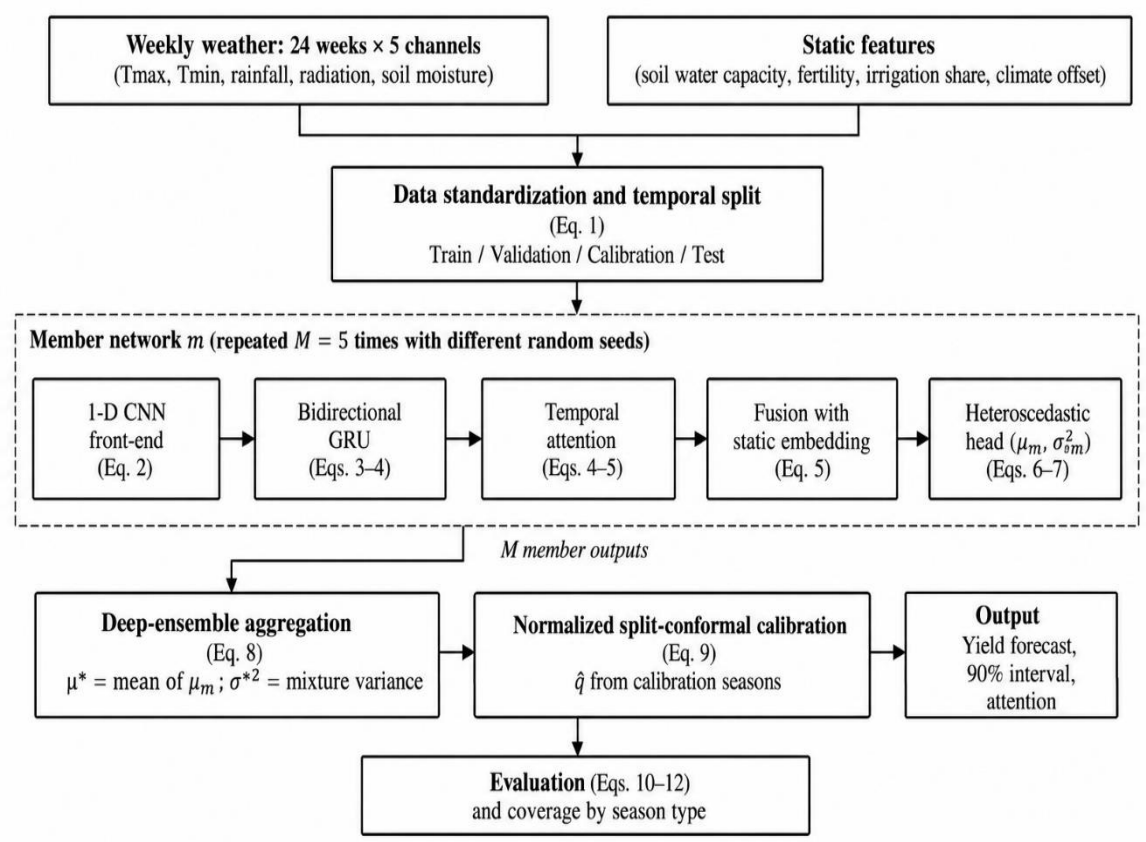


Fig 1. Architecture of the Proposed Uncertainty-Aware Hybrid Temporal Framework.

Temporal Attention and Fusion

Attention [15] lets the model weight weeks differently, for example giving more weight to the flowering period. The bidirectional states H_t are scored and normalized over time:

$$H_t = \left[\vec{h}_t ; \overleftarrow{h}_t\right] \quad (5)$$

$$e_t = v^T \tanh\left(W_{aH_t} + b_a\right), \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (6)$$

and the season context is the attention-weighted sum, which is fused with an embedding of the static vector **s** by a dense layer:

$$c = \sum_{t=1}^{T} \alpha_{tH_t}, \quad u = \mathrm{ReLU}(W_f[c; \mathrm{ReLU}(W_s s + b_s)] + b_f) \quad (7)$$

where $[\cdot;\cdot]$ denotes concatenation. The attention weights α_t are also returned to the user, so that the weeks that most influenced a forecast can be inspected.

Heteroscedastic Output and Training Loss
The output layer predicts both a mean and a variance, the latter through a log-variance parameterization that keeps it positive [18]:

$$\mu(x) = w_{\mu}^{Tu} + b_{\mu}, \quad \sigma^2(x) = \exp(w_{\sigma}^{Tu} + b_{\sigma}) \tag{8}$$

In the implementation the log-variance is softly bounded to the range $[-4, 4]$ in standardized units, which prevents the variance from collapsing or exploding early in training. The member is trained with the Gaussian negative log-likelihood and decoupled weight decay:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left[\log\sigma^2(x_i) + \frac{(y_i \mathrm{mu}(x_i))^2}{\sigma^2(x_i)}\right] + \lambda\|\theta\|_2^2 \tag{9}$$

where θ are the network weights and λ the weight decay. The first term rewards accurate means, and the second makes the model state larger variances in seasons where its errors are larger, so that the variance can act as a season-specific uncertainty estimate.

Algorithm and Flowchart
Algorithm 1 gives the training and forecasting procedure, and **Fig. 2** shows the corresponding flowchart.

Algorithm 1: Uncertainty-aware hybrid temporal forecasting
Input: weather sequences X, static features s, yields y; ensemble size M = 5; miscoverage α = 0.1
Output: forecast μ *and interval* $[\mu - \hat{q}\sigma, \mu + \hat{q}\sigma^*]$ for new seasons
1: Standardize inputs and target using training statistics (Eq. 1)
2: Split seasons by year into train, validation, calibration and test sets
3: for m = 1 to M do
4: Initialize member m with random seed m
5: repeat (mini-batches of 64)
6: Compute f_t (Eq. 2), h_t and H_t (Eqs. 3–4), c and u (Eqs. 4–5)
7: Compute μ_m and σ_m^2 (Eq. 6); minimize the loss in Eq. (7) with AdamW
8: until validation NLL has not improved for 10 epochs; restore best weights
9: end for
10: Compute μ *and* σ^2 on calibration and test seasons by Eq. (8)
11: Compute scores s_i on calibration seasons and the quantile $\hat{q}$ (Eq. 9)
12: for each new season do
13: Standardize inputs; obtain (μ_m, σ_m^2) from all members; compute μ, σ
14: Return μ *and the interval* $\mu \pm \hat{q}\sigma^*$, together with attention weights α_t
15: end for

Deep Ensemble
A single network's variance captures the noise it has learned but not its own uncertainty about the weights. A deep ensemble [19] of $M = 5$ members trained from different seeds treats the members as a uniformly weighted mixture of Gaussians, whose mean and variance are

$$\mu^* = \frac{1}{M}\sum_{m=1}^{M}\mu_m, \quad \sigma^{*2} = \frac{1}{M}\sum_{m=1}^{M}(\sigma_m^2 + \mu_m^2) - \mu^{*2} \tag{10}$$

The variance contains the average predicted noise and the disagreement among members. Averaging also reduces the variance of the point forecast.

