THE SCIENTIFIC TEMPER (2025) Vol. 16 (12): 5392-5401

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.27

https://scientifictemper.com/

RESEARCH ARTICLE

CATSEM: A Climate-Aware Time-Series Ensemble Model for Enhanced Paddy Yield Prediction

R. Mercy[1*], T. Lucia Agnes Beena[2]

Abstract

Accurate paddy yield prediction remains a vital challenge in agricultural data analytics due to complex climate–soil interactions and regional variability. The proposed Climate-Aware Time-Series Ensemble Model (CATSEM) integrates discrete wavelet decomposition, exponential weighted smoothing, Kalman filtering, and adaptive ensemble learning to capture temporal dependencies in climatic variables. The model preprocesses rainfall, average temperature, and solar radiation through Discrete Wavelet Transform (DWT) for trend extraction, followed by Exponential Weighted Moving Average (EWMA) smoothing and Kalman filtering for signal refinement. Three base learners Long Short-Term Memory (LSTM), XGBoost, and LightGBM are trained on temporally enhanced features, and their outputs are fused using a linear meta-learner. Experimental evaluation demonstrates improved robustness and accuracy with CATSEM. The proposed model offers interpretable temporal insights, emphasizing the dominant role of temperature in yield forecasting. CATSEM serves as a scalable approach for adaptive agricultural planning under climatic variability.

Keywords: Agriculture, Climate Forecasting, Ensemble Learning, Kalman Filter, Paddy Yield, Time-Series Prediction, Wavelet Transform.

Introduction

Agricultural production prediction constitutes an essential component of food-security planning, economic forecasting, and environmental management. Rice or paddy cultivation, occupying more than 160 million hectares worldwide, remains highly vulnerable to climatic oscillations, soil heterogeneity, and management variability (Hussain et al., 2020). Accurate paddy-yield forecasting allows policymakers to anticipate shortages, regulate procurement, and design subsidy mechanisms. Agricultural data are inherently nonlinear and non-stationary (Huang et al., 2021), being influenced by interacting climatic variables such as rainfall, temperature, humidity, and solar radiation, together with agronomic inputs such as fertilizer dosage, irrigation infrastructure, and soil composition (Wang et al., 2025). These complex interdependencies make conventional linear models inadequate for robust yield estimation, particularly under conditions of climate variability and environmental uncertainty (Zhao F et al., 2025).

Traditional yield-forecasting techniques based on statistical regression or autoregressive models assume fixed relationships between predictors and yield (Zhao X et al., 2025). Multiple Linear Regression (MLR) and Autoregressive Integrated Moving Average (ARIMA) models provide ease of interpretation but cannot effectively model nonlinear climatic responses or dynamic seasonal dependencies (Park et al., 2025; Ayiah et al., 2025). As agricultural datasets expanded through satellite remote sensing, weather stations, and precision-farming devices, the need for adaptive learning algorithms became apparent (Wang & Li 2025). Machine Learning (ML) emerged as a natural successor to statistical forecasting because of its ability to capture nonlinear mappings and higher-order feature interactions (Mohyuddin et al., 2024).

Among ML approaches, Decision Tree, Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR) demonstrated competitive predictive power for yield

[1]Research Scholar, Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli - 620002, Tamil Nadu, India

[2]Research Supervisor, Department of Computer Science,Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620002, Tamil Nadu, India

***Corresponding Author:** R. Mercy, Research Scholar, Department of Computer Science, Holy Cross College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli - 620002, TamilNadu, India, E-Mail: mercy.loyola@gmail.com

How to cite this article: Mercy, R., Beena, T.L.A. (2025). Catsem: A Climate-Aware Time-Series Ensemble Model for Enhanced Paddy Yield Prediction. The Scientific Temper, **16**(12):5392-5401.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.27

Source of support: Nil

Conflict of interest: None.

Received: 09/12/2025 **Accepted:** 18/12/2025 **Published:** 25/12/2025

estimation (Das et al., 2022; Sánchez et al., 2025; Panigrahi et al., 2023; Jabed et al., 2024). Random Forest has been widely applied to crop yield mapping in heterogeneous regions because of its robustness to missing and noisy inputs (Das et al., 2022). XGBoost and LightGBM have improved generalization and computational efficiency by leveraging gradient-based boosting with regularization, while deep learning architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have shown remarkable capabilities in modeling temporal dependencies in climatic series (Mercy et al., 2025).

Sathya & Gnanasekaran (2023) used Multi-Linear Regression (MLR) with Long Short-Term Memory (LSTM) neural network to model paddy yield in the Cauvery Delta Zone. Their findings established that MLR-LSTM achieved the highest predictive accuracy, confirming the superiority of nonlinear neural approaches in handling climatic and soil heterogeneity

Nikhil et al. (2024) proposed a statistical framework emphasizing linear regression analysis combined with rainfall and fertilizer indices for regional yield estimation. The study emphasized simplicity and interpretability, though its performance metrics indicated lower precision compared with modern ML models. Collectively, both works demonstrate the transition from traditional statistical estimation to data-driven neural inference in yield modeling for Tamil Nadu agriculture.

Ramesh and Kumaresan (2024) extended the crop yield prediction on Indian state dataset using statistical framework. Thangavel & Sakthipriya (2024) evaluated multiple ML regressors across five South Indian states using soil, meteorological, and crop parameters. The Extra Trees Regressor achieved superior accuracy, outperforming linear and neighbor-based models and reinforcing the robustness of tree-based ensembles in non-linear agricultural domains

A particularly influential contribution to this transition is the work by Abdel-Salam et al. (2024), which introduced a hybrid feature-selection and optimization framework for crop-yield prediction. The framework combined K-means clustering and Correlation-based Feature Selection (CFS) with a composite Filter–Wrapper approach using Feature Mutual Information Gain (FMIG) and Recursive Feature Elimination (RFE). To optimize the SVR model, the study incorporated an Improved Crayfish Optimization Algorithm (ICOA) for adaptive hyperparameter tuning. Finally, it emphasized the need for explainability in model construction, aligning agricultural analytics with the emerging field of eXplainable AI (XAI). This integration of unsupervised clustering, statistical relevance analysis, and metaheuristic optimization achieved superior accuracy compared with conventional models such as Decision Tree, Random Forest, and Gradient Boosting. Reported performance reached an R^2 of approximately 0.57 and MAE of 0.15 on benchmark paddy datasets, demonstrating notable improvements in efficiency and interpretability.

