Journal of Agrometeorology

(A publication of Association of Agrometeorologists)

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 28 (2) : 237-248 (June - 2026)
https://doi.org/10.54386/jam.v28i2.3400
https://journal.agrimetassociation.org/index.php/jam

Research paper

Performance Evaluation of Regional Heatwave Prediction Using Statistical and Deep Learning Models

SULOCHANA DEVI[1, 2*] and RADHIKA KOTECHA[1]

[1]*Department of Information Technology, K J Somaiya Institute of Technology, University of Mumbai, Mumbai, India*
[2]*Department of Information Technology, Xavier Institute of Engineering, University of Mumbai, Mumbai, India*
[*]*Corresponding author's email: bishnoisulochana@gmail.com*

ABSTRACT

Rising temperatures are increasing the frequency and impact of heat extreme events in India. So, this is increasing the need for early warning and climate-risk management. Heat risks and behaviors vary across regions, so national model can miss local temperature dynamics. We present a season-aware, region-wise framework to forecast daily maximum temperature and identify heatwave days in IMD defined Central India region. In Central India's severe pre-monsoon heat is recurrent and closely linked to crop-weather stress, irrigation demand, and outdoor labor exposure. Heatwaves are first categorized at grid-cell level using operational threshold-and -departure rules. Then heatwaves are aggregated to a region-day label using a spatial coverage threshold and minimum 2-day duration. In this paper, we benchmark conventional statistical time-series models against recurrent sequence models under same splits and evaluation window using national meteorological temperature data. Performance is evaluated with temperature-error metrics and event-based measures for heatwave-day detection. Results show that recurrent models give the strongest overall skill, with the LSTM model delivering the promising improvements. Seasonal statistical modeling improves over non-seasonal baselines by capturing the seasonal cycle. This framework offers a compact regional benchmark and practical guidance for region specific early warning systems. This work supports agrometeorological advisories for farm operations, water planning and heat-stress risk management during extreme heat.

Keywords: Heatwave Prediction, Seasonal Segmentation, Statistical Methods, Regional Segmentation, Climate Adaptation, Sustainability, Weather-Based Decision Support

Heatwaves are long spells of abnormally high temperature usually experienced over a number of days that cause stress beyond what is normal for the season and place. Extreme heat is increasingly becoming more common and severe in most parts of the world due to climate warming and so the need for early warning is growing (Perkins-Kirkpatrick & Lewis, 2020; Davis *et al.,* 2021). In India, heat waves are not only a weather issue but are also affecting the health outcomes, working capacity, electricity requirements, and other services that are very sensitive to peak heat (Fischer *et al.,* 2021; Somanathan *et al.,* 2021). Evidence from India shows clear connections between heatwave conditions and the risk of mortality. This reinforces the need for prediction methods that are reliable and easy to apply in practice (Davis *et al.,* 2021; Somanathan *et al.,* 2021).

Extreme heat is closely associated with crop stress, irrigation demand and when outdoor farm work can be undertaken (Heino *et al.,* 2023). A location-specific forecasting of agrometeorological advisories would be enhanced by region-wise forecasting of heatwave-days, to provide earlier, location-specific warnings during hot days.

The main problem is that the heat hazard does not spread evenly on the national level. The same temperature value may be associated with various degrees of risks in various regions due to background climatic conditions, housing and access to air conditioning. This makes region-based early warning significant as compared to a single nation-based model that averages the very dissimilar regimes (Ratnam *et al.,* 2016; Mandal *et al.,* 2025). Central India is an important target as it is part of an inland belt

Article info - DOI: *https://doi.org/10.54386/jam.v28i2.3400*
Received:24 February 2026; Accepted: 13 May 2026; Published online : 04 June 2026
Suplimentary Table S1 : https://journal.agrimetassociation.org/index.php/jam/article/view/3400/1814

where heat could accumulate and linger particularly towards the pre-monsoon period and it is recurrently known to undergo harsh hot temperatures. Regional focus also reduces mixed controls on climates which emerge when both coastal and inland regions are modeled together which proves useful when the aim is to convert climate temperature behavior into operational heatwave-day decision-making (Ganeshi *et al.,* 2023; Cho *et al.,* 2014).

The study develops a compact, season-aware baseline of predicting heat waves in the region of Central India based on the temperature of the national meteorological agency. The first stage is the identification of heatwave conditions at the grid level based on an operational level threshold-and-departure rule and converted into a regional heatwave-day label based on a spatial coverage requirement and a minimum-duration requirement. In order to establish a clear benchmark across model families, classical statistical time-series models are compared to recurrent sequence models using the same data splits and evaluation windows. The temperature forecast errors are used along with event metrics of confusion matrices, which means that not only average error can be seen but also missed heatwave days and false alarms. By combining grid-based labeling with region-day evaluation and a season-aware setup, the paper provides a clear reference point for Central India and a practical basis for extending the approach to other IMD regions and increased lead times.

In this study, we use two approaches, statistical and deep learning to develop a dependable heatwave prediction model. The key contributions of this paper are given as:

1. This study provides framework with implementation to predict heatwave days using IMD temperature records.
2. This study is focused on Central India region defined by IMD. This region faces severe pre-monsoon heat repeatedly.
3. The study compares classical statistical models and sequence-based deep learning models using the same data and labelling rules. It covers both next-day T_{max} forecasting and heatwave-day prediction. Results are evaluated using temperature error and event-based heatwave metrics. A sensitivity check is used to ensure the findings do not depend on minor, reasonable changes in the labelling choices.

