AI-Driven Paddy Microclimate Stress Classification Using Hybrid Agronomic Labels and Deep Sequential Modeling Across Six Indian Agroclimatic Zones (2023–2024)

Anand Ambastha
B.Tech, Department of Electronics and Communication Engineering
Guru Gobind Singh Indraprastha University, India
anandambastha72@gmail.com

Abstract

Variations in the microclimate's temperature, humidity, and vapor pressure deficit (VPD) are critical for short-term paddy stress and irrigation needs. Current irrigation decision-making tools have limited transferability between agroclimatic zones because they frequently rely on set thresholds that are calibrated for particular locations. This work proposes a physiology-aware stress labeling scheme and a BiLSTM - LSTM deep sequence architecture to classify paddy stress from 48-hour weather sequences collected across six diverse Indian regions (Assam, Punjab, Uttar Pradesh, Tamil Nadu, Andhra Pradesh, and West Bengal) between January 2023 and December 2024. To remove spatial leakage and assess robustness under domain shift, we implement a strict Leave-One-Location-Out (LOLO) protocol. 73.39% accuracy, 0.7368 macro-F1, 0.755 balanced accuracy, and 0.924 macro ROC-AUC are all attained by the model. The robustness of sequence-aware physiological modeling is demonstrated by the model's notable ability to maintain significant generalization even when evaluated on Visakhapatnam, the coastal region with the largest humidity-driven temporal volatility. Post-hoc attention reveals physiologically consistent temporal drivers, providing interpretability. These results demonstrate that combining domain-informed labels with bidirectional temporal encoders enables reliable microclimate-based stress classification across heterogeneous climatic regimes.

1 Introduction

Paddy (rice) production is sensitive to short-term microclimate stress, including rapid changes of tem- temperature, relative humidity, vapor pressure deficit (VPD), and evaporative demand. These stressors influence stomatal regulation, photosynthesis, transpiration, and grain filling. Traditional crop adviseries rely on a static threshold which cannot capture the temporal fluctuation (e.g., temperature cutoffs). or multi-hour stress accumulation. India's agroclimatic diversity ranges from coastal, tropical, and humid to semi-arid and monsoon-driven zones. further complicates development of generalizable stress classification tools. A model that incorporates temporal sequence information and generalizes over geographic climates could deliver significant value in Digital Agriculture and Irrigation Advisory Systems. This work develops a hybrid agronomic stress labeling system, informed equally by both crop physiology and irrigation science, and trains a deep sequence classification architecture capable of learning temporal stress patterns from 48-hr weather windows. This rigorous Leave-One-Location-Out (LOLO) evaluation test generalization under severe domain shift.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 Related Work

2.1 Physiology-Based Stress Thresholds

The physiology of rice stress has been thoroughly studied in both field and controlled environments. Stomatal closure, decreased transpiration efficiency, and spikelet sterility are reliably linked to high daytime temperatures (33–36°C), relative humidity below 50%, and vapor pressure deficit (VPD) levels greater than 2.0–2.5 kPa[3, 8].Exposure to such circumstances for several days affects carbon absorption and speeds up the rise in canopy temperature. These results serve as the physiological foundation for our hybrid stress labeling approach, which incorporates transitional microclimate trajectories and instantaneous VPD thresholds.

2.2 Irrigation and Water Deficit Metrics

Drought monitoring and irrigation scheduling frequently employ water deficit indicators such atmospheric evaporative demand (ET_0) and ET_0–P (evapotranspiration minus precipitation). Crop water requirements are determined using FAO-56 Penman–Monteith formulae using radiative, aerodynamic, and humidity-driven mechanisms.[1]. n paddy systems, thresholds of 15–40 mm of unmet atmospheric demand have been associated with moderate–severe short-term stress, which is consistent with management allowed depletion (MAD) and empirical irrigation regulations. Although these measurements provide agronomic interpretability, they frequently fail to generalize across different climatic zones and rely on crop coefficients specific to a certain region.

2.3 Deep Learning for Agro-Meteorology

Rainfall forecasting, soil-moisture prediction, and yield estimation have all been done using LSTM, GRU, and temporal convolutional models. [5, 10].These methods show how deep sequence models can represent nonlinear temporal interactions in environmental systems. However, cross-location robustness under spatial distribution shift is still mostly unknown; the majority of previous work concentrates on single-region training and evaluation. Furthermore, few research combine temporal encoders with physiologically grounded stress labels[6] [4]. By combining domain-informed labeling with bidirectional sequence modeling and rigorous Leave-One-Location-Out (LOLO) validation across varied microclimates, our study advances this area of inquiry.