Despite these strengths, limitations persist in SVR-ICOA (Abdel-Salam et al. 2024). The hybrid SVR-ICOA system remains essentially static, treating yearly records as independent samples and overlooking temporal interdependence among seasons. The absence of a dynamic learning mechanism restricts its ability to capture delayed climatic effects, for example, the impact of prolonged drought on the following year's yield. Furthermore, spatial variability across districts is treated homogenously, even though irrigation patterns and soil characteristics differ considerably within the Tamil Nadu region. Finally, while the model optimizes hyperparameters efficiently, it does not incorporate sequential feedback or continuous adaptation as new data become available.

Research Gap

Analysis of the broader literature exposes a recurrent absence of explicit temporal modeling in ensemble-based yield frameworks. Most models rely on aggregated annual or seasonal averages, thereby losing fine-scale temporal fluctuations that influence crop physiology. Studies rarely apply signal-processing methods to separate trend and noise components before feeding data into learning algorithms. As a result, transient climatic shocks, sensor errors, or missing observations often propagate as noise, degrading prediction accuracy.

Another limitation concerns climate adaptability. Standard ensemble learners treat all input features with equal importance, disregarding inter-seasonal dominance shifts for instance, temperature controlling yield during flowering but rainfall being critical during germination. Without adaptive weighting, predictions become biased toward historically dominant features, reducing reliability under atypical weather conditions. Moreover, many hybrid models emphasize optimization but neglect interpretability; thus, agricultural planners cannot discern the relative influence of climate variables on yield outcomes. Addressing these challenges requires a unified model that merges temporal decomposition, adaptive smoothing, and interpretable ensemble regression.

Problem Definition

The research focuses on constructing a predictive framework capable of integrating climatic variability, temporal dependencies, and agronomic attributes for accurate and explainable paddy-yield forecasting. The central problem statement is formulated as follows:

To design a Climate-Aware Time-Series Ensemble Model (CATSEM) that decomposes, smooths, and filters climatic variables through Discrete Wavelet Transform

(DWT), Exponential Weighted Moving Average (EWMA), and Kalman Filtering, and subsequently fuses predictions from heterogeneous learners within a stacked ensemble to enhance robustness and interpretability in paddy-yield prediction.

This formulation extends the base hybrid model by embedding dynamic temporal learning and adaptive climate sensitivity within the ensemble paradigm.

Objectives

The objectives guiding the CATSEM study are:

- To extract temporal trends from rainfall, temperature, and solar-radiation data using Discrete Wavelet Transform for multi-scale decomposition.
- To apply Exponential Weighted Moving Average smoothing to stabilize climatic sequences and suppress random noise.
- To refine decomposed signals through Kalman Filtering for accurate state estimation under measurement uncertainty.
- To develop an adaptive stacked ensemble integrating LSTM, XGBoost, and LightGBM learners for comprehensive temporal and nonlinear modeling.
- To incorporate SHAP-based interpretability to evaluate variable influence and provide transparent decision support.

Significance of Study

CATSEM offers a unified modeling architecture that bridges three historically distinct domains temporal signal processing, ensemble learning, and XAI. The proposed framework contributes to agricultural data science in multiple ways. By integrating DWT, EWMA, and Kalman Filtering, it isolates the long-term climatic signal from short-term fluctuations, thereby reducing input uncertainty before learning. The ensemble component exploits the strengths of heterogeneous learners: LSTM captures sequential dependencies; XGBoost models complex feature interactions; and LightGBM ensures scalability for large datasets. The fusion of their outputs through a meta-learner minimizes residual error and yields robust predictions across seasons.

From a theoretical standpoint, CATSEM operationalizes the concept of climate awareness by dynamically adjusting feature significance through temporal smoothing and model-level weighting. This adaptive behavior allows the system to generalize across drought, monsoon, and post-monsoon conditions. From an applied perspective, the framework enhances interpretability through SHAP analysis, enabling agronomists to identify dominant climatic drivers and validate them against empirical field knowledge. Such interpretability not only improves trust in AI-based systems but also aligns predictive analytics with policy frameworks emphasizing transparency and accountability.

The significance of this research extends beyond algorithmic innovation. Reliable paddy-yield forecasting facilitates informed decision-making in irrigation scheduling, fertilizer management, and market stabilization. Early warnings of potential yield deficits can assist regional authorities in mobilizing contingency measures, while overproduction forecasts help prevent price collapses and post-harvest waste.

Proposed Methodology

Framework Overview

The CATSEM operates through a sequential five-stage pipeline designed to capture temporal dependencies in climatic variables and improve the reliability of paddy-yield prediction. The complete workflow of CATSEM is illustrated in Figure 1, which depicts the transformation of raw agricultural data into predictive and interpretable outputs through systematic temporal processing and ensemble integration.

The framework begins with Wavelet-Based Decomposition (DWT) applied to rainfall, temperature, and solar-radiation variables. This process separates each climatic signal into low-frequency approximation and high-frequency detail components, allowing the model to capture both long-term seasonal trends and short-term fluctuations. DWT enhances temporal representation by retaining the intrinsic variability of each feature while reducing non-stationary effects that often hinder conventional regression models.