MATERIALS AND METHODS

Temperature regime of Central India

The study area is IMD Central India which is defined using the IMD meteorological subdivision system and is being used as a single spatial mask for the gridded analysis. The $1° \times 1°$ grid cells with the centre points within the boundary are kept after the dissolution of the 4 polygons (West Madhya Pradesh, East Madhya Pradesh, Vidarbha and Chhattisgarh). It is an inland belt. The influence of the sea is limited in this belt. Land-surface conditions and regional circulations are important factors in determining the warm-season T_{max}. This can lock multi-day hot spells in the interior region which is physically plausible and operationally relevant for an analysis based on regions. A strong annual cycle is also visible. T_{max} generally climbs through the pre-monsoon months (March–May), drops during the southwest monsoon as cloud cover and rainfall increase, and then shifts again in the post-monsoon transition; the monthly mean maps over the Central India mask capture this structure clearly as shown in Fig. 1. What tends to stand out during heat events is the spatial coherence, neighboring grid cells often warm together under the same forcing, rather than producing scattered, isolated spikes. This behavior supports the use of a region-scale mask for consistent modelling and heatwave labelling (Ratnam *et al.,* 2016; Mandal *et al.,* 2019; Das *et al.,* 2020). The seasonal cycle in Fig. 1 sets the background for heatwave formation. Since heatwave labels depend on grid-level exceedances and persistence, we also summarize how often heatwave conditions occur inside each subdivision of Central India. Table 1 shows that heatwave burden is not uniform inside Central India (2000–2024; $\theta = 0.10$). West Madhya Pradesh and Vidarbha have high heatwave-day totals and positive trends, while Chhattisgarh shows a negative trend.

These supports keeping grid-level labeling and then building a region-day label from spatial coverage instead of treating the whole region as uniform. Fig. 2 makes the year-to-year variability visible and shows that trends differ across the four subdivisions, which is important when interpreting region-level results. Heatwave days are computed using grid-level detection and aggregated using $\theta = 0.10$ with a minimum two-day duration. The subdivision-wise analysis of heatwave characteristics across Central India reveals significant spatial variability in both frequency and temporal trends. Some regions are clearly getting more heatwaves, while others are not changing much or even improving. West Madhya Pradesh shows the strongest increase, with heatwave days rising steadily over time. Vidarbha shows a similar pattern, so both of these regions are becoming more prone to extreme heat.

In the other hand, the area of East Madhya Pradesh is almost flat with no appreciable variation with time. Chhattisgarh is interesting as it indeed recorded a reduction in heatwave days. That indicates that heat waves are occurring a bit less frequently there, which may be caused by local climate or land conditions. The mean heatwave coverage further highlights spatial heterogeneity, with Vidarbha showing the highest spatial extent (0.138), followed by West Madhya Pradesh (0.125), reinforcing their vulnerability to widespread heatwave conditions. The findings strongly support using a spatial analysis framework based on a grid because aggregating at a regional level would conceal the local variations in heatwave behavior. The main baseline is based solely on T_{max}. This choice matches the target label, because the operational IMD heatwave definition is expressed through T_{max} thresholds and departures from a daily normal. Some other factors, such as soil moisture and humidity, also play an important role in impacts and persistence but they address a different question: Soil moisture deficits can influence heatwave persistence through land-atmosphere coupling (Seneviratne *et al.,* 2010; Ganeshi *et al.,* 2023; Lorenz *et al.,* 2010) while humidity mainly affects human heat stress (apparent temperature) rather than air temperature itself (Cvijanovic *et al.,* 2023; Lu & Romps, 2022). For this paper, we keep a T_{max} -only design to provide a clean, reproducible benchmark aligned to the IMD definition; multi-variable extensions are left for future work.

Table 1: Subdivision-wise heatwave frequency and trend within Central India

Subdivision	Total Days	HW Days	HW Events	Mean Heatwave Coverage (Fraction 0–1)	Trend (HW Days/ Decade)
West Madhya Pradesh	9132	1655	124	0.125	1.869
East Madhya Pradesh	9132	1394	102	0.115	0.062
Vidarbha	9132	1582	84	0.138	1.615
Chhattisgarh	9132	1173	97	0.075	−1.292