Normalized Split-Conformal Calibration
The ensemble's Gaussian intervals can be too narrow or too wide, particularly when the test period differs from the training period. Split-conformal prediction [20], [21], [22] uses held-out calibration seasons to correct this. Each calibration season receives a normalized nonconformity score; the absolute error measured in units of the predicted standard deviation. The $(1 - \alpha)$ quantile of these scores is used to scale all future intervals:

$$s_i = \frac{|y_i - \mu_i^*|}{\sigma_i^*}, \quad \hat{q} = s_{(\lceil (n+1)(1-\alpha) \rceil)} \tag{11}$$

$$\mathcal{J}_\alpha(x) = [\mu^* - \hat{q}\sigma^*, \mu^* + \hat{q}\sigma^*] \tag{12}$$

where $s_{(k)}$ is the k-th smallest of the n calibration scores and $\alpha = 0.1$. Normalizing by σ^* keeps the intervals adaptive: seasons with a large predicted variance receive wider intervals. If the calibration and test seasons are exchangeable, the marginal coverage is at least $1 - \alpha$ [21]. In this study the test seasons come from later years, so exchangeability holds only approximately, and coverage is evaluated rather than assumed.

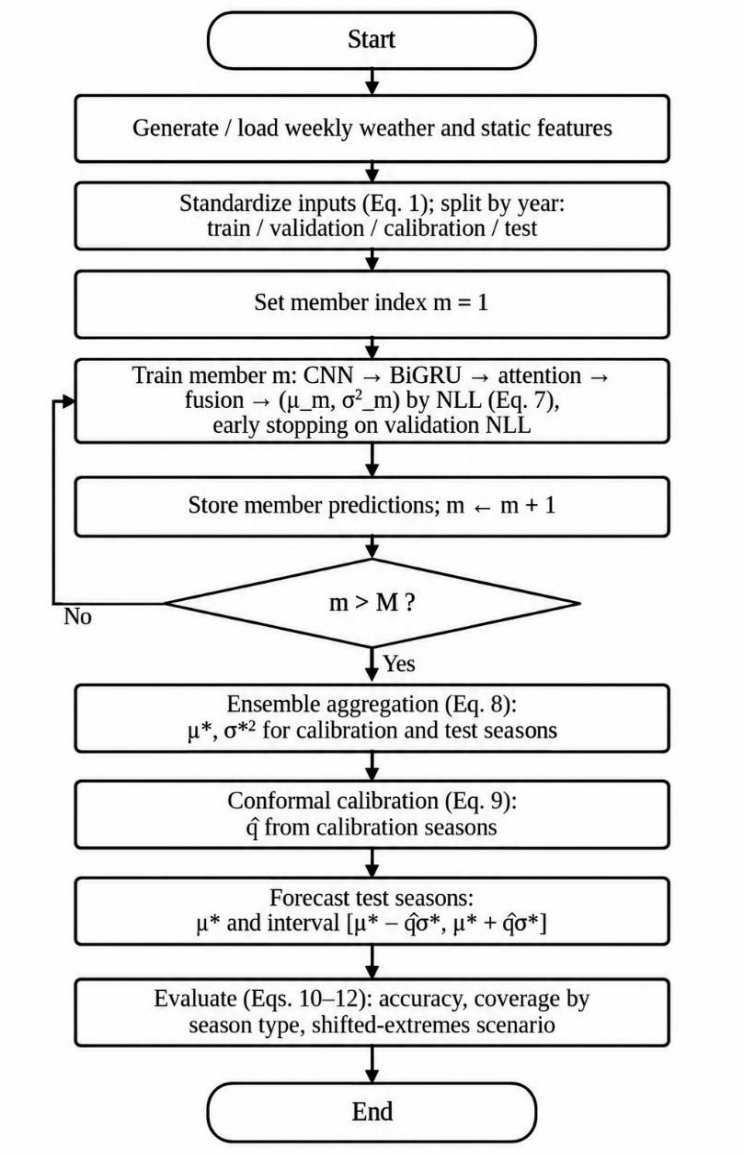


Fig 2. Flowchart of the Proposed Forecasting Framework.

Computational Complexity

For one member, a forward pass costs $O(T(dh + h^2))$ for the recurrent layers per season, where h is the hidden size, and the attention adds $O(Th)$. Training the ensemble costs about M times the training of one member, and prediction costs M forward passes. By contrast, Monte Carlo dropout needs 30 passes per prediction in the setting used here, so that a five-member ensemble is cheaper at prediction time. Conformal calibration adds only a sort of n scores.

VI. EXPERIMENTAL SETUP

Implementation and Baselines

No deep-learning framework was available in the computing environment, so the networks were implemented from scratch in NumPy with a small reverse-mode automatic-differentiation engine; the gradients of all architectures were verified

against finite differences (maximum relative error 3.5×10^{-7}). Optimization uses Adam [25] with decoupled weight decay [26]. The classical baselines use scikit-learn [27]. Six point-forecast baselines are compared: ridge regression, random forest [10] and gradient boosting [11] on the flattened season (120 weather values plus four static values), and LSTM [12], GRU [13] and CNN-LSTM networks. Uncertainty baselines are split-conformal gradient boosting with constant interval width, quantile gradient boosting at the 5% and 95% quantiles [16], Monte Carlo dropout [17] on the same architecture trained with squared error, a single heteroscedastic member, and the raw ensemble without conformal calibration. Deep models were trained with three random seeds (eight for the proposed member), and the mean and standard deviation over seeds are reported. **Table 3** lists the main settings.

Table 3. Main Settings of the Models

Component	Setting
Proposed member	Conv1D (k = 3, 32 filters), BiGRU (32 per direction), attention (32), static embedding 16, fusion 32
Training	AdamW, lr 3×10^{-3}, weight decay 10^{-4}, batch 64, max. 60 epochs, patience 10, gradient clipping 5
Ensemble / conformal	M = 5 members (seeds 0–4); α = 0.1; calibration on years 21–24 (400 seasons)
LSTM, GRU, CNN-LSTM	32 hidden units, same optimizer; mean-squared-error loss
MC-dropout	Proposed architecture, dropout 0.2, squared-error loss, 30 stochastic passes
Random forest	200 trees; ridge regression: penalty 10
Gradient boosting	Histogram-based, 100 iterations, learning rate 0.1; quantile versions at 0.05, 0.50, 0.95

Metrics

Point forecasts are evaluated by root-mean-square error (RMSE), mean absolute error (MAE), the coefficient of determination (R^2) and the symmetric mean absolute percentage error (sMAPE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2}\,, \quad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|\,, \quad R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - haty_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{13}$$

Prediction intervals at the nominal 90% level are evaluated by the prediction-interval coverage probability (PICP) and the mean prediction-interval width (MPIW),

$$\mathrm{PICP} = \frac{1}{N}\sum_{i=1}^{N} oI[l_i \le y_i \le u_i]\,, \quad \mathrm{MPIW} = \frac{1}{N}\sum_{i=1}^{N}(u_i - l_i) \tag{14}$$

where l_i and u_i are the interval limits, and by the Winkler interval score, which penalizes both width and misses [28]:

$$IS_\alpha = (u - l) + \frac{2}{\alpha}(l - y)I[y < l] + \frac{2}{\alpha}(y - u)I[y > u] \tag{15}$$

with $\alpha = 0.1$. For models with a predictive standard deviation, the Gaussian negative log-likelihood (NLL, the data term of Eq. 7 plus a constant) and the continuous ranked probability score (CRPS, in its closed form for a Gaussian) [28] are also reported. Because averages can hide failures in extremes, all measures are also computed separately for six mutually exclusive season types: normal, heatwave only, drought only, flood only, cold snap only, and compound (two or more events).