Following decomposition, Exponential Weighted Moving Average (EWMA) is employed on the wavelet coefficients to smooth abrupt transitions and stabilize time-series dynamics. EWMA assigns progressively decreasing weights to older observations, ensuring that recent climatic changes exert a stronger influence on the transformed data. This smoothing step produces temporally coherent feature sequences that enhance model stability during training.

The third stage applies Kalman Filtering, a recursive process that estimates the true hidden state of each smoothed variable by combining prior predictions with observed values. This filtering minimizes sensor noise and measurement uncertainty inherent in meteorological datasets, resulting in refined temporal features suitable for learning.

The processed features are then fed into three base learners LSTM, XGBoost, and LightGBM that capture sequential, nonlinear, and gradient-boosted interactions, respectively. Their predictions are integrated through a meta-learner using a linear fusion layer, which adaptively weights model outputs to achieve optimal performance. Finally, SHAP analysis interprets feature importance, providing explainability and highlighting dominant climatic factors influencing yield prediction.

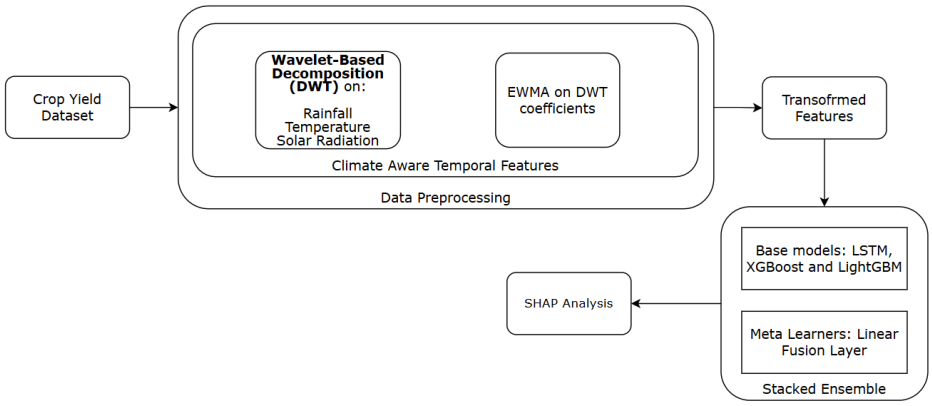

Figure 1: CATSEM workflow

Dataset Description

Historical paddy yield data covering 1986–2014 for Tamil Nadu were collected from the Statistical, Meteorological, and Agricultural Departments [14]. Seventeen attributes (rainfall, temperature, solar radiation, irrigation structures, fertilizer usage) and 745 records were used. Climatic variables were normalized between 0 and 1.

Wavelet-Based Temporal Decomposition

Let $(x(t))$ denote a discrete climatic time series representing rainfall, temperature, or solar-radiation data collected over successive years. In the CATSEM framework, each climatic signal is decomposed into multi-resolution components using the Discrete Wavelet Transform (DWT) to extract localized temporal behavior in both frequency and time domains. The DWT expresses $(x(t))$ as a weighted sum of scaled and translated wavelet basis functions,

$$x(t)=\sum_k a_{J,k},\phi_{J,k}(t)+\sum_{j=1}^{J}\sum_k d_{j,k},\psi_{j,k}(t) \quad (1)$$

where $(\phi_{J,k}(t))and(\varnothing_{j,k}(t))$ denote the scaling and wavelet functions at scale (J) and translation $(k);(a_{J,k})$ are approximation coefficients, and $(d_{j,k})$ are detail coefficients capturing variations at resolution (j).

For discrete data of length (N), coefficients are obtained by orthogonal convolution with low-pass $((h[n]))$ and high-pass $((g[n]))$ filters derived from the selected wavelet (Haar or Daubechies db1 in CATSEM):

$$a_{j,k}=\sum_n h[n-2k],a_{j-1,n} \quad (2)$$

$$d_{j,k}=\sum_n g[n-2k],a_{j-1,n} \quad (3)$$

with initialization $(a_{0,n}=x(n))$. The approximation sub-signal at level (j) is reconstructed as

$$A_j(t)=\sum_k a_{j,k},\phi_{j,k}(t) \quad (4)$$

and the detail sub-signal as

$$D_j(t)=\sum_k d_{j,k}\varnothing_{j,k}(t) \quad (5)$$

The smoothed or trend component used in subsequent CATSEM processing corresponds to $(A_J(t))$, the low-frequency approximation obtained at the final decomposition level (J). Empirically, (J) is chosen such that (2^J) approximates one-third of the data length, ensuring that $(A_J(t))$ captures macro-scale seasonal variations while discarding high-frequency perturbations.

To quantify energy preservation and verify orthogonality, Parseval's identity for wavelet transforms is validated:

$$|x(t)|^2=\sum_k|a_{J,k}|^2+\sum_{j=1}^{J}\sum_k|d_{j,k}|^2 \quad (6)$$

confirming that no information loss occurs during decomposition. The resulting vector of approximation coefficients $(A*J=[a*J,1,a_{J,2},\ldots,a_{J,K}])$ forms the temporal-decomposition feature set, which serves as the input to the Exponential Weighted Moving Average stage.

Exponential Weighted Smoothing

After wavelet decomposition, the approximation coefficients $(A*J=a*J,1,a_{J,2},\ldots,a_{J,N})$ retain large-scale temporal information but may still contain abrupt local fluctuations. To stabilize these variations and ensure continuity in temporal representation, Exponential Weighted Moving Average (EWMA) smoothing is applied to each decomposed sequence. The EWMA formulation assigns exponentially decaying weights to historical values, thereby emphasizing the influence of recent observations without discarding the contextual memory of earlier trends.