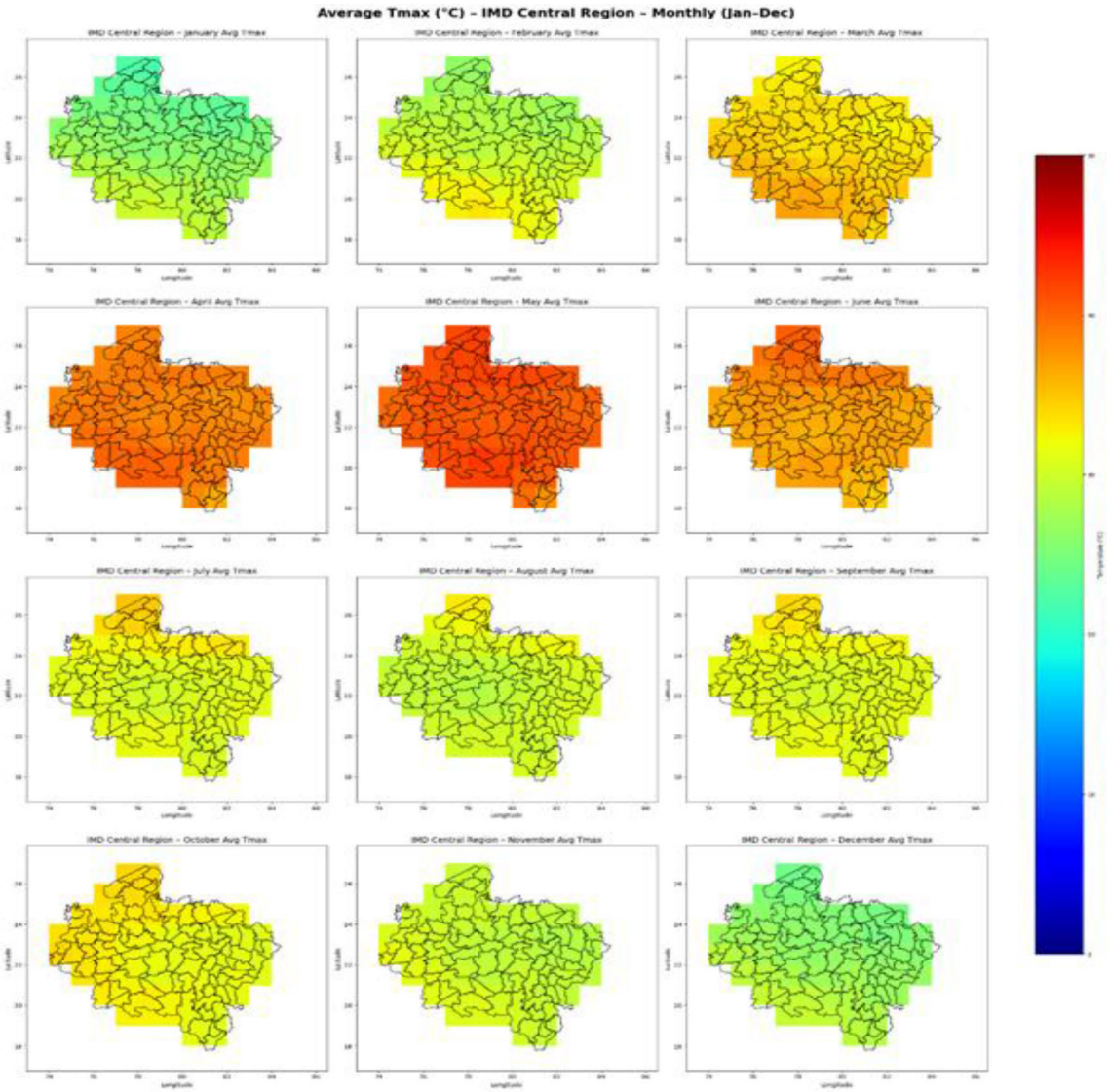

Fig. 1: Monthly mean T_{max} (°C) over IMD Central India (Jan–Dec) from the IMD 1° × 1° gridded archive, showing the pre-monsoon peak and monsoon cooling.

Proposed Approach

To predict the heatwave for the focused area under study, a detailed approach is discussed phase-wise presented in Fig. 3 with end-to-end workflow to generate heatwave predictions over IMD Central India from daily maximum temperature (T_{max}).

The approach has six phases: (i) retrieve IMD gridded T_{max}, (ii) preprocess and fill missing values, (iii) apply the Central India spatial mask, (iv) split the time series into meteorological seasons, (v) train forecasting models to predict next-day T_{max}, and (vi) convert predicted T_{max} into heatwave-day predictions using IMD-consistent rules and event persistence constraints. The modelling is deliberately focused to one observable variable (T_{max}) and it keeps the heatwave definition fully transparent and consistent with operational practice over inland India (Satyanarayana & Bhaskar Rao, 2020; Akaike, 1974).

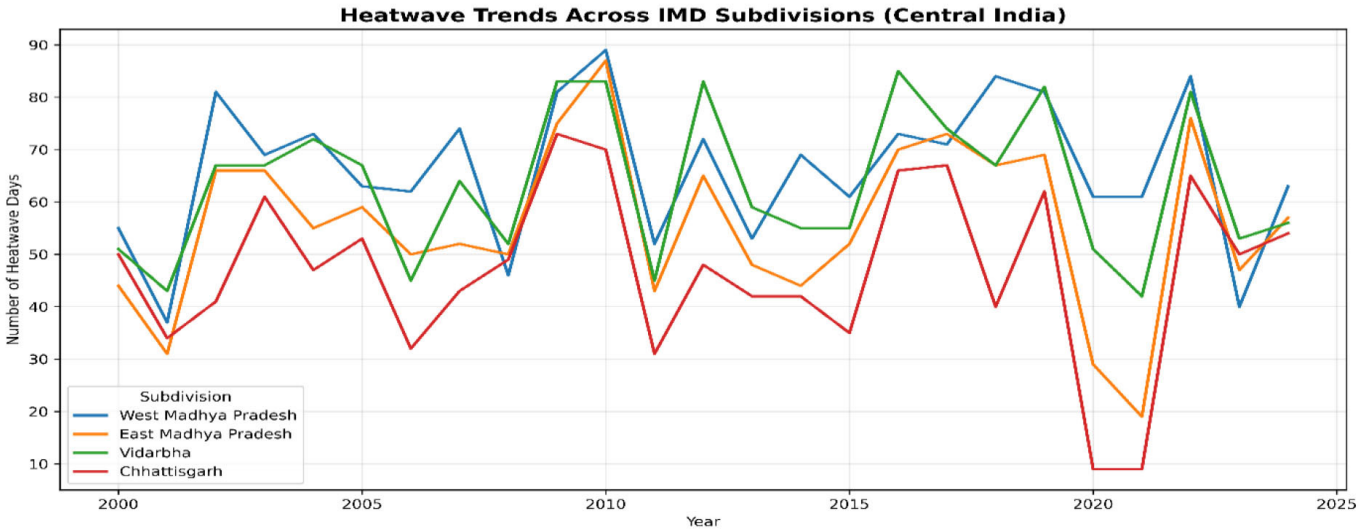

Fig. 2: Annual heatwave days (2000–2024) across the four IMD subdivisions within Central India