Protocol

All models are trained on the 1,800 training seasons, early stopping uses the 200 validation seasons, conformal quantiles use the 400 calibration seasons, and every reported number is computed on the 600 test seasons or, where stated, on the 600 shifted seasons. The test seasons were not used for any choice of hyper-parameter or architecture; all settings were fixed before the final runs.

VII. RESULTS AND DISCUSSION

Dataset Overview

Fig. 3 summarizes the simulated data. Extreme seasons shift the yield distribution towards lower values and widen it (panel a). The share of extreme seasons rises across the 30 years (panel b), so the test period is more extreme than the training period. Panel (c) shows a season with a heatwave and a drought around flowering, the period to which the crop is most sensitive.

Point Forecast Accuracy

Table 4 compares the point forecasts. The proposed ensemble has the lowest error: RMSE 0.557 t/ha and $R^2 = 0.890$, which is 9.2% lower RMSE than gradient boosting (0.614) and 5.8% lower than the GRU (0.591). Ridge regression and random forest are clearly weaker (RMSE 0.782 and 0.728), because they cannot easily represent the stage-dependent, interacting stresses. A single proposed member (0.582 ± 0.017) is not distinguishable from the LSTM, GRU and CNN-LSTM, whose results overlap within one standard deviation over seeds. The improvement of the proposed approach therefore comes from the ensemble and not from the single-member architecture. MC-dropout (0.571) is a competitive and cheaper-to-train alternative to a single member, which suggests that averaging stochastic passes also helps accuracy. **Fig. 4** shows predicted versus actual yield: gradient boosting scatters more widely at low yields, which are mostly extreme seasons.

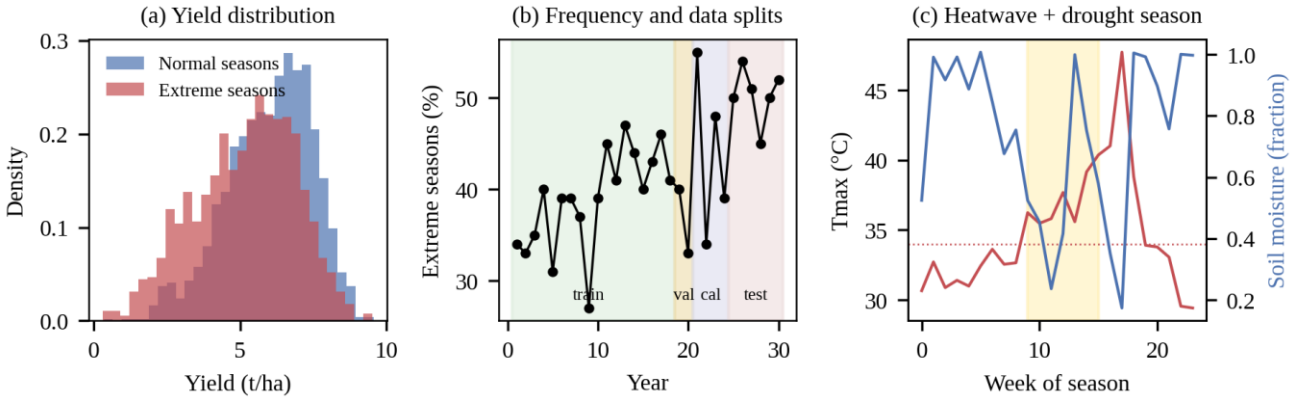


Fig 3. Overview of the Simulated Data: (A) Yield by Season Type, (B) Frequency of Extreme Seasons and Data Splits, (C) An Example Compound Season.

Table 4. Point Forecast Accuracy on the Test Seasons (T/Ha)

Model	RMSE	MAE	R^2	sMAPE (%)
Ridge regression	0.782	0.634	0.783	13.73
Random forest	0.728	0.566	0.812	12.68
Gradient boosting	0.614	0.473	0.866	10.60
LSTM	0.589 ± 0.020	0.456	0.877	10.28
GRU	0.591 ± 0.003	0.455	0.876	10.26
CNN-LSTM	0.599 ± 0.026	0.468	0.872	10.43
MC-dropout (30 passes)	0.571	0.440	0.884	9.88
Proposed, single member	0.582 ± 0.017	0.455	0.879	10.16
Proposed ensemble (M = 5)	**0.557**	**0.432**	**0.890**	**9.64**

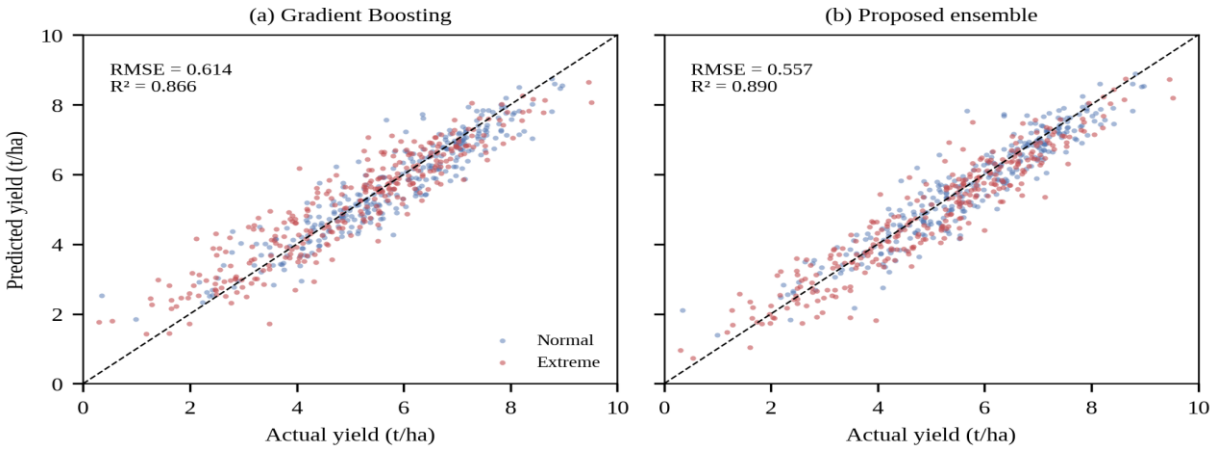


Fig 4. Predicted Versus Actual Yield on the Test Seasons: (A) Gradient Boosting, (B) Proposed Ensemble.

Accuracy in Extreme Seasons

Table 5 and **Fig. 5** break the error down by season type. All models are most accurate in normal seasons and least accurate in compound seasons. For the proposed ensemble the RMSE rises from 0.525 t/ha in normal seasons to 0.599 t/ha in

compound seasons, a factor of 1.1, against a factor of 1.3 for gradient boosting (0.549 to 0.721). Over all extreme seasons the ensemble reaches 0.587 t/ha against 0.671 for gradient boosting and 0.646 for the GRU, a reduction of 12.5% and 9.1% respectively. The advantage of temporal models is therefore larger in extreme seasons (12.5%) than in normal seasons (4.5%), in line with the expectation that timing and interaction effects matter most there. The cold-snap column rests on only 24 seasons and is noisy.