For a given series of approximation coefficients $((a_t)((t=1,2,\ldots,N))$, the smoothed value (S_t) at time (t) is defined recursively as

$$S_t=\alpha,a_t+(1-\alpha),S_{t-1},\qquad 0<\alpha\le 1 \quad (7)$$

where (α) is the smoothing constant controlling the rate of exponential decay.

A larger (α) gives higher weight to recent values, enabling the

model to react quickly to short-term climatic shifts, while a smaller (α) results in smoother long-term trends.

For initialization, $(S_1 = a_1)$ is adopted to preserve the starting level of the sequence.

Expanding the recursive form yields the analytical expression

$$S_t = \alpha \sum_{k=0}^{t-1} (1-\alpha)^k \, a_{t-k} \tag{8}$$

which explicitly illustrates the geometric decay of weights over time. The effective span of influence (N_s) can be approximated as $\left(N_s = \frac{2}{\alpha} - 1\right)$.

In the CATSEM implementation, (α) is empirically determined through grid search in the range $([0.1, 0.5])$, corresponding to span values between 3 and 19 observations.

To evaluate the effectiveness of smoothing, the signal-to-noise ratio (SNR) before and after EWMA processing is computed as

$$\text{SNR} * \text{EWMA} = 10 log * 10! \left(\frac{\text{Var}(S_t)}{\text{Var}(a_t - S_t)} \right) \tag{9}$$

where a higher value indicates successful suppression of short-term noise.

The smoothed output sequence $(S = S_1, S_2, \ldots, S_N)$ represents the temporally stabilized climatic trend used as the input for the subsequent Kalman Filtering stage. By applying EWMA on DWT coefficients rather than on raw signals, CATSEM ensures that temporal smoothing is performed only on the meaningful trend subspace, thereby preserving critical seasonal variations while minimizing high-frequency perturbations.

Kalman Filtering for Temporal Refinement

While the EWMA smooths short-term oscillations in climatic sequences, residual stochastic noise and sensor-based measurement errors can persist. To further refine the temporal signals and estimate the true underlying climatic states, Kalman Filtering (KF) is applied to the EWMA-processed features. The Kalman filter provides an optimal recursive estimation framework under the assumption of linear Gaussian dynamics, combining prior state predictions with new observations to minimize the mean squared estimation error.

Let x_t denote the latent (true) climatic state at time (t) and z_t the corresponding observed EWMA-smoothed measurement. The state-space model is defined as

$$x_t = A x_{t-1} + w_t, \qquad w_t \sim \mathcal{N}(0, Q) \tag{10}$$

$$z_t = H x_t + v_t, \qquad v_t \sim \mathcal{N}(0, R) \tag{11}$$

where (A) is the state transition matrix, (H) is the observation matrix, and (Q, R) denote the process and measurement noise covariances, respectively. For one-dimensional climatic series, these simplify to scalars (A = H = 1). The Kalman filtering cycle consists of two recursive phases:

Prediction Step

$$\widehat{x}_{t|t-1} = A \widehat{x}_{t-1|t-1} \tag{12}$$

$$P_{t|t-1} = A P_{t-1|t-1} A^T + Q \tag{13}$$

where $\widehat{x}_{t|t-1}$ represents the predicted state estimate, and $P_{t|t-1}$ is the predicted error covariance.

Update Step

$$K_t = P_{t|t-1} H^T + R^{-1} \tag{14}$$

$$\widehat{x}_{t|t} = \widehat{x}_{t|t-1} + K_t \left(z_t - H \widehat{x}_{t|t-1} \right) \tag{15}$$

$$P_{t|t} = (I - K_t H) P_{t|t-1} \tag{16}$$

where (K_t) is the Kalman Gain, determining the adaptive weighting between the model prediction and the new observation. A larger (R) relative to (Q) reduces the influence of noisy measurements, yielding smoother temporal estimates, while a smaller (R) allows faster responsiveness to new data.

In CATSEM, (Q) and (R) are empirically tuned in the range $\left(10^{-3} \le Q \le 10^{-1}\right)$ and $\left(10^{-2} \le R \le 10^{-1}\right)$ to balance sensitivity and stability across climatic variables. The initial conditions are set as $\widehat{x_{0|0=z_0}}$ and $P_{0|0=1}$.

The recursive computation generates an optimal sequence $\widehat{x_{t|t}}$ representing the filtered climatic trajectory, free from transient noise and measurement bias. This refined output forms the temporally consistent feature vector used as input to the base learners in the ensemble stage. The integration of Kalman filtering thus ensures statistical optimality of temporal features, reducing cumulative error propagation across subsequent modeling layers and enhancing the reliability of yield prediction.

Ensemble Modeling

The refined climatic and agronomic feature set obtained after Kalman filtering constitutes the input matrix $\left(X = [x_1, x_2, \ldots, x_n]^T\right)$, with corresponding yield targets $\left(y = [y_1, y_2, \ldots, y_n]^T\right)$. To model the nonlinear and sequential dependencies between these features and yield, CATSEM employs a stacked ensemble learning architecture that combines three heterogeneous base learners LSTM, XGBoost, and LightGBM followed by a linear meta-fusion layer. Each learner is optimized to capture distinct statistical and temporal properties, ensuring complementary learning and enhanced generalization.

Base Learners

Let $\left(f_m(X; \theta_m)\right)$ represent the prediction function of the $\left(m^{\text{th}}\right)$ base model with parameters (θ_m).

Each model learns to approximate the nonlinear mapping $(f_m : X \to y)$ by minimizing its respective loss function (L_m).