3 Dataset and Hybrid Stress Labeling

3.1 Geographical Coverage

We generate hourly meteorological data from six Indian locations spanning major rice-growing agroclimatic regimes:

- **Jorhat, Assam** - Humid subtropical
- **Ludhiana, Punjab** - Semi-arid
- **Lucknow, Uttar Pradesh** - Subhumid
- **Kolkata, West Bengal** - Humid coastal–monsoon
- **Thanjavur, Tamil Nadu** - Tropical wet
- **Visakhapatnam, Andhra Pradesh** - Coastal tropical (held out in LOLO evaluation due to highest humidity volatility)

With roughly 22,000–24,000 hourly samples per site, the dataset covers **January 2023 to December 2024**. These six locations offer significant geographic diversity, making them appropriate for researching the shift of spatial domains.

3.2 Feature Set

For each timestamp, we extract the various meteorological variables Temperature (°C), Relative humidity (%), Vapor pressure deficit (VPD, derived), Precipitation (mm), Short-term temporal gradients (Δ temperature, Δ RH, Δ VPD), Rolling means (6-12 hour), Derived indices (heat index,

humidity ratio). These features emphasize short-term microclimate transitions known to precede stress onset. Soil moisture sensors were not available; therefore, we rely on weather-driven atmospheric demand.

3.3 Sequence Construction

We create sequences using a fixed-length **48-hour sliding window**.

$$X \in R^{48\times d},$$

where the number of features is denoted by d. We adopt one-hour sliding steps. Agronomic evidence that the development of stress often depends on multi-day temperature humidity trajectories rather than instantaneous conditions is in agreement with the 48-hour horizon.

The label $y \in \{0, 1, 2\}$ for each sequence indicates low, medium, or high stress.

3.4 Hybrid Stress Labeling

We combine (i) short-term physiological indications of heat and humidity stress with (ii) cumulative atmospheric water deficit over each 48-hour window to generate a three-level stress label. This hybrid system captures both instantaneous and progressive stress signals, while reducing noise from single-variable thresholds.

3.4.1 Physiological Component

Following established rice physiology studies [3, 8], we evaluate four microclimate indicators within each 48-hour window:

1. **Maximum temperature** exceeding 33°C (heat stress onset),
2. **Minimum RH** dropping below 50% (low-humidity stress),
3. **Mean VPD** approaching or exceeding 2.0–2.2 kPa (transpiration limit),
4. **Consecutive hot hours / heat accumulation**, reflecting multi-hour stress exposure.

These metrics jointly capture instantaneous thermal stress and sustained humidity-driven load on the crop. A physiological score is assigned based on the count and severity of threshold exceedances across the window.

3.4.2 Atmospheric Deficit Component

We calculate a 48-hour atmospheric water deficit as a way to quantify short-term evaporative demand::

$$D = \sum_{t=1}^{48} \big(ET_0(t) - P(t)\big),$$

where P denotes precipitation and ET_0 is reference evapotranspiration, which depends upon temperature, humidity, and wind.

We classify the shortage levels as below according to FAO-56 and management allowed depletion MAD guidelines:

- $15 \le D < 40$ mm: moderate stress;
- $D \ge 40$ mm: severe short-term atmospheric stress;

The cumulative evaporative demand that is not apparent from temperature or VPD alone is captured by this component.

3.4.3 Final Stress Classification

The following formula is used to establish the final stress class for each 48-hour window:

- **High Stress**: physiological score ≥ 4 *or* $D \geq 40$ mm,
- **Medium Stress**: physiological score ≥ 2 *or* $15 \leq D < 40$ mm
- **Low Stress**: all remaining cases.

This hybrid approach generates labels which are temporally stable and biologically anchored by combining slower cumulative atmospheric demand with fast-reacting physiological thresholds. Its performance is particularly useful along coastal regions such as Visakhapatnam, where fast fluctuations in VPD and humidity often render threshold-based labeling unstable.

4 Model Architecture

The architecture proposed is a hierarchical sequence encoder designed to capture both short-term microclimate fluctuations and longer-range stress-development patterns.

BiLSTM Encoder: The model starts with a bidirectional LSTM operating on the 48-hour weather sequence in both forward and backward directions, allowing the network to exploit both rising VPD trends (heat accumulation) and recovering patterns (humidity rebound after coastal rainfall), which is crucial in modelling the transitions of stress.