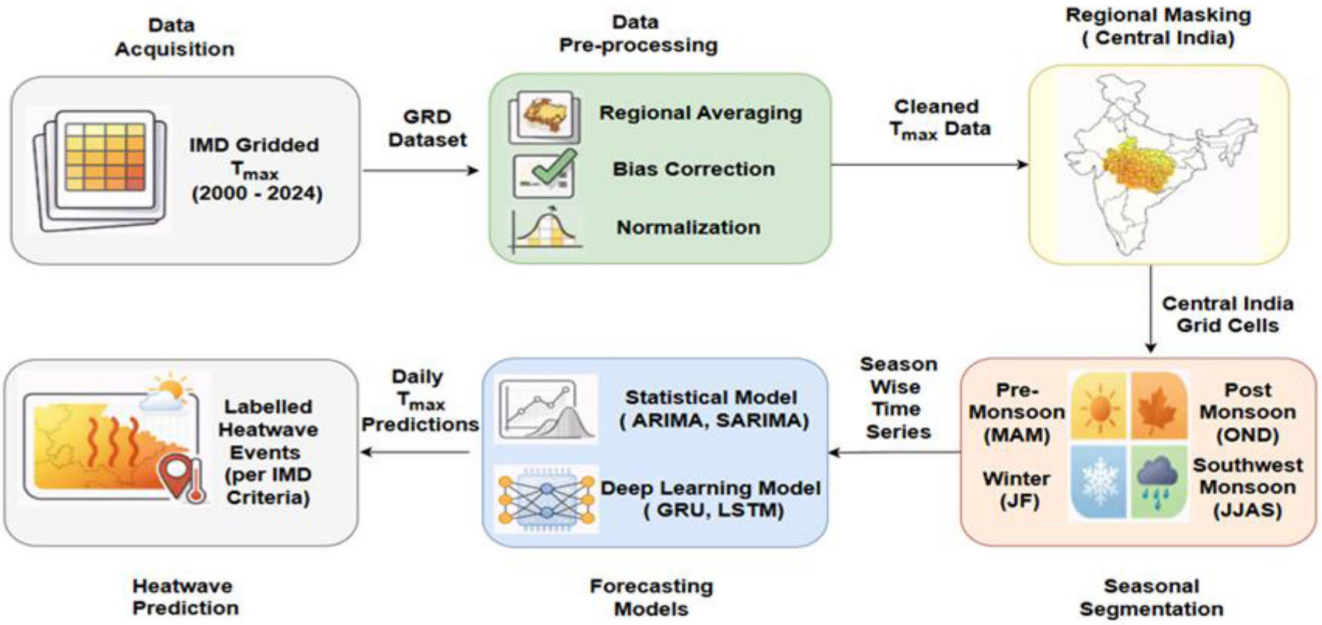

Fig. 3: Overall workflow for T_{max}-based heatwave prediction over Central India

Data Source, Resolution, And Temporal Coverage

Daily maximum temperature (T_{max}, °C) is taken from the IMD gridded daily temperature dataset at 1° × 1° resolution. We use the 2000–2024 subset. The gridded fields are derived from quality-controlled station observations and interpolated to a regular grid (Srivastava *et al.,* 2009). Data are ingested and converted into analysis-ready NetCDF/Xarray format using IMD Python tools to keep processing reproducible (Nandi *et al.,* 2022).

Region-Wise Data Extraction and Analysis Representations

The Central India mask is applied to the IMD national grid to keep only grid cells within Central India. We use two data representations. First, we compute a daily region-mean T_{max} series across all selected grid cells; this single time series is used for univariate forecasting with ARIMA/SARIMA and sequence models (LSTM/GRU). Second, we retain the grid-level T_{max} series for all Central India cells to apply IMD heatwave rules at the grid-cell level and to confirm that region-day heatwave labels reflect spatially consistent heating rather than isolated grid artifacts.

Data Cleaning

Invalid or missing data can cause inaccurate results, so careful handling of missing or invalid is important in first step, we treat placeholder entries like 99.9 as missing (NaN) in IMD T_{max} data. The aim is to avoid manufacturing rare hot values from bad inputs and keeping the time series usable for sequence model. In next step, these values are filled with a regional-average daily climatology value. These climatology values are computed over 1980-2010 temperature data for the same calendar day. These

Table 2: Sensitivity of Central India heatwave labels to coverage threshold θ

θ (Grid-cell fraction)	Heatwave Days	Heatwave Events	Mean Heatwave Coverage (fraction, 0-1)
0.05	449	99	0.18
0.10	305	66	0.24
0.20	137	36	0.35

climatology values replace the gap with temperature data that is physically reasonable for that place and date. This process aligns with how multi-decadal normals are commonly used in Indian heat related work (Srivastava *et al.,* 2022; Akaike, 1974). In the end, simple range checks are applied to catch values that are unlikely before training and heatwave labeling.

Grid-to-Region Link

Heatwave conditions are first computed at the grid-cell level using the IMD HW/SHW rule. For each day, the fraction of Central India grid cells flagged as HW/SHW is computed as the coverage p(t). A day is treated as a regional heatwave day when $p(t) \geq \theta$ (here $\theta=0.10$), and only spells of at least two consecutive days are kept. This converts grid-wise labels into a single region-day label that matches the regional-mean plots, while still ensuring that each marked event is supported by spatially coherent heating across the region.

Grid-Level IMD Heatwave Labeling Under Plains Criteria

Heatwave labels are constructed using IMD-style operational logic based on T_{max} thresholds and departures from a daily normal, along with intensity categories (India Meteorological Department [IMD], n.d.). Since the Central India mask excludes coastal subdivisions, plains criteria are applied.

For each grid cell, a daily "normal" T_{max} is computed as the long-term day-of-year climatology (1980–2010). The daily departure is computed as:

$$\Delta T(t) = T_{max}(t) - T_{max,normal}(t) \quad (1)$$

A grid-cell day is labeled as Heatwave (HW) if it satisfies either:

1. Departure-based (plains): $T_{max}(t) \geq 40^0 C$ and $4.5^0 C \leq \Delta T(t) \leq 6.4^0 C$, or
2. Absolute: $T_{max}(t) \geq 45^0$

A grid-cell day is labeled as Severe Heatwave (SHW) if it satisfies either:

1. Departure-based severe (plains): $T_{max}(t) \geq 40^0 C$ and $4.5^0 C > \Delta T(t) \leq 6.4^0 C$
2. Absolute severe: $T_{max}(t) \geq 47^0$

Regional aggregation, Minimum duration, and Sensitivity analysis

A day is labelled as a regional heatwave if $p(t) \geq \theta$; otherwise, it is non-heatwave. We use in the main experiments and test θ from 0.05 to 0.20 because IMD guidance does not define a gridded region-level coverage threshold. As θ increases, the number of labelled heatwave days and events fall (449 days/99 events to 137 days/36 events), while mean coverage on labelled heatwave days rises (0.180 to 0.350), indicating that stricter thresholds retain fewer but more spatially extensive events (Table 2).

After aggregation, only heatwave spells lasting at least two consecutive days are retained, consistent with common practice in India heatwave studies (India Meteorological Department [IMD], n.d.; Srivastava *et al.,* 2022). Overall, $\theta=0.10$, provides a practical balance between spatial coherence and having enough events for training and evaluation.

Example-Year Visualization of Central India Heatwave Labeling

To check whether labelling yields coherent multi-day events during the expected hot season, an example year plot is used which is shown in Fig. 4.