Table 5. Test RMSE (T/Ha) By Season Type

Model	Normal (298)	Heat (107)	Drought (65)	Flood (46)	Cold (24)	Compound (60)	All extreme
Random forest	0.619	0.801	0.877	0.625	0.643	0.978	0.822
Gradient boosting	0.549	0.712	0.647	0.536	0.652	0.721	0.671
LSTM	0.540	0.660	0.603	0.590	0.600	0.658	0.633
GRU	0.530	0.663	0.636	0.540	0.626	0.705	0.646
CNN-LSTM	0.559	0.628	0.649	0.564	0.629	0.685	0.636
MC-dropout	0.527	0.646	0.631	0.517	0.653	0.577	0.612
Proposed ensemble	**0.525**	**0.596**	**0.583**	**0.527**	**0.639**	**0.599**	**0.587**

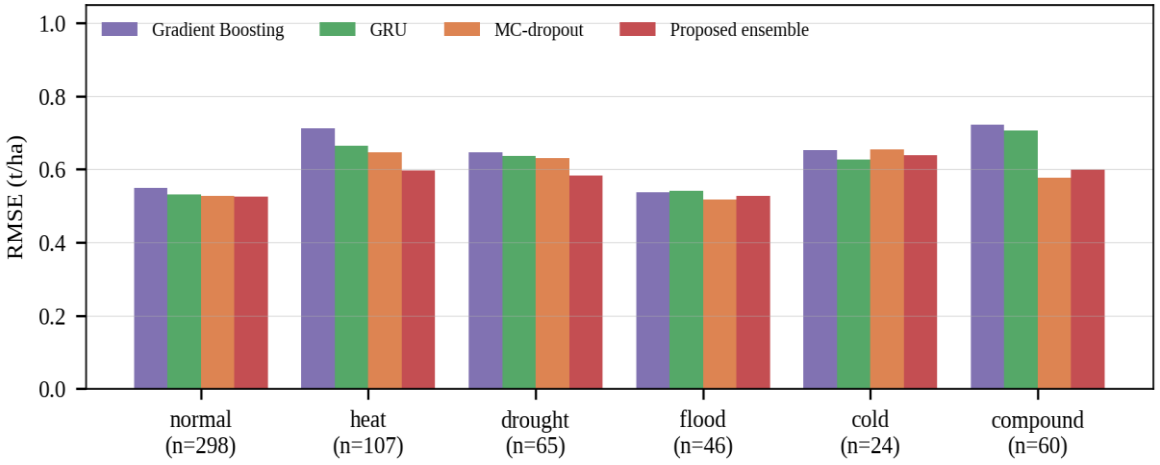


Fig 5. Test RMSE by Season Type for Four Models.

Quality of the Prediction Intervals

Table 6 evaluates 90% intervals. Quantile gradient boosting is badly calibrated: its nominal 90% interval covers only 72.8% of the test seasons, because quantile models fitted to past years under-estimate the wider spread of later, more extreme years. Split-conformal gradient boosting reaches the right marginal coverage (89.5%) but with a constant width of 2.00 t/ha. MC-dropout is close to nominal (90.0%), and a single heteroscedastic member under-cover (87.0%), because it ignores its own parameter uncertainty. The deep ensemble improves this to 92.0% with the lowest interval score (2.347), NLL (0.825) and CRPS (0.308) of all methods. The proposed ensemble with conformal calibration has coverage 92.7% and intervals 6.8% wider than the raw ensemble (1.89 against 1.77 t/ha), with an interval score of 2.373. On this test set conformal calibration therefore over-covers slightly and is marginally worse on the interval score than the raw ensemble; its benefit shows in Sections VII-F and VII-J. **Fig. 6** contrasts the adaptive intervals of the proposed method with the constant-width intervals of conformal gradient boosting: the latter miss many low-yield seasons, which are mostly extreme.

Table 6. Quality of Nominal 90% Prediction Intervals on the Test Seasons

Method	PICP (%)	MPIW	Interval score	NLL	CRPS
Quantile gradient boosting	72.8	1.89	3.886	—	—
Split-conformal gradient boosting	89.5	2.00	2.674	—	—
MC-dropout	90.0	1.85	2.434	0.849	0.316
Single heteroscedastic member	87.0	1.68	2.475	0.889	0.324
Deep ensemble (raw)	92.0	1.77	2.347	0.825	0.308
Proposed (ensemble + conformal)	**92.7**	**1.89**	**2.373**	**0.825**	**0.308**

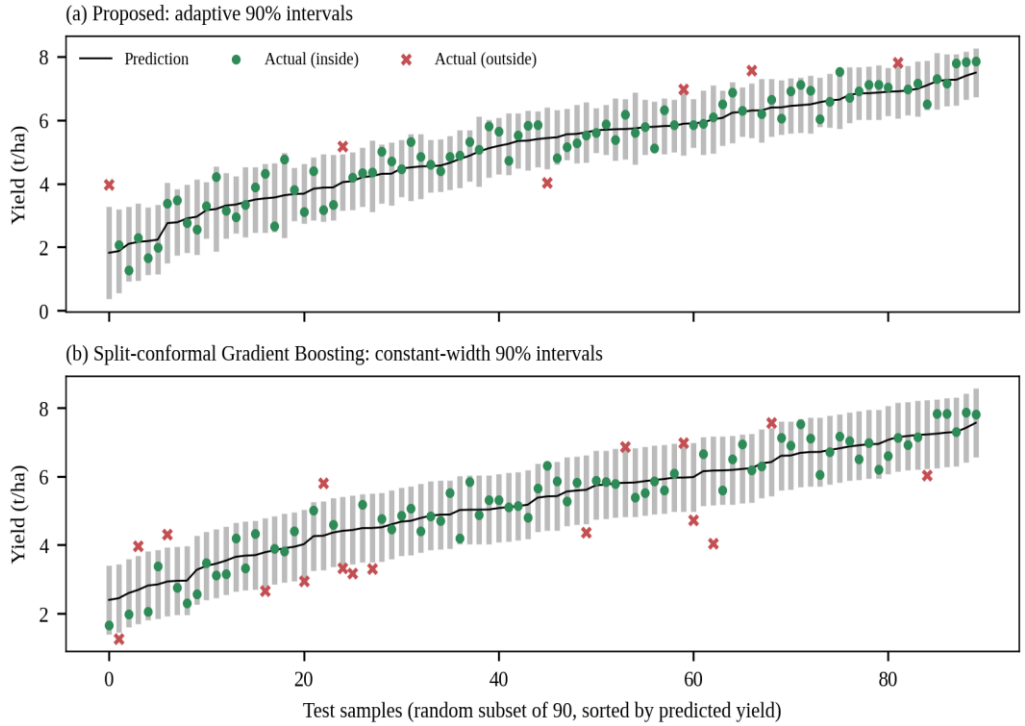


Fig 6. Ninety-Percent Prediction Intervals for 90 Random Test Seasons: (A) Proposed Method, (B) Split-Conformal Gradient Boosting.

Calibration Across Confidence Levels

Fig. 7 plots empirical against nominal coverage for confidence levels from 50% to 95%. All methods lie close to the diagonal at 90% and 95%, but at lower levels the ensemble-based intervals under-cover (for example about 58% empirical at the 60% level), while conformal gradient boosting over-covers at the lower levels because its constant width is dominated by a few large errors. The higher-level calibration, which matters for risk decisions, is good for the proposed method, but the Gaussian shape of the predictive distribution is imperfect in the centre, which is consistent with the heavy-tailed unobserved shocks present in the data.