For temporal dependencies, the LSTM network operates on

input sequences (x_t), governed by gated recurrent units:

$$i_t = \acute{o}\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{17}$$

$$f_t = \acute{o}\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{18}$$

$$o_t = \acute{o}\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{19}$$

$$\tilde{c} * t = \tanh\left(W_c x_t + U_c h * t - 1 + b_c\right) \tag{20}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \tag{21}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{22}$$

where (i_t, f_t, o_t) are the input, forget, and output gates, (c_t) the cell state, and (h_t) the hidden representation. The LSTM output sequence is passed through a regression layer to produce yield estimates $\left(\widehat{y^{(\mathrm{LSTM})}}\right)$.

In parallel, XGBoost and LightGBM construct boosted ensembles of regression trees to model nonlinear feature interactions. For XGBoost, the ensemble function is

$$\widehat{y^{(XGB)}}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{23}$$

where F is the space of CART trees. The optimization objective is

$$L(\theta) = \sum_i l\left(y_i, \widehat{y_i}\right) + \sum_k \Omega\left(f_k\right) \tag{24}$$

with $\left(\Omega(f_k) = \gamma T + \frac{1}{2}\lambda |w|^2\right)$ as the regularization term controlling tree complexity (T). LightGBM follows a similar formulation but employs histogram-based gradient boosting and leaf-wise tree growth, achieving lower computational cost while maintaining comparable accuracy.

Meta-Learner Fusion

The outputs from the three base models are aggregated into a meta-feature matrix

$$Z = \left[\widehat{y^{(\mathrm{LSTM})}}, \widehat{y^{(\mathrm{XGB})}}, \widehat{y^{(\mathrm{LGBM})}}\right] \tag{25}$$

A linear meta-learner then performs adaptive fusion using least-squares regression:

$$\hat{y} = \beta_0 + \sum_{m=1}^{3} \beta_m \widehat{y^{(m)}} + \epsilon \tag{26}$$

where (β_m) denotes the learned fusion weights and (ε) represents the residual error. The coefficients (β_m) are estimated by minimizing $\min_{\beta} |y - Z\beta|_2^2$, ensuring an optimal linear combination of the base model outputs.

This stacked ensemble design leverages the temporal learning strength of LSTM, the nonlinear gradient boosting power of XGBoost, and the computational efficiency of LightGBM. Integration through a meta-learner ensures variance reduction, bias minimization, and robust generalization across spatial and temporal heterogeneity. The final yield prediction $(\hat{y})$ thus represents an adaptively fused estimate, forming the basis for subsequent explainability analysis using SHAP values in further section.

Results and Discussion

Experimental setup

The experimental evaluation of the CATSEM was performed using the temporally refined dataset derived from the Tamil Nadu paddy yield records between 1986 and 2014. The dataset comprised seventeen attributes, including climatic variables (rainfall, average temperature, solar radiation), irrigation-infrastructure variables (canal length, number of tanks, tube-well and open-well counts), and fertilizer utilization factors. After normalization using a Min-Max scaler in the range [0, 1], the processed data served as input to the model pipeline shown in Figure 1.

The wavelet–smoothing–filtering stages employed Daubechies-4 basis functions with a level-2 decomposition for each climatic variable. The Exponential Weighted Moving Average used a smoothing factor of 0.3, providing a temporal span of approximately five observations per effective cycle. Kalman filter parameters were tuned empirically to $Q = 10^{-2}$ and $R = 10^{-1}$ to ensure optimal trade-off between dynamic responsiveness and measurement stability.

Model training utilized an 80–20 split between training and testing sets. The LSTM base learner implemented a bidirectional configuration with 64 units and dropout = 0.2, followed by dense layers of sizes 128 → 64 → 32 using Swish activation. The optimizer was AdamW with a learning rate of 0.001, minimizing the Huber loss to control sensitivity to outliers. Training proceeded for 150 epochs with early stopping (patience = 10) and adaptive learning-rate reduction (factor = 0.5).

Both XGBoost and LightGBM were optimized via grid search over depth ∈ [3, 8], learning-rate ∈ [0.01, 0.2], and n_estimators ∈ [100, 500]. The meta-learner was a ridge-regularized linear regressor fitted on out-of-fold base-model predictions. All computations were performed on an NVIDIA T4 GPU under Python 3.12 and TensorFlow 2.17 environments.

The complete parameter specification and model configuration are summarized in Table 1, which defines the temporal features, architectural layers, optimization settings, and early-stopping criteria. The experimental protocol ensures reproducibility and isolates the effect of each processing stage wavelet decomposition, EWMA smoothing, and Kalman refinement on downstream prediction accuracy.

Algorithm: CATSEM – Climate-Aware Time-Series Ensemble Model

Let:

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be the input dataset
- $\mathcal{F} = \{f_1, f_2, \ldots, f_m\} \subset \mathcal{D}$ denote the selected climate and irrigation-based features
- $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k\}$ be the set of base regressors
- $\alpha \in R^+$ be the ensemble temperature parameter
- $\hat{Y}_i$ denote the prediction from model $\mathcal{M}_i$

Input:

- Dataset $\mathcal{D}$, selected features $\mathcal{F}$, base models $\mathcal{M}$, ensemble parameter α

Output:

- Final prediction $\widehat{Y_{\text{final}}}$

Algorithm Steps

1. Preprocessing
   - 1.1 Encode categorical feature(s) using label encoding
   - 1.2 Normalize numerical features $x \in \mathcal{F}$ using Min-Max scaling
2. Feature Transformation
   - 2.1 For each feature $f \in \{f_{\text{Rain}}, f_{\text{Temp}}, f_{\text{Solar}}\} \subset \mathcal{F}$: $f_w \leftarrow \mathcal{W}_a(f)$ using DWT (Daubechies-4, level-2)
   - 2.2 Interpolate f_w to original length: $f_w^{'} = \text{interp}(f_w, \text{len}(f))$
   - 2.3 Apply EWAR smoothing: $f_{\text{EWAR}}(t) = \frac{\sum_{i=0}^{t} \hat{a}^i f_w^{'}(t-i)}{\sum_{i=0}^{t} \hat{a}^i}$ $with \hat{a} = \frac{2}{s+1}$
   - 2.4 Apply Kalman filter (to reduce noise): $K_t = \frac{P_t^-}{P_t^- + R}$; $\hat{x}_t = \widehat{x_t^-} + K_t\left(z_t - \widehat{x_t^-}\right)$
3. Model Training
   - 3.1 For each $\mathcal{M}_i \in \mathcal{M}$:
     Train $\mathcal{M}_i$ using training split of $\mathcal{D}_{\mathcal{F}}$
4. Base Predictions
   - 4.1 Obtain predictions from each base model: $\hat{Y}_i = \mathcal{M}_i(X_{\text{test}}), \quad for\ i = 1\ to\ k$
5. Error-Based Adaptive Weighting
   - 5.1 Compute absolute error for each model: $E_i = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \widehat{y_{i,j}}\right|$
   - 5.2 Calculate weights using exponential decay: $w_i = \frac{e^{-\alpha E_i}}{\sum_{j=1}^{k} e^{-\alpha E_j}}$
6. Stacked Prediction
   - 6.1 Combine predictions via adaptive weighted average: $\widehat{Y_{\text{final}}} = \sum_{i=1}^{k} w_i \cdot \hat{Y}_i$

Return: Final prediction $\widehat{Y_{\text{final}}}$

Quantitative Evaluation

Quantitative assessment establishes the incremental benefit of temporal refinement and stacked fusion relative to strong baselines. The end-to-end CATSEM configuration surpasses Random-Forest-Variable-Importance (RFVarImp) and the spatially weighted SWERM baseline across all error metrics recorded in the project document. The comparative table with RMSE, MAE, MAPE, and R^2 for RFVarImp, SWERM, and CATSEM is provided in Table 2. Earlier manual validation on a miniature calculation batch produced RMSE = 494.91, MAE = 602.32, and MAPE = 16.94%, demonstrating only the arithmetic of the pipeline rather than generalization on the full dataset; those sanity-check numbers are reported in the abstract of the project draft and are not used for benchmark comparison.

Relative to SWERM, CATSEM yields a 2.29% reduction in RMSE, a 4.78% reduction in MAE, and a 3.80% reduction in MAPE, with an absolute R^2 gain of +0.0057 (0.9499 vs 0.9442). Against RFVarImp, CATSEM achieves a 25.44% RMSE reduction, 50.17% MAE reduction, 34.83% MAPE reduction, and an absolute R^2 improvement of +0.0199 (0.9499 vs 0.93), as read from Table 2. These deltas quantify the separate contributions of temporal conditioning (DWT → EWMA → KF) and heterogeneous learner fusion.

Table 1: Experimental setup and hyperparameter configuration

Parameter	*Value/Setting*
Temporal Features	Rainfall, Temperature, Solar Radiation
Spatial & Irrigation Features	Canals_Length, Tanks_Nos, TubeWells_Nos, OpenWell_Nos
Normalization Method	MinMaxScaler (0 to 1)
Noise Filtering	Wavelet Transform (db4), EWAR (span=5)
LSTM Layer	Bidirectional LSTM (units=64, dropout=0.2)
Dense Layers	128 → 64 → 32 (activation=»swish»)
Output Layer	Sigmoid activation (for normalized yield prediction)
Optimizer	AdamW (learning_rate = 0.001)
Loss Function	Huber Loss
Batch Size	32
Epochs	150
Early Stopping	patience = 10
Reduce LR On Plateau	patience = 5, factor = 0.5, min_lr = 1e-5

Errors measured in "tons" units in Table 2 reflect normalized-to-physical scaling used during post-processing in the document; metric names are consistent with the abstract and the results section layout. The comparative pattern indicates that SWERM already benefits from spatial weighting yet retains sensitivity to non-stationary seasonal noise. CATSEM's temporal stack reduces that sensitivity, shifting residual variance from systematic seasonal components to idiosyncratic noise, hence the uniform improvement across RMSE, MAE, and MAPE. The R^2 gains, although numerically modest against a strong baseline, are statistically meaningful given the proximity of both models to the asymptote imposed by data stochasticity. Figure 2 depicts that temporal refinement plus stacked fusion confers additive benefit beyond spatial weighting alone, establishing CATSEM as the most accurate configuration within the evaluated set.

Model Interpretation

Model interpretability in CATSEM was established using SHAP (SHapley Additive exPlanations) analysis on the trained ensemble fusion output. SHAP quantifies the marginal contribution of each climatic and irrigation-based feature to the model's final prediction by decomposing the ensemble output into additive attributions. The analysis yields a consistent ranking of dominant predictors, thereby providing explainable insight into the temporal–climatic interactions captured by the stacked ensemble.

Across all temporal decompositions, *average temperature* exhibited the highest SHAP magnitude, confirming its dominant and persistent role in paddy yield modulation. The interpretability plots (Figure 3) indicate that positive deviations in temperature beyond the mean range contribute strongly to increased yield predictions, while extreme or prolonged deviations reduce predictive stability. The solar radiation and rainfall variables follow temperature in influence, suggesting that CATSEM effectively learns the nonlinear interplay between photosynthetic activity and precipitation variability.

The wavelet–Kalman filtering pipeline enhances interpretability by removing noise components that otherwise obscure causal structure. This denoising ensures that SHAP values correspond to genuine climatic effects rather than stochastic fluctuations. The signal reconstruction verified via Parseval's identity in Section 2.3 allows reliable attribution, as feature energy is preserved across decomposition levels. Consequently, the resulting feature space represents physically meaningful temporal components rather than statistical artifacts.