Heatwave days are marked using IMD criteria for plains applied at the grid-cell level. Then it is aggregated to a regional heatwave-day indicator using the two-day minimum duration rule. 2022 is selected as an illustrative year because it has the most observed heatwave activity in 2000–2024, which makes the labeling rule produce several clear multi-day spells for a quick visual check

Forecasting Models

Two model families, statistical models and recurrent neural networks are used and tested under same seasonal division and same evaluation windows. ARIMA and SARIMA are used as low-cost baselines because daily T_{max} is strongly seasonal and autocorrelated. These models keep the diagnostics and produce useful forecasts (Aksoy *et al.,* 2025). LSTM and GRU can learn nonlinear behavior. These models can also learn multi-day persistence that often shows up during hot spell using only univariate Tamx historical data (Ratnam *et al.,* 2023; He *et al.,* 2022). Central India masking and season wise design is kept fixed for fair comparison and to test whether deep learning models add clear gains over seasonal linear structure (Aksoy *et al.,* 2025). Other options such as ETS/Prophet and tree/boosting methods (RF/XGBoost/SVR) are not included to avoid expanding the scope and because they often require additional feature design or exogenous predictors for fair benchmarking; we note them as future extensions.

Statistical Baselines

Statistical models ARIMA and SARIMA are used to give classical baselines for univariate T_{max} forecasting. SARIMA keeps seasonality in view by adding explicit seasonal terms, while ARIMA serves as the simplest reference point when only short-memory structure is needed.

When the temperature series is highly auto-correlated and seasonally structured the selection of appropriate model orders is done with a rule of thumb based on the information criterion (AIC)

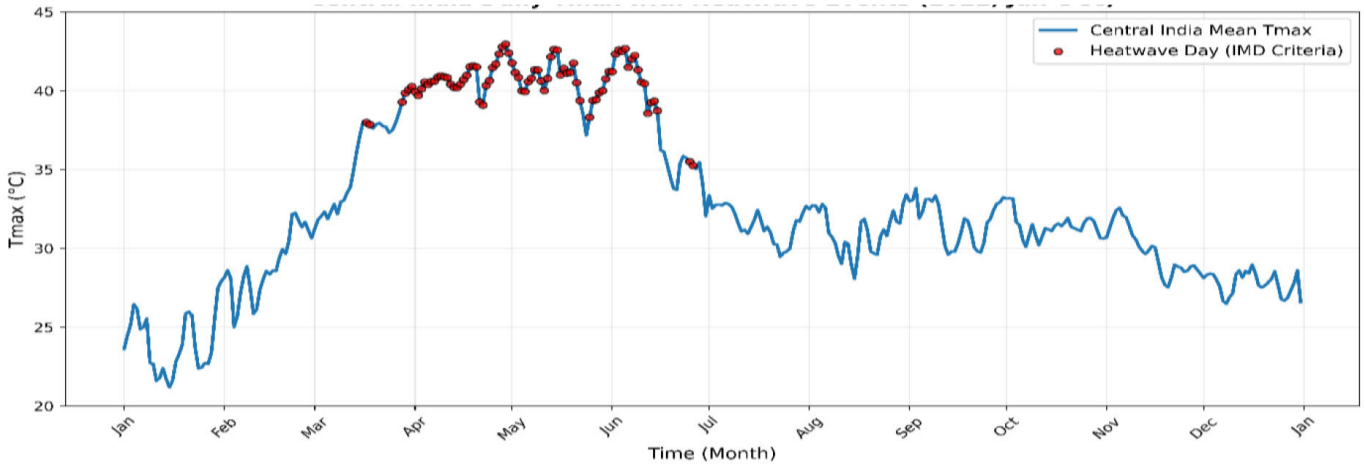

Fig. 4: Central India mean T_{max} and heatwave days for year 2022

that balances goodness of fit with the complexity of the model (Akaike 1974), and is considered as a modelling decision and not a hidden tuning step, as the value of these baselines is as much in their transparency as in their accuracy.

After the main autocorrelation, residual behavior is checked to make sure that there is no remaining serial dependence at lag 10 and 20, commonly used screeners for residual serial dependence (Ljung & Box, 1978). The tests for residual stationarity are also performed by ADF and KPSS tests, and the diagnostic p-values along with a short explanation are provided in appendix table S1 (Dickey & Fuller, 1979; Kwiatkowski *et al.,* 1992). These are all commonly automated forecasting workflows for maintaining baseline tuning reproducibility (Hyndman & Khandakar, 2008; Rohini & Rajeevan, 2023). Comparing with ARIMA, SARIMA adds seasonal term with explicit period m and all other settings are kept as same as in ARIMA. The ARIMA and SARIMA parameters for the season-wise Central India T_{max} experiments are given in Table 3. For reproducibility purposes it is included. The table also shows the data protocol used for sharing data (T_{max} is the target variable, season-based partitioning, chronological train-test split, and climatology-based filling missing data), as well as the model selection criterion.

The comparison remains fair between ARIMA and SARIMA as both of them search candidate orders on a grid and the winner of the search is the model with the lowest AIC on the training period; differences in results due to tuning efforts are more likely to be due to the model assumptions than to the tuning effort, if this effort is not even (Akaike, 1974).

Deep Learning

Recurrent neural networks (RNN) are well suited to time-series forecasting because they process data in sequence and can retain information from earlier time steps, which helps capture temporal dependence. RNN deep learning models GRU and LSTM are used to learn nonlinear temporal dependence and multi-day persistence from T_{max} sequences (Hochreiter & Schmidhuber, 1997; Cho *et al.,* 2014; Hou *et al.,* 2022; Kan *et al.,* 2025; Li *et al.,* 2023; Choudary *et al.,* 2025). The input to each model is a rolling window of the previous N days of T_{max} (with N=30), and the target is next-day T_{max}. The recurrent layer output is mapped to the forecast through a linear regression head. Inputs are scaled using parameters fit on the training period only, and predictions are transformed back to °C before evaluation. Architecture and training configuration of deep learning models, LSTM and GRU used for next-day forecasting of Central India regional-mean T_{max} is given in Table 4.

This supports reproducibility and state main design choices. Look-back window(N) of 30 days and train-test split of 80/20 is used in current implementation, and the network layer layouts. Both models are trained to predict univariate sequence on the same region-mean T_{max} series from Central India mask. Any difference in performance is mainly due to recurrent architecture and model capacity rather than data handling or inputs. The time order is preserved during training (no shuffling). T_{max} is scaled before training and the same scaling is applied to validation and test. Dropout is not used in the final run; generalization is controlled through chronological validation and early stopping. The input window is set to 30 days to capture short-term persistence and multi-day build-up typical of heat spells while keeping the input length stable. Two recurrent layers (64 units each) are used as a balance between capacity and overfitting risk, consistent with common practice in temperature sequence forecasting (Hochreiter & Schmidhuber, 1997; Hou *et al.,* 2022; Aksoy *et al.,* 2025).