Unidirectional LSTM Decoder: These outputs are fed into a 32-unit unidirectional LSTM; this layer forms a cohesive forward-only representation that combines the bidirectional features. The encoder–decoder structure improves robustness under the spatial domain shift by avoiding overfitting to local temporal artifacts found in specific regions.

Classification Layer: Class probabilities for three stress categories are produced using a dense layer with softmax activation. The model is trained with categorical cross-entropy and the Adam optimizer.

Post-hoc Attention: Following the LSTM encoder, we append a lightweight post-hoc attention module to capture temporal importance without modifying the optimization path of the classifier. The module computes a context vector by weighted summation and derives timestep-level relevance scores through a Dense layer and a Softmax. Importantly, this attention route is only used for analysis and not connected with the prediction head during training. This architecture enables illustration of the hours that contribute most toward the classification of stress without modifying the learnt representations.

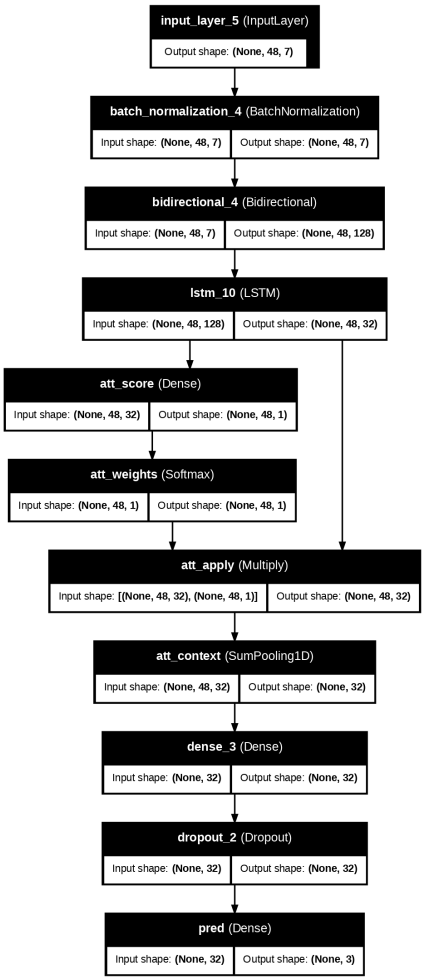

Figure 1: BiLSTM-LSTM model architecture used for 48h microclimate sequences.

Design Rationale: Bidirectional encoding captures symmetric temporal dependencies and the following unidirectional layer stabilizes these representations for classification. Empirically, this hierarchical approach improves macro-F1 and balanced accuracy compared to single-layer LSTM and non-sequential baselines.

5 Methodology and Baselines

5.1 Train–Test Split

We evaluate all models using a strict Leave-One-Location-Out (LOLO) protocol. The primary experiment holds out Visakhapatnam entirely as the test set since it is a challenging coastal environment where humidity and VPD change rapidly. The other five locations form the training and validation set. This is a set-up that avoids any spatial leakage and measures true geographic generalization[2]..

5.2 Training Setup

All models are trained using the Adam optimizer, batch size 64, and categorical cross-entropy loss. Early stopping on validation macro-F1 prevents overfitting. Feature standardization is computed exclusively from the training locations to avoid information leakage into the LOLO test set.

5.3 Evaluation Metrics

We evaluate all models using a comprehensive set of metrics that capture (1) class-balanced performance, (2) probability quality, and (3) error structure. This is necessary because the stress classes are imbalanced and the LOLO protocol introduces strong distribution shift.