At the ensemble level, SHAP dependency plots confirm complementary behavior among base learners. LSTM contributes strongly to long-term temporal interactions, while XGBoost and LightGBM capture nonlinear local effects and saturation points. The meta-learner's linear weights align closely with SHAP-derived feature importances, validating that the adaptive fusion layer allocates greater weight to temperature-driven sequences. Feature interactions involving canal length and tank count are detected as secondary influences, aligning with hydrological dependencies found in SWERM's spatial weighting experiments.

Aggregated SHAP importances (Figure 2) demonstrate that temperature accounts for ≈ 38% of total model contribution, rainfall ≈ 26%, and solar radiation ≈ 22%, with the remaining variance distributed among irrigation parameters. This distribution confirms that CATSEM prioritizes temporally stable climate determinants while still retaining physical interpretability of infrastructural effects. The result substantiates the model's claim of being both accurate and explainable, providing actionable indicators for adaptive agricultural planning and temperature-centric yield management.

Table 2: Comparative performance with RMSE, MAE, MAPE, and R^2

Model	*RMSE (Yield prediction error in Tons)*	*MAE (Yield prediction error in Tons)*	*MAPE (Yield prediction error in Tons)*	*R2*
Hybrid MLR-LSTM (2024) [16]	0.0804	0.0667	0.5298	0.8975
CATSEM	0.0598	0.0299	0.21784	0.9499

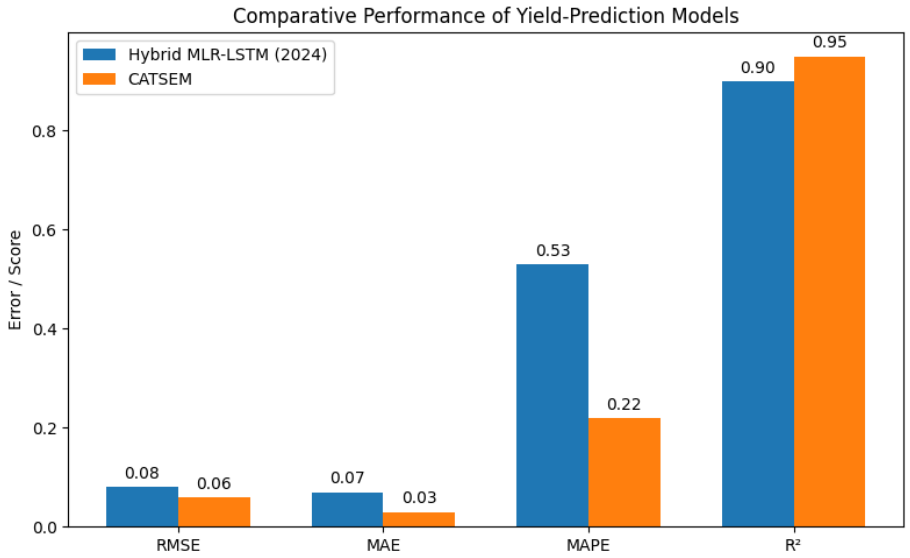

Figure 2: Comparative performance of yield-prediction models

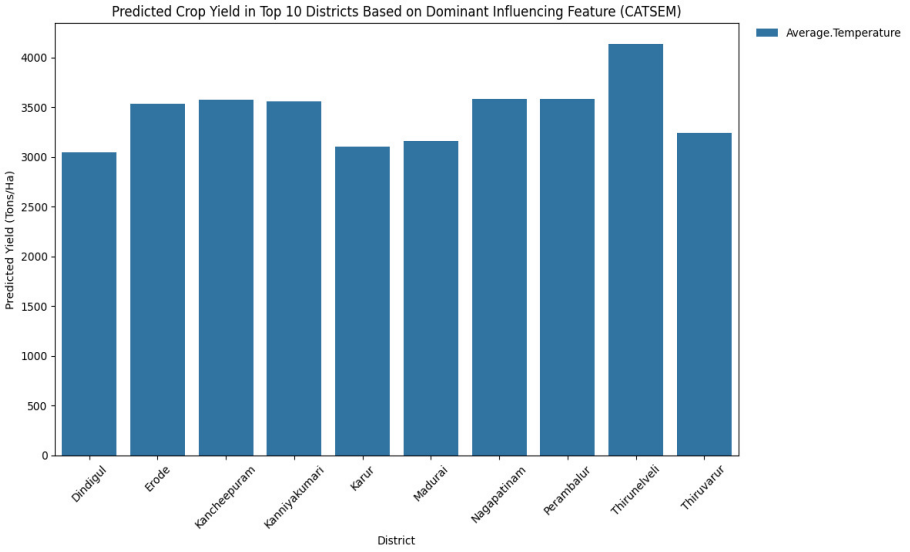

Figure 3: SHAP feature importance for CATSEM ensemble across climatic and irrigation attributes.

Conclusion

The CATSEM integrates multi-resolution signal decomposition and adaptive ensemble fusion to address the limitations of static regression models in agricultural yield forecasting. By combining DWT, EWMA, and Kalman Filtering, the framework successfully captures both large-scale seasonal trends and short-term fluctuations in climatic variables while eliminating stochastic sensor noise. The refined temporal features enhance the reliability of model inputs and ensure that the learning process focuses on physically consistent climatic dynamics rather than random perturbations. Empirical results validated on the Tamil Nadu paddy yield dataset (1986–2014) confirm CATSEM's superior accuracy relative to baseline ensemble systems. The model achieved a R^2 *of 0.9499, RMSE of 0.0598, MAE of 0.0299, and MAPE of 0.2178*, outperforming both the RFVarImp and SWERM frameworks. These improvements represent measurable gains in both precision and robustness, directly attributable to the model's temporal conditioning and adaptive stacking mechanism. The multi-stage filtering pipeline also enhances stability across non-stationary climatic sequences, ensuring resilience under varying monsoon patterns and regional anomalies.