RESULTS AND DISCUSSION

Training and Evaluation

Time-ordered splits are used to preserve causality in all experiments. For each season, first models are trained and then tested. Statistical models are fit only on the training segment and generate forecasts on the test segment. Fig. 5 shows the training and validation loss curves for LSTM and GRU. Hyperparameter settings are kept constant across all seasons for fair comparison within page constraints of the study. Mini batch optimization is used to train recurrent models. Recurrent model validation is performed using held-out tail of the training period when required for early monitoring.

Models are trained using chronological validation and early stopping based on validation loss. Heatwave-day predictions are derived from model outputs by applying the same heatwave definition used for observed labels. Specifically, predicted T_{max} is passed through the IMD-consistent grid-level HW/SHW logic, then

Table 3: ARIMA/SARIMA configuration and hyperparameter

Item	ARIMA	SARIMA
Model form	ARIMA (p,d,q)	SARIMA (p,d,q) (P,D,Q,m)
Model selection	AIC-min grid search (train)	AIC-min grid search (train)
Non-seasonal bounds	$p \leq 5, d \leq 2, q \leq 5$	$p \leq 3, d \leq 2, q \leq 3$
Seasonal bounds	Not applicable	$P \leq 2, D \leq 1, Q \leq 2$
Seasonal period (m)	Not applicable	Median seasonal samples/year.
Constraints	stationarity = False; invertibility = False	stationarity = False; invertibility = False

Table 4: Hyperparameter and configuration of LSTM and GRU

Item	LSTM	GRU
Input window	(30, 1)	(30, 1)
Split (time-ordered)	Train+Val first 80%, Test last 20%; Validation = last 20% of Train	Same
Architecture	2 layers: LSTM (64, tanh, return_sequences=True) → LSTM (64, tanh) → Dense (32, ReLU) → Dense (1)	2 layers: GRU (64, tanh, return_sequences=True) → GRU (64, tanh) → Dense (32, ReLU) → Dense (1)
Optimizer	Adam (lr = 1×10^{-3})	Adam (lr = 1×10^{-3})
Batch size	32	32
Training control	Up to 100 epochs; early stopping on val_loss (patience=6; restore best)	Same

aggregated to a Central India region-day label using θ=0.10, and finally filtered using the minimum two-day event duration rule, exactly as described in Section 3.4. This ensures that the event prediction stage is deterministic and identical across models.

Two categories of evaluation metrics are used:

1. T_{max} forecasting accuracy: RMSE, MAE, and R2 are computed on daily T_{max} forecasts. RMSE and MAE are both retained because they summarize error differently, with RMSE giving more weight to larger deviations (Ratnam *et al.,* 2023).

2. Heatwave-day detection skill: Accuracy, precision, recall (POD), and F1-score are computed from the region-day confusion matrix. Heatwave days are less frequent than non-heatwave days. So, precision-recall measures are highlighted as more informative indicators of missed events and false alarms than accuracy alone (Hodson, 2022). In the test split, heatwave days account for 305 of 1827 days (16.7%), and in the training split they account for 1408 of 7305 days (19.3%). Because accuracy can be inflated under class imbalance, we report precision, recall (POD), F1, FAR and CSI derived from the confusion matrix. Thresholds are selected using chronological validation to avoid bias toward the majority

In this section, models are checked in two ways. First, we look at how well they forecast daily temperature using RMSE, MAE, and R^2. Second, we check whether those forecasts translate into correct heatwave-day decisions using event metrics like accuracy, precision, recall (POD), and F1. Table 5 brings these two views together so the comparison is not based on temperature error alone. Table 5 gives the headline results, but it does not show where mistakes happen. For that, Table 6 reports the train and test confusion matrices along with FAR and CSI. This makes it easy to see how many heatwave days are detected (TP), how many are missed (FN), and how often the model raises false alarms (FP). On this split, LSTM keeps the best balance, with strong recall while holding false alarms at a manageable level, and it stays stable from training to testing compared with the statistical baselines. Central India T_{max} forecasting errors and ranking of all models is shown in table 5.

ARIMA has the highest RMSE=4.5 °C and MAE = 3.5 °C errors. SARIMA improves both and has RMSE = 3.5 °C and MAE = 2.8 °C error values. SARIMA is consistent with strong seasonality in T_{max} using AR and MA terms. Recurrent models outperform statistical models. GRU achieves RMSE = 2.4 °C and MAE = 1.8 °C. LSTM further reduces the error to RMSE = 2.2 °C and MAE = 1.6 °C. Compared to ARIMA, LSTM reduces RMSE by 2.3 °C and MAE by 1.9 °C. Results are indicating that gated networks capture nonlinear persistence and longer temporal dependence beyond differencing and seasonal adjustments. R^2 along with heatwave-day accuracy from the fixed labelling approach is shown in table 5. The R^2 values match the error results and summaries explained variance. ARIMA reaches R^2=0.65, SARIMA improves to 0.75. R^2 value for GRU is 0.88 and 0.92 for LSTM. Gains by SARIMA over ARIMA confirms that seasonality drives much of the predictability in regional T_{max}. Further improvement with GRU and LSTM suggest that remaining variance comes from nonlinear short-term dynamics, which are not captured by linear seasonal models.

Heatwave-day accuracy increases from 80%(ARIMA) to 84.5%(SARIMA) for statistical models. This accuracy is further increased by GRU to 91% and by LSTM to 93%. However, accuracy alone is limited for heatwave prediction because heatwave days are rarer than non-heatwave days. So, a conservative model can score well by often predicting non-heatwave days. Therefore, event metrics are given importance, and accuracy is used mainly to check

Table 5: Overall performance summary on Central India (2000–2024)

Model	RMSE (°C)	MAE (°C)	R^2	Accuracy	Precision	Recall (POD)	F1 Score
ARIMA	4.515	3.567	0.654	0.801	0.410	0.449	0.429
SARIMA	3.521	2.827	0.754	0.845	0.529	0.620	0.571
GRU	2.432	1.841	0.882	0.911	0.740	0.810	0.773
LSTM	2.223	1.683	0.924	0.928	0.751	0.849	0.797