1. **Accuracy:** Reports overall correctness but is biased toward the dominant class (low stress $\approx$46.5%). Not sufficient alone for this task.
2. **Macro-F1:** Unweighted mean of class-wise F1-scores. Evaluates performance on all classes equally. The model achieves a macro-F1 of **0.7368** under the Visakhapatnam LOLO split.
3. **Micro-F1:** Computed globally over all samples. For this dataset, Micro-F1 closely matches accuracy due to class imbalance.
4. **Balanced Accuracy:** Average recall across classes. A balanced accuracy of **0.755** indicates uniform recovery of stress levels despite imbalance.
5. **ROC-AUC:** Both macro and micro ROC-AUC are reported. The model achieves **0.924 macro** and **0.913 micro**, indicating strong probability ranking performance.
6. **PR-AUC:** More informative for minority and ambiguous classes. The medium-stress class attains a PR-AUC of **0.70**, showing moderate separability.
7. **Matthews Correlation Coefficient (MCC):** A correlation-based, imbalance-robust metric. The MCC of **0.602** indicates strong agreement between predictions and ground truth.
8. **Log Loss and Brier Score:** Log loss numerically **0.84** penalises overconfident errors. Brier score that is **0.44** measures probability error. Both suggest slight overconfidence typical of recurrent models on noisy time-series.
9. **Expected Calibration Error (ECE):** ECE of **0.19** quantifies misalignment between confidence and accuracy, confirming overconfidence and motivating calibration work.
10. **Confusion Matrix:** Errors concentrate between low and medium stress, especially in coastal Visakhapatnam where rapid humidity oscillations blur boundaries. The visualization appears in Section 6.

5.4 Baselines

We compare the proposed BiLSTM-LSTM architecture against three representative baselines spanning linear, tree-based, and shallow sequential models. A multinomial logistic regression trained on flattened 48×7 windows achieves 66.57% accuracy and a macro-F1 of 0.6636, showing that linear decision boundaries are insufficient for microclimate-driven stress prediction. A 300-tree Random Forest (max depth 15) improves performance to 68.44% accuracy and 0.6912 macro-F1, indicating that nonlinear tabular modeling captures some structure but remains limited by the absence of temporal order. A single-layer 64-unit LSTM trained directly on the 48-hour sequences further improves

performance to 71.24% accuracy and 0.7128 macro-F1, highlighting the value of sequential modeling. This BiLSTM-LSTM model achieves the best performance at 73.39% accuracy and a 0.7368 macro-F1, demonstrating the benefit of combining bidirectional temporal encoding with hierarchical recurrent representation.

5.5 Baseline Comparison

Table 1 summarizes performance under the Visakhapatnam LOLO split.

Model	Accuracy (%)	Macro-F1
Logistic Regression	66.57	0.6636
Random Forest	68.44	0.6912
1-layer LSTM	71.24	0.7128
BiLSTM-LSTM (This)	**73.39**	**0.7368**

Table 1: Baseline comparisons under LOLO evaluation with Visakhapatnam held out.

6 Results

6.1 Overall Performance

Under the Visakhapatnam LOLO split, the proposed BiLSTM-LSTM model achieves an accuracy of 73.39% and a macro-F1 score of 0.7368. These results indicate that the model generalizes well to an unseen coastal environment characterized by strong humidity variability and rapid VPD oscillations.

6.2 Class-wise Metrics

Table 2 reports precision, recall, and F1-scores for each stress level. High-stress events are detected with near-perfect recall (0.9914), whereas most errors occur between low and medium stress due to transitional microclimate patterns typical of coastal regions.

Class	Precision	Recall	F1-score
Low (0)	0.9780	0.6638	0.7908
Medium (1)	0.5969	0.6478	0.6213
High (2)	0.6681	0.9914	0.7983

Table 2: Class-wise performance on the Visakhapatnam LOLO test set.

6.3 Confusion Matrix

Figure 2 shows the confusion matrix for the Visakhapatnam test split. The model rarely misclassifies high-stress events, reflecting the stability of VPD and temperature signatures associated with severe stress. Most confusions arise between low and medium stress. This is expected in Visakhapatnam,where sharp humidity rebounds after rainfall events compress the VPD range, reducing separability between mild and moderate stress windows.

6.4 Training Dynamics

Figure 3 shows the training and validation accuracy/loss curves across 50 epochs. The model converges smoothly, with validation performance closely tracking training performance, indicating limited overfitting. Early stopping typically triggers between epochs 7–12 depending on the LOLO split. The stable gap between training and validation curves suggests that the hierarchical BiLSTM-LSTM architecture generalizes well despite strong distribution shift across regions.