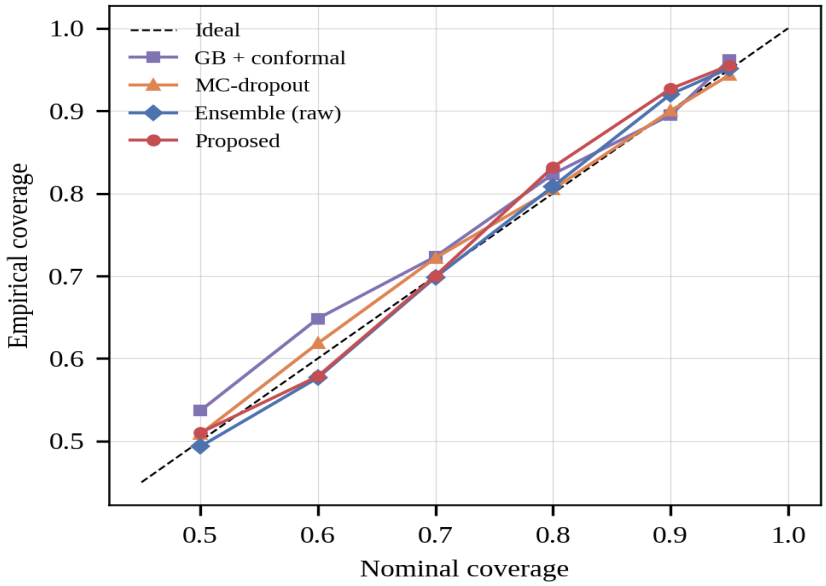


Fig 7. Calibration Curves: Empirical Versus Nominal Coverage.

Conditional Coverage by Season Type

Marginal coverage can mask failures in the seasons that matter most. **Table 7** and **Fig. 8** show coverage by season type. Constant-width conformal gradient boosting covers 93.3% of normal seasons but only 84.1% of heatwave, 83.1% of drought and 83.3% of compound seasons, so it is over-confident exactly when the stakes are high. MC-dropout shows the

same pattern in milder form (84.1% for heatwaves). The heteroscedastic ensemble keeps coverage within 92.3–93.5% across all season types with at least 46 seasons, and its mean interval width is larger for extreme seasons: for the proposed method the average width is 1.79 t/ha for normal seasons and 2.08 t/ha for compound seasons. Cold-snap coverage (87.5%) rests on only 24 seasons and cannot be distinguished from 90% (a deviation of one season changes it by four points).

Table 7. Empirical Coverage (%) of Nominal 90% Intervals by Season Type

Method	Normal (298)	Heat (107)	Drought (65)	Flood (46)	Cold (24)	Compound (60)
GB + conformal	93.3	84.1	83.1	95.7	87.5	83.3
MC-dropout	93.3	84.1	86.2	91.3	87.5	88.3
Deep ensemble (raw)	92.3	91.6	92.3	93.5	87.5	91.7
Proposed	**93.0**	**92.5**	**92.3**	**93.5**	**87.5**	**93.3**

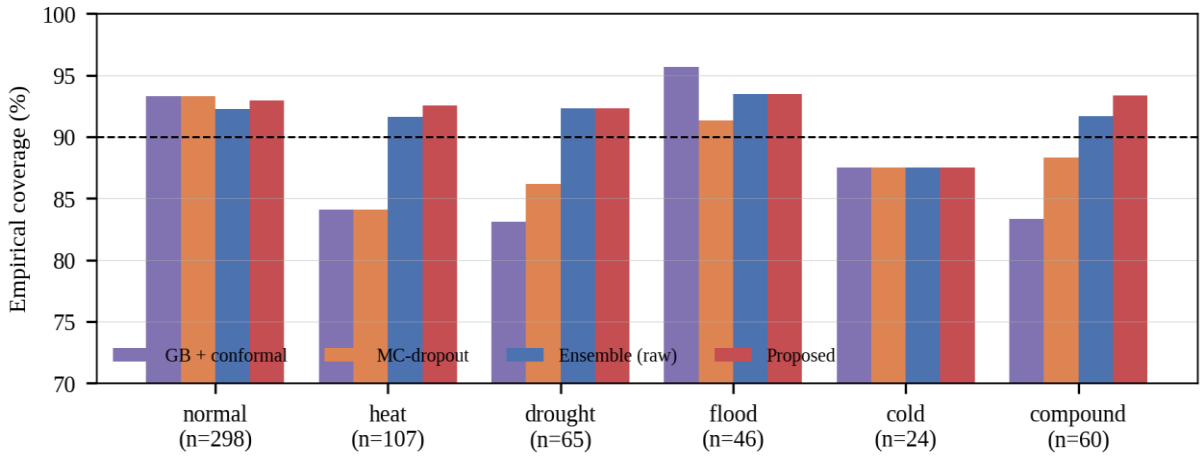


Fig 8. Empirical Coverage of 90% Intervals by Season Type (Dashed Line: Nominal Level).

Fig. 9 shows the mean attention weight by week for each season type, averaged over the ensemble members. The attention is concentrated between weeks 11 and 14, which includes the flowering period to which the simulated crop is most sensitive, for all season types, and the curves differ only slightly between season types (peak weeks 13, 13, 14, 13, 12, 11 for normal, heatwave, drought, flood, cold and compound seasons). The attention weights therefore mainly reflect the stage sensitivity of the crop and not the timing of individual events. This limits their value as an explanation of single events, and, as the ablation in Section VII-H shows, they do not by themselves improve accuracy.

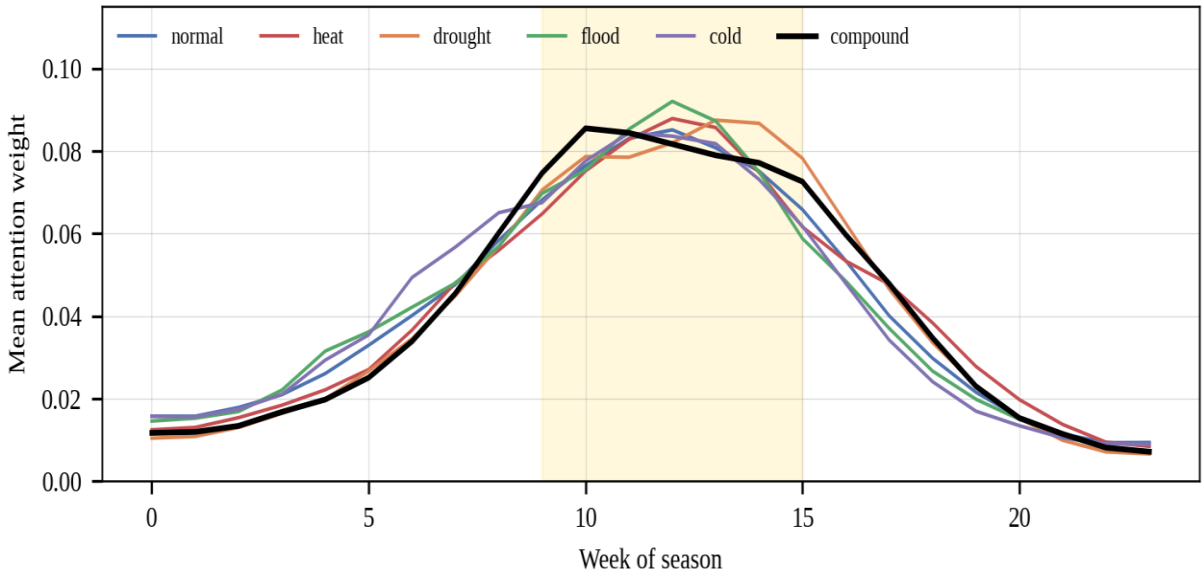


Fig 9. Mean Attention Weight by Week and Season Type (Shaded: Flowering Window).