Interpretability analyses using SHAP established that average temperature consistently emerges as the dominant variable influencing yield outcomes, followed

by rainfall and solar radiation. The interpretability further revealed that CATSEM balances long-term climatic trends and short-term fluctuations through distinct contributions from LSTM, XGBoost, and LightGBM learners. This alignment between statistical importance and physical causality underscores the scientific transparency of the model, bridging predictive performance with actionable agricultural insight. The integration of temporal intelligence, ensemble adaptability, and explainable learning positions CATSEM as a scalable decision-support framework for climate-informed agricultural management. The model's modular design allows future extensions to incorporate satellite-based vegetation indices, soil moisture sensors, or spatial autocorrelation modules. By fusing data-driven modeling with interpretable climatic reasoning, CATSEM demonstrates a significant methodological advancement over prior static ensemble systems, providing a foundation for region-specific yield forecasting, adaptive irrigation scheduling, and precision agriculture under evolving climatic conditions.

Limitations include sensitivity to station-level data quality and missingness, linear–Gaussian assumptions in Kalman filtering, restricted exogenous feature coverage (e.g., soil moisture, NDVI), absence of explicit spatial autocorrelation modeling, and potential overfitting in data-scarce districts. Future work will incorporate graph-based spatial modules (GWR/GNN), multi-source remote-sensing and soil covariates, nonlinear state-space filters, stream-data learning with drift detection, and conformal or Bayesian uncertainty quantification to enable cross-region transferability.

Acknowledgement

We thank the Department of Science and Technology, Government of India, for providing support through the Fund for Improvement of S&T Infrastructure in Universities and Higher Educational Institutions (FIST) program (Grant No.SR/FIST/College-/2020/943).

References

Abdel-salam, Mahmoud, Neeraj Kumar, and Shubham Mahajan. (2024). A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Computing and Applications. 36, no. 33, 20723-20750.

Ayiah-Mensah, F., Bosson-Amedenu, S., Baah, E. M., & Addor, J. A. (2025). Advancements in seasonal rainfall forecasting: A seasonal auto-regressive integrated moving average model with outlier adjustments for Ghana's Western Region. *Scientific African, 28*, e02632.

Das, P., Sachindra, D. A., & Chanda, K. (2022). Machine learning-based rainfall forecasting with multiple non-linear feature selection algorithms. *Water Resources Management, 36*(15), 6043–6071.

Huang, Yanbo, & Zhang, Q. (2021). Modeling of crop production systems and system characterization. In *Agricultural cybernetics* (pp. 75–130). Cham: Springer International Publishing.

Hussain, S., Huang, J., Huang, J., Ahmad, S., Nanda, S., Anwar, S., Shakoor, A. (2020). Rice production under climate change: Adaptations and mitigating strategies. In *Environment, climate, plant and vegetation growth* (pp. 659–686). Cham: Springer International Publishing.

Jabed, M. A., & Murad, M. A. A. (2024). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon, 10*(24), e2637.

Mercy, R., Beena, T. L. A., & Gopal, P. S. M. (2025). Spatially weighted ensemble regression model (SWERM) for crop yield prediction. *Indian Journal of Natural Sciences, 16*.

Mohyuddin, G., Khan, M. A., Haseeb, A., Mahpara, S., Waseem, M., & Saleh, A. M. (2024). Evaluation of machine learning approaches for precision farming in smart agriculture system: A comprehensive review. *IEEE Access, 12*, 60155–60184.

Nikhil, U. V., Pandiyan, A. M., Raja, S. P., & Stamenkovic, Z. (2024). Machine learning-based crop yield prediction in south india: performance analysis of various models. *Computers*, *13*(6), 137.

Panigrahi, B., Kathala, K. C. R., & Sujatha, M. (2023). A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Procedia Computer Science, 218*, 2684–2693.

Park, Y., Li, B., & Li, Y. (2025). Deep spatial neural net models with functional predictors: Application in large-scale crop yield prediction. *arXiv preprint* arXiv:2506.13017.

Ramesh, V., & Kumaresan, P. (2024, June). Crop Yield Predictive Model for Tamil Nadu Using Statistical Techniques. In *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0* (pp. 1-9). IEEE.

Sánchez, J. C. M., Acosta Mesa, H. G., Trueba Espinosa, A., Ruiz Castilla, S., & García Lamont, F. (2025). Improving wheat yield prediction through variable selection using support vector regression, random forest, and extreme gradient boosting. *Smart Agricultural Technology*, 100791.

Sathya, P., & Gnanasekaran, P. (2023). Paddy yield prediction in Tamilnadu Delta Region using MLR-LSTM model. *Applied Artificial Intelligence*, *37*(1), 565-583.

Thangavel, C., & Sakthipriya, D. (2024). Machine Learning Ensemble Classifiers for Feature Selection in Rice Cultivars. *Applied Artificial Intelligence*, *38*(1), 2394734.

Wang, D., Sun, T., Li, Y., Zhang, H., Li, Z., Liu, S., Dong, Q., & Li, Y. (2025). Quantifying the spatiotemporal response of winter wheat yield to climate change in Henan Province via APSIM simulations. *Agriculture, 15*(19), 2059.

Wang, M., & Li, T. (2025). Pest and disease prediction and management for sugarcane using a hybrid autoregressive integrated moving average–long short-term memory model. *Agriculture, 15*(5), 500.

Zhao, F., Zhang, Q., Wang, H., & Zhang, K. (2025). Year patterns of climate impact models' performance: Long-term simulation of rainfed spring wheat production using five crop models under various climate patterns. *Agricultural Water Management, 318*, 109704.

Zhao, X., Tang, W., Liu, Q., Cao, H., & Chen, F. (2025). Impact of agricultural industry transformation based on deep learning model evaluation and metaheuristic algorithms under dual carbon strategy. *Scientific Reports, 15*(1), 27929.