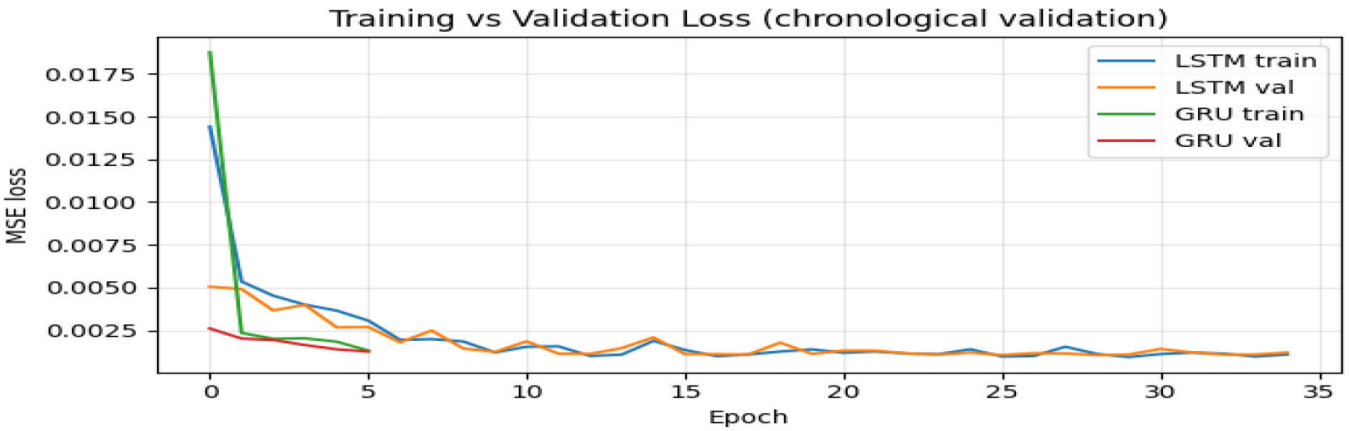

Fig. 5: Training and validation loss curves for LSTM and GRU under chronological validation with early stopping.

that gains in event skill are not driven by broad misclassification (Hodson, 2022).

Precision, recall (POD), and F1 score reflect false alarms and missed events. The statistical baselines show weak event discrimination. ARIMA achieves precision=0.40, recall=0.45 and F1 score=0.42. SARIMA perform better than ARIMA with precision=0.52, recall=0.62 and F1=0.57. SARIMA's higher recall represents fewer missed heatwave days. SARIMA's capability of capturing seasonal baseline reduces warm-season bias and improves detection around event onset. The recurrent models improve gains in both missed-event reduction and false-alarm control. GRU reaches precision = 0.74 and recall = 0.81 (F1 = 0.77), while LSTM performs best with precision = 0.75 and recall = 0.84 (F1 = 0.79). The rise in both, precision and recall suggest that LSTM and GRU are able to separate sustained heatwave conditions from non-event warm days under same IMD-consistent definition. Practically, higher recall reduces missed warnings and precision limits false alarms and warning fatigue. LSTM gives the best balance. GRU remains a strong option with slightly lower F1 and typically lower computational complexity in similar recurrent setting. Table 6 shows the train and test confusion-matrix results for all four models. It makes the trade-off clear: LSTM keeps high detection skill on the test set, while ARIMA and SARIMA miss more events and also raise more false alarms.

Table 7 adds 95% confidence intervals for RMSE, F1 and CSI on the test set, so the main results are shown with uncertainty, not just single numbers. RMSE is reported in °C; F1 and CSI are heatwave-day detection metrics. Confidence intervals are computed using a 7-day moving block bootstrap with 2000 resamples.

On the test set, LSTM and GRU keep a similar recall (0.85 and 0.81) but LSTM has better precision and fewer missed days overall. ARIMA performs weakest, mainly because false alarms are high (FAR ≈ 0.59), while SARIMA improves event detection but still stays behind the deep models. Along with accuracy, precision/recall and CSI are reported because the heatwave class is rare and accuracy alone can be misleading. Along with TP/FP/FN/TN, Table 6 also lists the total observed and predicted heatwave days, so it is clear whether a model tends to under-predict or over-predict heatwave periods. Observed and LSTM-predicted regional-mean T_{max} for Central India during March-June 2024 is shown in Fig. 6. We show 2024 because it is the most recent out-of-sample year in our dataset and provides a direct check of model behaviour on current heat conditions. IMD-consistent pipeline (grid-level HW/SHW, region-day aggregation with θ=0.10, and minimum 2-day duration) is used to mark observed heatwave days.

Predicted heatwave days are derived from the predicted regional-mean T_{max} using the same persistence rule. The plot aids as a quantitative check that model capture the timing and multi-days persistence of heat spells linked to heatwave events. Mandal *et al.,* (2019) assess extended-range heatwave prediction over India (week-1 to week-4) using ACC for T_{max} and event metrics such as SEDI and reliability, with skill decreasing as lead time increases (Mandal *et al.,* 2019). Ratnam *et al.,* (2023) report that 10-day T_{max} anomaly prediction remains challenging in March and June (e.g., persistence ACC ≈ 0.38; CFS ACC 0.72 in March and 0.62 in June; best ML ACC 0.27 in March), which supports using short-horizon regional baselines when the goal is day-to-day heatwave-day warning (Ratnam *et al.,* 2023).

The gap between ARIMA/SARIMA and GRU/LSTM indicates both the T_{max} series structure and the heatwave labelling rule.

1. Central India heatwaves last many days and cause sustained warming. Recurrent networks learn dependencies over prior days and can represent gradual build-up and persistence

Table 6: Confusion matrix–based evaluation of heatwave detection models showing training and testing performance.