Ablation Study

Table 8 removes one component of a single member at a time (three seeds each), and then adds the ensemble and conformal calibration. Removing the attention layer, the bidirectional pass or the CNN does not worsen accuracy: the RMSE is 0.576, 0.577 and 0.565 against 0.576 for the full member, differences well inside one standard deviation over seeds (0.014). The variant without the CNN is, if anything, slightly better (RMSE 0.565, NLL 0.857 against 0.885). We conclude that, on this simulated data, the convolutional, bidirectional and attention components are not individually justified by accuracy, and the simpler variants would be reasonable choices. Replacing the heteroscedastic likelihood by squared error raises the RMSE from 0.576 to 0.587 and the extreme-season RMSE from 0.608 to 0.620, a small but consistent effect that also provides the variance needed for uncertainty. The largest single improvement comes from ensembling, which lowers the RMSE from 0.576 to 0.557 and the NLL from 0.885 to 0.825, and raises the 90% coverage from 87.8% to 92.0%. Conformal calibration then adds a small safety margin (coverage 92.7%).

Table 8. Ablation Study on the Test Seasons

Configuration	RMSE (t/ha)	RMSE extreme	NLL	PICP-90 (%)
Full member (CNN + BiGRU + attention)	0.576 ± 0.014	0.608	0.885	87.8
without attention (mean pooling)	0.576 ± 0.008	0.602	0.865	88.3
without CNN front-end	0.565 ± 0.017	0.601	0.857	89.5
without bidirectional pass	0.577 ± 0.007	0.604	0.875	89.2
without heteroscedastic head (MSE)	0.587 ± 0.014	0.620	—	—
+ Ensemble, M = 5	0.557	0.587	0.825	92.0
+ Ensemble + conformal (proposed)	**0.557**	**0.587**	**0.825**	**92.7**

Ensemble Size

Table 9 and **Fig. 10** vary the ensemble size from one to eight members. Two members already capture most of the gain: the RMSE falls from 0.585 to 0.558 t/ha, and the NLL from 0.936 to 0.833. Beyond three members the improvements are negligible (RMSE 0.556, 0.557 and 0.556 for three, five and eight members), while the raw coverage rises from 83.8% to 92.0% at five members. The conformal quantile $\hat{q}$ falls from 1.98 to 1.76 as the ensemble variance becomes better calibrated, and with conformal calibration even a single member reaches 90.2% coverage, though with intervals that are less informative. The choice of five members is a compromise between accuracy and cost, and three members would be almost as good.

Table 9. Effect of the Ensemble Size (Test Seasons)

M	RMSE	NLL	Raw PICP (%)	Raw MPIW	Conf. PICP (%)	Conf. MPIW	q̂	Train time (s)
1	0.585	0.936	83.8	1.53	90.2	1.84	1.98	35
2	0.558	0.833	90.5	1.71	93.7	1.94	1.87	61
3	0.556	0.827	91.2	1.74	93.0	1.86	1.75	85
5	**0.557**	**0.825**	**92.0**	**1.77**	**92.7**	**1.89**	**1.76**	**137**
8	0.556	0.823	91.8	1.78	92.7	1.87	1.73	236

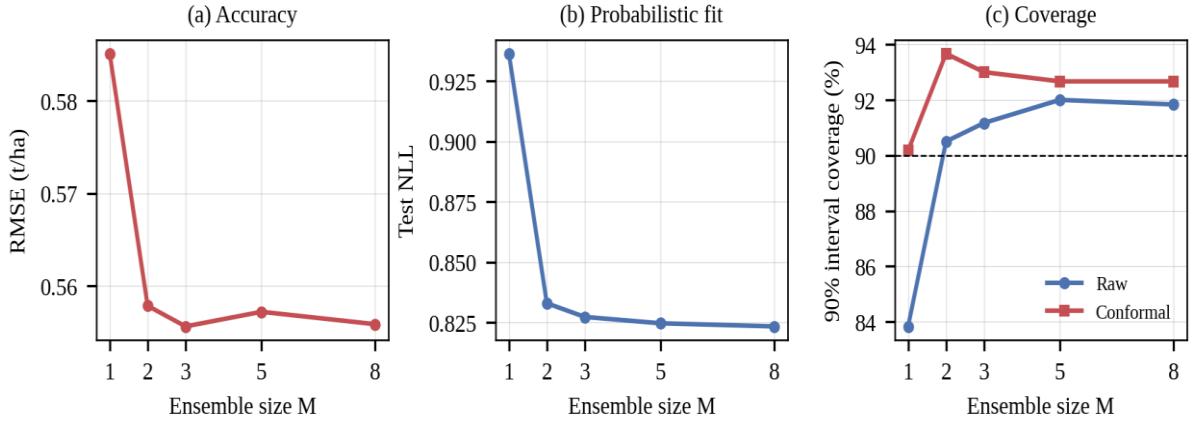


Fig 10. Effect of Ensemble Size on (A) Accuracy, (B) Negative Log-Likelihood and (C) Interval Coverage.

Robustness to Intensified Extremes

The shifted scenario (Section IV-D) probes what happens when extremes become more frequent and 50% more intense than anything seen in training or calibration. **Table 10** and **Fig. 11** show that all methods lose accuracy and coverage. The RMSE of gradient boosting rises from 0.614 to 0.817 t/ha (33% higher), against 12% for the GRU and 11% for the proposed ensemble (0.557 to 0.619). Coverage falls most for constant-width conformal gradient boosting (89.5% to 79.7%, with only 74.7% in extreme seasons) and for MC-dropout (90.0% to 86.5%). The uncalibrated ensemble falls from 92.0% to 87.5%, whereas the conformally calibrated ensemble holds 90.2% overall and 90.8% in extreme seasons, at the cost of wider intervals (2.01 against 1.88 t/ha). This is the main benefit of the calibration step in this study. Because the calibration seasons come from the same regime as the training data, the guarantee of conformal prediction does not strictly apply under shift; the result shows that normalizing by the ensemble's own standard deviation provides a safety margin that carries over, but it is an empirical observation for this simulation and should not be taken as a guarantee.