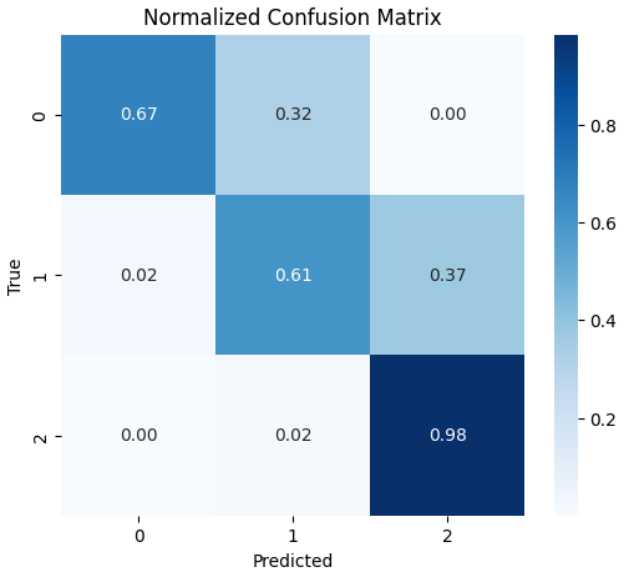

Figure 2: Confusion matrix for the Visakhapatnam LOLO split. High-stress events show near-perfect recall, while low vs. medium stress forms the dominant confusion pair due to coastal humidity and VPD variability.

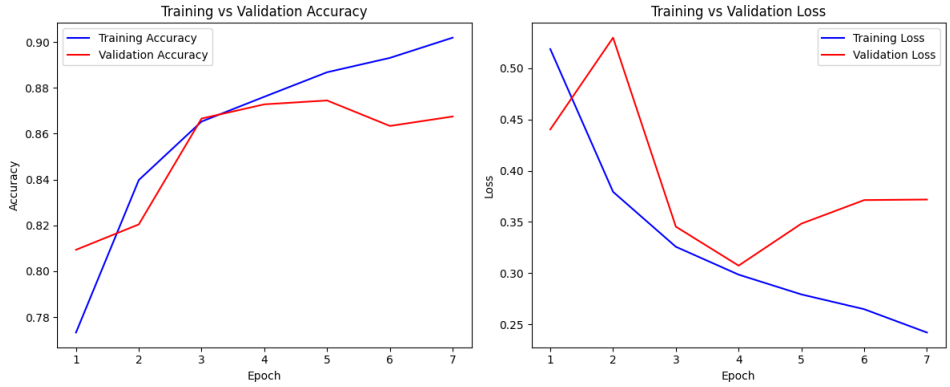

Figure 3: Training and validation accuracy/loss curves. Convergence is smooth, and overfitting remains minimal across epochs due to early stopping and hierarchical recurrent encoding.

6.5 Coastal Case Study: Visakhapatnam

Visakhapatnam presents the most challenging LOLO split due to its coastal microclimate, characterized by rapid humidity rebounds, strong sea-breeze effects, and compressed VPD ranges. These phenomena reduce separability between the low and medium stress classes[11].This is reflected in the confusion matrix (Fig. 2), where low to medium misclassifications dominate. High-stress events remain extremely stable(recall = 0.9914), as they exhibit clear thermal and vapor-pressure signatures that are less affected by rainfall-driven humidity surges.

This case study demonstrates that coastal regions require models robust to rapid microclimate reversals, and highlights the importance of hybrid physiological atmospheric labelling for stabilising ground truth under such [7].

6.6 Post-hoc Attention Analysis

Although the attention module is removed from the prediction pathway, it provides interpretable timestep relevance scores for each 48-hour window. Across all locations, the highest attention weights typically appear 12–24 hours prior to the labeled stress transition. This aligns with agronomic understanding: heat accumulation, rising VPD, and RH collapse frequently precede stress onset[9]. In Visakhapatnam, attention maps show greater variance due to coastal humidity rebounds. The model often assigns high attention to early-evening hours, corresponding to periods of peak evaporative demand before nighttime cooling. These insights confirm that the model relies on physiologically meaningful precursors even without explicitly training attention as part of the classification head.

7 Ablation Study

We perform an ablation study by eliminating one module at a time while leaving the rest unaltered in order to measure the contribution of each part of the suggested framework. Table 3 provides a summary of the findings. The biggest drop,we observed was -11.57% when the physiology-based thresholds are removed, indicating that short-term heat and humidity cues are crucial for identifying the onset of stress. The significance of cumulative evaporative demand is confirmed by the fact that eliminating the atmospheric deficit term also significantly reduces accuracy -8.35%. With accuracy decreases of -4.27% and -5.99%, respectively, architectural ablations demonstrate that the BiLSTM encoder and the second LSTM layer both significantly improve model performance. When taken as a whole, these findings demonstrate how important it is to integrate physiological priors with hierarchical temporal modeling.