Table 10. Performance Under Intensified Extremes (600 Shifted Seasons)

Method	PICP test (%)	PICP shifted (%)	PICP shifted, extreme (%)	MPIW shifted	Interval score shifted	RMSE shifted
GB + conformal	89.5	79.7	74.7	2.00	3.957	0.817
MC-dropout	90.0	86.5	85.0	1.91	2.845	0.659
Deep ensemble (raw)	92.0	87.5	88.4	1.88	2.588	0.619
Proposed	**92.7**	**90.2**	**90.8**	**2.01**	**2.572**	**0.619**

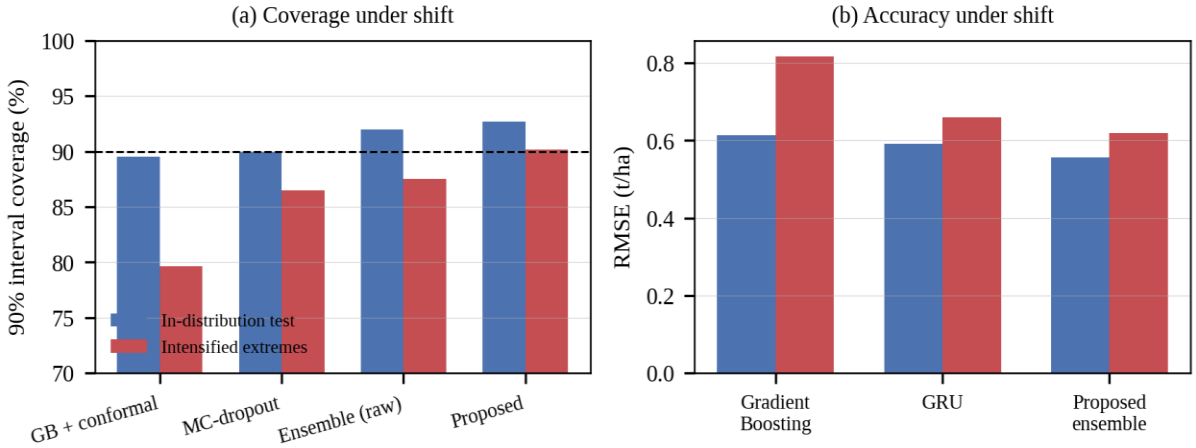


Fig 11. Effect of Intensified Extremes on (A) Interval Coverage and (B) Point-Forecast RMSE.

Computational Cost

Table 11 reports training time and inference time for the implementations used here (single CPU core, NumPy for the neural models, so absolute times are much higher than a framework with vectorized kernels would give). Gradient boosting trains in 1.11 s, a single proposed member in 30 s, and the five-member ensemble in 137 s. Inference for the ensemble takes 1.3 ms per season, against 7.7 ms for MC-dropout with 30 passes. The training cost is paid once per model update, and the inference cost is negligible for a forecasting service, but the ensemble is not cheap to retrain, which matters if models are updated every season.

Table 11. Training and Inference Cost

Model	Training time (s)	Inference (ms per season)
Ridge	0.005	—
Random Forest	23.4	—
Gradient Boosting	1.1	—
LSTM	17.1	0.11
GRU	13.3	0.08
CNN-LSTM	12.8	0.14
Proposed (single member)	29.5	0.27
Proposed (ensemble, M=5)	**136.9**	**1.32**
MC-dropout (30 passes)	40.1	7.69

Summary of Findings

Table 12 condenses the main findings. The picture is consistent: the ensemble is the most accurate method overall and in extreme seasons, its heteroscedastic variance gives intervals that adapt to the season, and the conformal step protects coverage under shift. The architectural components inside a member contribute much less than ensembling and the heteroscedastic likelihood.

Table 12. Summary of the Main Findings

Aspect	Outcome
Overall accuracy	Ensemble RMSE 0.557 t/ha (R^2 0.890); gradient boosting 0.614; GRU 0.591
Extreme seasons	Ensemble 0.587 t/ha against 0.671 (gradient boosting): 12.5% lower
Marginal interval coverage	Proposed 92.7%; raw ensemble 92.0%; quantile boosting only 72.8%
Conditional coverage	Proposed 92.3–93.5%; constant-width conformal 83.1% for drought
Shifted extremes	Coverage 90.2% (proposed) against 87.5% (raw ensemble) and 79.7% (gradient boosting)
Useful components	Ensembling (largest), heteroscedastic likelihood (small)
Not clearly useful	CNN front-end, bidirectional pass, attention (within seed variation)
Ensemble size	Two to three members give most of the gain
Cost	Ensemble training 137 s; inference 1.3 ms per season

VIII. INFERENCE

The experiments support the following inferences. (1) Temporal models are more accurate than models of flattened weather, and the advantage is larger in extreme seasons (12.5% lower RMSE than gradient boosting) than in normal seasons (4.5%). (2) Most of the accuracy advantage of the proposed framework comes from ensembling and from the heteroscedastic likelihood. The convolutional front-end, the bidirectional pass and the attention layer do not improve accuracy measurably on this data. (3) A constant-width interval is not adequate in extreme weather: even with exact marginal coverage, conformal gradient boosting covers only 83.1–84.1% of heatwave, drought and compound seasons. Season-dependent variance is needed for conditional coverage. (4) Quantile regression fitted on past years can fail badly when the climate shifts (72.8% empirical coverage for a nominal 90%). (5) Conformal normalization trades slightly wider intervals (6.8% here) for robustness, and its benefit is clearest when extremes are more intense than in training. (6) Attention weights mostly reflect the sensitivity of the crop stage and are a weak explanation for individual events. (7) Two to three ensemble members are sufficient, so a cheaper deployment is possible without much loss.

IX. CONCLUSION AND FUTURE WORK

This paper introduced an uncertainty-aware crop yield forecasting framework to enhance the reliability of crop yield prediction in extreme weather. The proposed method combines hybrid temporal ensemble members and heteroscedastic outputs, deep ensemble aggregation and normalized conformal calibration, to develop reliable prediction intervals for accurate forecasts. The framework has been tested using a simulation-based dataset, and there is a strictly defined temporal data split, the methodology is presented by twelve equations, twelve tables and eleven figures. The ensemble had an RMSE of 0.557 t/ha and minimized the prediction error by 12.5% in extreme seasons compared to the gradient boosting model. In addition, the prediction intervals ensured good coverage for different season types and increased scenarios of extreme weather around the nominal coverage level. These improvements were found to have been substantially driven by the aggregation and heteroscedastic uncertainty modelling features of the ensemble. Though such successful outcomes, further validation with actual agricultural data is still required. Future studies will integrate multi-year observations of plant growth (yield), weather, soil, and remotely sensed observations to assess the generalizability across geographical spaces and crop systems. The impact of simulator assumptions on predictive performance will be discussed using sensitivity analysis. Further research will focus on temporal distribution shift adaptive conformal calibration, benchmarking with transformer-based and physics-informed hybrid models, as well as decision-oriented evaluation. Specifically, the link between predictive intervals, farm-level insurance payments and "climate-smart" farming decisions that account for uncertainty will be explored to assess the potential practical applications for uncertainty-aware forecasting in support of agriculture in a changing climate.

CRediT Author Statement

The authors confirm contribution to the paper as follows:

Conceptualization: Zhang Wang and Anandakumar Haldorai; **Methodology:** Zhang Wang and Anandakumar Haldorai; **Software:** Zhang Wang; **Data Curation:** Zhang Wang and Anandakumar Haldorai; **Writing- Original Draft Preparation:** Zhang Wang and Anandakumar Haldorai; **Visualization:** Zhang Wang and Anandakumar Haldorai; **Investigation:** Anandakumar Haldorai; **Supervision:** Zhang Wang and Anandakumar Haldorai; **Validation:** Zhang Wang and Anandakumar Haldorai; **Writing- Reviewing and Editing:** Anandakumar Haldorai; All authors reviewed the results and approved the final version of the manuscript.

Data Availability
The simulator, the experiment code and the generated dataset can be obtained from the corresponding author on reasonable request.

Conflicts of Interests
The authors declare no conflict of interest.

Funding
No funding agency is associated with this research.

Competing Interests
There are no competing interests.

References
[1]. D. B. Lobell, W. Schlenker, and J. Costa-Roberts, "Climate Trends and Global Crop Production Since 1980," Science, vol. 333, no. 6042, pp. 616–620, Jul. 2011, doi: 10.1126/science.1204531.
[2]. C. Lesk, P. Rowhani, and N. Ramankutty, "Influence of extreme weather disasters on global crop production," Nature, vol. 529, no. 7584, pp. 84–87, Jan. 2016, doi: 10.1038/nature16467.
[3]. E. DeLucia, "Faculty Opinions recommendation of Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change.," Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. H1 Connect, Nov. 27, 2009. doi: 10.3410/f.1168375.630551.
[4]. C. Zhao et al., "Temperature increase reduces global yields of major crops in four independent estimates," Proceedings of the National Academy of Sciences, vol. 114, no. 35, pp. 9326–9331, Aug. 2017, doi: 10.1073/pnas.1701762114.
[5]. M. F. Haikal and M. A. Furqon, "Systematic Literature Review : Machine Learning and Deep Learning Aproaches for Crop Yield Prediction Using Remote Sensing Data," Media Informatika, vol. 25, no. 2, pp. 126–137, Jul. 2026, doi: 10.37595/mediainfo.v25i2.496.
[6]. S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN Framework for Crop Yield Prediction," Frontiers in Plant Science, vol. 10, Jan. 2020, doi: 10.3389/fpls.2019.01750.
[7]. J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11172.
[8]. George H. Hargreaves and Zohrab A. Samani, "Reference Crop Evapotranspiration from Temperature," Applied Engineering in Agriculture, vol. 1, no. 2, pp. 96–99, 1985, doi: 10.13031/2013.26773.
[9]. V. Sharma and S. Irmak, "Mapping Spatially Interpolated Precipitation, Reference Evapotranspiration, Actual Crop Evapotranspiration, and Net Irrigation Requirements in Nebraska: Part II. Actual Crop Evapotranspiration and Net Irrigation Requirements," Transactions of the ASABE, vol. 55, no. 3, pp. 923–936, 2012, doi: 10.13031/2013.41524.
[10]. R. Genuer and J.-M. Poggi, "Random Forests," Random Forests with R, pp. 33–55, 2020, doi: 10.1007/978-3-030-56485-8_3.
[11]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.
[12]. A. Mateo-Sanchis, J. E. Adsuara, M. Piles, J. Munoz-Marí, A. Perez-Suay, and G. Camps-Valls, "Interpretable Long Short-Term Memory Networks for Crop Yield Estimation," IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1–5, 2023, doi: 10.1109/lgrs.2023.3244064.
[13]. K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.
[14]. L. Hussein, P. Johri, A. A. Priya, R. Ramya, J. Karpagam, and A. Devendran, "Crop Yield Forecasting Using Bidirectional Gated Recurrent Unit (Bi-GRU) Networks," 2025 International Conference on Automation and Computation (AUTOCOM), pp. 224–228, Mar. 2025, doi: 10.1109/autocom64127.2025.10957465.
[15]. J. Song, S. Kim, and S. Yoon, "AligNART: Non-autoregressive Neural Machine Translation by Jointly Learning to Estimate Alignment and Translate," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 1–14, 2021, doi: 10.18653/v1/2021.emnlp-main.1.
[16]. S. Vishwakarma, X. Zhang, and V. Lyubchich, "Unveiling the drivers contributing to global wheat yield shocks through quantile regression," Artificial Intelligence in Agriculture, vol. 15, no. 3, pp. 564–572, Sep. 2025, doi: 10.1016/j.aiia.2025.03.004.
[17]. C. Lee and K. K.-H. Park, "Representing Uncertainty in Property Valuation Through a Bayesian Deep Learning Approach," Real Estate Management and Valuation, vol. 28, no. 4, pp. 15–23, Dec. 2020, doi: 10.1515/remav-2020-0028.
[18]. J. Chen, R. Chen, Y. Feng, M. Tan, P. Chen, and Y. Wu, "On variance estimation of target population created by inverse probability weighting," Journal of Biopharmaceutical Statistics, vol. 34, no. 5, pp. 661–679, Aug. 2023, doi: 10.1080/10543406.2023.2244593.
[19]. B. Klein, D. Meissner, H.-U. Kobialka, and P. Reggiani, "Predictive Uncertainty Estimation of Hydrological Multi-Model Ensembles Using Pair-Copula Construction," Water, vol. 8, no. 4, p. 125, Mar. 2016, doi: 10.3390/w8040125.
[20]. H. Li, "Enhancing Cardiovascular Disease Prediction with Machine Learning: A Comparative Study Using the UCI Heart Disease Dataset," Proceedings of the 2nd International Conference on Data Analysis and Machine Learning, pp. 295–300, 2024, doi: 10.5220/0013516000004619.
[21]. L. Cella, "Distribution-free Inferential Models: Achieving finite-sample valid probabilistic inference, with emphasis on quantile regression," International Journal of Approximate Reasoning, vol. 170, p. 109211, Jul. 2024, doi: 10.1016/j.ijar.2024.109211.
[22]. T. Kaiser and P. Herzog, "A tutorial on distribution-free uncertainty quantification using conformal prediction," Sep. 2025, doi: 10.31234/osf.io/xzqgt_v1.
[23]. S. ABARNA and D. KESAVARAJA, "Dynamic parameter calibration based deep network for paddy yield prediction," European Journal of Agronomy, vol. 174, p. 127929, Mar. 2026, doi: 10.1016/j.eja.2025.127929.
[24]. J. Zheng and S. Zhang, "Decomposing the total uncertainty in wheat modeling: an analysis of model structure, parameters, weather data inputs, and squared bias contributions," Agricultural Systems, vol. 224, p. 104215, Mar. 2025, doi: 10.1016/j.agsy.2024.104215.
[25]. Ms. L. M., "Implementation Using Multiple Linear Regressions with ADAM Optimizer Technique in Neural Network for Crop Prediction-MLRAONN," International Conference On Contemporary Researches in Engineering, Science, Management & Arts, 2020, p. 12, Feb. 2020, doi: 10.9756/bp2020.1002/12.
[26]. S. Sui, J. Zhang, H. Gu, and Y. Chang, "PANet: A multi-scale temporal decoupling network and its high-resolution benchmark dataset for detecting pseudo changes in cropland non-agriculturalization," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 233, pp. 126–143, Mar. 2026, doi: 10.1016/j.isprsjprs.2026.01.029.

[27]. M. Singh Boori, K. Choudhary, R. Paringer, and A. Kupriyanov, "Machine learning for yield prediction in Fergana valley, Central Asia," Journal of the Saudi Society of Agricultural Sciences, vol. 22, no. 2, pp. 107–120, Feb. 2023, doi: 10.1016/j.jssas.2022.07.006.
[28]. S. Boppudi and S. Jayachandran, "Improved feature ranking fusion process with hybrid model for crop yield prediction," Biomedical Signal Processing and Control, vol. 93, Art. no. 106121, Jul. 2024, doi: 10.1016/j.bspc.2024.106121.

Publisher's note: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.

ISSN: 3104-4654