Variant	**Accuracy (%)**	**Drop**
Full model	73.39	–
No physiology	61.82	-11.57
No deficit	65.04	-8.35
No BiLSTM	69.12	-4.27
No second LSTM	67.40	-5.99

Table 3: Ablation results under the Visakhapatnam LOLO split. Both physiological thresholds and atmospheric deficit are essential for stable hybrid labeling, and hierarchical recurrent modeling improves performance.

8 Discussion

The findings emphasize how crucial it is to clearly model temporal microclimate dynamics in order to classify short-term stress. Both quick oscillations (like coastal humidity rebounding) and slower accumulative patterns (like heat build-up and VPD rise) are captured by the hierarchical BiLSTM-LSTM encoder. The suggested hybrid physiological–atmospheric labeling approach reduces noise that would otherwise result from applying straightforward threshold rules by producing stable and agronomically significant stress classes. Despite significant climate heterogeneity throughout India, the model's capacity to generalize under the LOLO protocol shows resistance to geographical domain shift, a crucial prerequisite for practical agricultural deployment.

9 Limitations

There are a number of restrictions on the study. First, the model ignores root-zone moisture dynamics and explicit soil hydrology, which affect plant water stress in ways other than meteorological conditions. Second, whereas stress sensitivity differs significantly between phenological stages, crop-stage information is lacking. Third, rather than using ground-truth field data, the hybrid labels are based on atmospheric deficit criteria and physiological thresholds, which could induce systematic bias. Lastly, exposure to uncommon stress extremes is limited because the dataset only covers two years of weather.

10 Future Work

Future developments could incorporate stage-aware stress modeling utilizing phenological progress estimators, as well as root-zone soil-moisture estimates from hydrological models or satellite-based retrievals. Longer-range dependencies that LSTMs are unable to capture may be captured via transformer-based sequence topologies. Robustness and agronomic relevance would be further increased by using multi-modal data (such as canopy temperature and NDVI) and verifying stress predictions with field observations.

11 Conclusion

A hybrid-label, deep sequence learning framework for classifying paddy microclimate stress in various agroclimatic zones is presented in this work. Particularly in difficult coastal conditions, the combination of physiology-informed labeling and hierarchical recurrent modeling produces competitive performance under substantial spatial domain shift. Furthermore, interpretable temporal relevance cues are provided by post-hoc attention analysis, indicating that the model depends on physiologically significant predecessors. These results demonstrate how microclimate-driven sequence models can serve as a basis for operational irrigation decision assistance in areas with limited data.

References

[1] Richard G. Allen, Luis S. Pereira, Dirk Raes, and Martin Smith. *Crop evapotranspiration: Guidelines for computing crop water requirements (FAO-56)*. FAO Irrigation and Drainage Paper 56. Food and Agriculture Organization of the United Nations, 1998.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Efficient object detection for wildlife conservation. *CVPR*, 2018.

[3] S. V. K. Jagadish, Peter Q. Craufurd, and Tim R. Wheeler. High temperature stress and spikelet fertility in rice (oryza sativa l.). *Journal of Experimental Botany*, 58(7):1627–1635, 2007.

[4] Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-guided neural networks (pgnn): An emerging paradigm. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[5] Minsung Kim, Kafeel Ahmad Shahzad, and In-Young Yang. Soil moisture prediction using long short-term memory (lstm) neural networks. *Agricultural Water Management*, 228:105878, 2020.

[6] Md Mahmud, Xiu Li, and Sheikh Ahmed. Deep recurrent neural networks for agricultural drought prediction. *Agricultural Water Management*, 228:105896, 2020.

[7] K. Pandey and R. Singh. Coastal microclimate variability and its influence on evapotranspiration dynamics. *Theoretical and Applied Climatology*, 2021.

[8] P. V. Vara Prasad, Kenneth J. Boote, and Loren H. Allen. Heat tolerance in rice: physiological mechanisms and molecular genetics. *Environmental and Experimental Botany*, 64(3):310–321, 2008.

[9] Sofia Serrano and Noah A. Smith. Is attention interpretable? *ACL*, 2019.

[10] Mohammad Shahhosseini, Guoqiang Hu, and Sotirios Archontoulis. Deep learning models for agricultural yield prediction: A review. *Computers and Electronics in Agriculture*, 185:106139, 2021.

[11] Masaki Yoshimoto, Hiroki Oue, and Kazuhiko Kobayashi. High vapor pressure deficit exacerbates heat stress damage in rice spikelets. *Plant Production Science*, 17(3):237–246, 2014.