Model	Split	TP	FP	FN	TN	Observed HW Days	Pred HW Days	Accuracy	Precision	Recall	F1	FAR	CSI
ARIMA	Train	634	910	774	4987	1408	1544	0.816	0.41	0.45	0.42	0.59	0.27
ARIMA	Test	137	197	168	1325	305	334	0.797	0.41	0.45	0.43	0.59	0.27
SARIMA	Train	873	775	535	5122	1408	1648	0.854	0.53	0.62	0.55	0.47	0.40
SARIMA	Test	189	168	116	1354	305	357	0.800	0.53	0.62	0.57	0.47	0.40
GRU	Train	1165	365	243	5532	1408	1530	0.891	0.74	0.81	0.75	0.26	0.64
GRU	Test	247	87	58	1435	305	334	0.921	0.74	0.81	0.77	0.26	0.64
LSTM	Train	1197	399	211	5498	1408	1596	0.92	0.75	0.85	0.83	0.25	0.68
LSTM	Test	259	86	46	1436	305	345	0.926	0.75	0.85	0.80	0.25	0.69

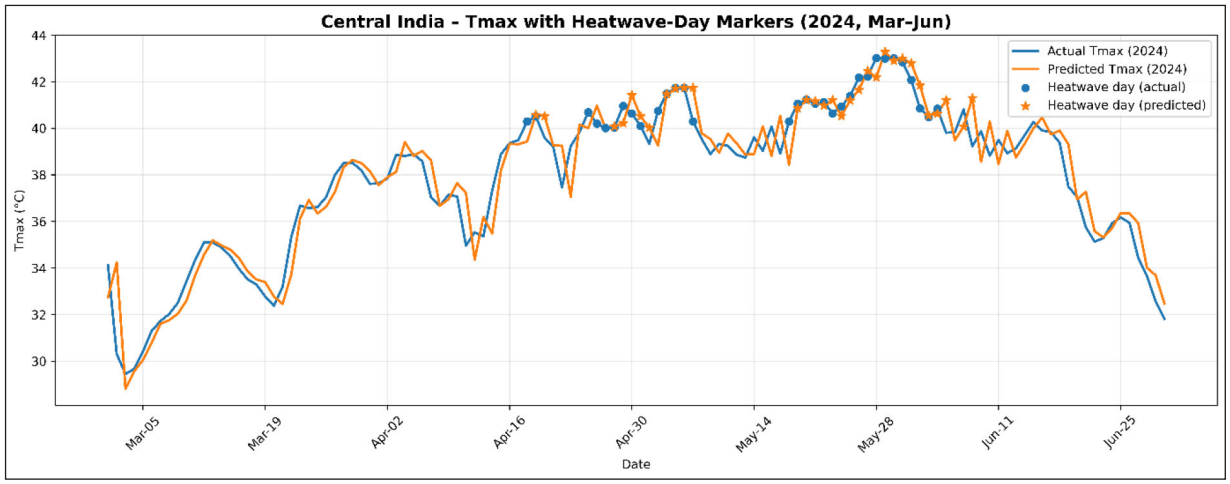

Fig. 6: Central India regional-mean T_{max} for the most recent out-of-sample year (March–June 2024): observed vs LSTM-predicted T_{max} with heatwave-day markers

more directly than linear stochastic terms (Hochreiter & Schmidhuber, 1997; Cho *et al.,* 2014; Li *et al.,* 2023; Hou *et al.,* 2022; Choudary *et al.,* 2025). This helps with large recall gains, better detection of events and continuation under the same fixed definition.

2. Central India T_{max} is directly affected by changes in weather systems that can cause sharp day-to-day changes around a high seasonal baseline. Some linear models can capture the seasonality. But these models struggle when the link between recent history and next day temperature data is nonlinear. Sequence models can capture this nonlinear mapping within the same input window and have lower error rates.

3. Heatwave labelling is dependent on threshold and persistence. So, when forecast error is dropped, fewer days are pushed to the wrong side of the threshold under the same labelling rule. This implies that there is a relationship between enhanced regression skill and increased event F1, while not altering the definition of heatwave. Fewer heatwave days classified under an operational rule set for a lower temperature (Akaike, 1974; Satyanarayan & Bhaskar Rao, 2020; Mandal *et al.,* 2019; Das *et al.,* 2020).

Based on these results, two practical conclusions are drawn for a region based baseline study:

1. SARIMA is a better baseline reference than ARIMA for Central India T_{max} because modelling seasonality reduces error and improves event recall. It qualifies to become the main linear benchmark.

2. Overall LSTM is the best method for both the T_{max} regression and heatwave day detection. It is the most appropriate model for further development, calibration and multi-region-season utilization.

The output of the heatwave-day is used as a regional risk flag for agrometeorological advisories. India already has similar short-term advisory services which translate forecasts into short-term guidance for timely irrigation and farm operations (Rathore *et al.,* 2025; Nannewar *et al.,* 2023). Heat spells in the central Indian region are significant for wheat around March-April as brief duration of heat around 35 °C during flowering and grain filling can cause a significant reduction in yield (Talukder *et al.,* 2014). Soybean was reported to have similar heat sensitivity during its reproductive stages (Alsajri *et al.,* 2020) and cotton during flowering and boll development under very high temperatures (Majeed *et al.,* 2021). For practical purposes, a region-level heatwave warning can support decisions, including planning irrigation schedules to start earlier in the hot spell, shifting labour-intensive tasks to cooler times of day, and postponing stress-sensitive tasks to cooler days. The signal is regional, so it is best used for early screening and then refined into

Table 7: Uncertainty of test-set performance using 95% confidence intervals (block bootstrap)

Model	RMSE (°C)	RMSE 95% CI	F1 (test)	F1 95% CI	CSI (test)	CSI 95% CI
ARIMA	4.505	(4.364, 4.656)	0.429	(0.380, 0.476)	0.273	(0.235, 0.312)
SARIMA	3.501	(3.391, 3.618)	0.571	(0.525, 0.615)	0.400	(0.356, 0.444)
GRU	2.412	(2.336, 2.493)	0.773	(0.736, 0.807)	0.630	(0.582, 0.677)
LSTM	2.213	(2.144, 2.287)	0.797	(0.761, 0.829)	0.662	(0.615, 0.709)

district- or crop-stage guidance where local information is available (Rathore *et al.,* 2025; Nannewar *et al.,* 2023).

CONCLUSION AND FUTURE WORK

This work introduces a reproducible T_{max} -only framework of predicting heatwave for Central India based on the IMD defined label and region-day aggregation with sensitivity checking to derive the threshold for heatwave with balanced coverage. The results indicate that season-aware baselines provide improved forecasts, and that the combined skill of predicting T_{max} and detecting heatwave days is most promising with recurrent sequence models. This is a practical limit for this baseline. While LSTM/GRU models are black-boxes as they do not offer any physical interpretation, they do need more compute to retrain than statistical models. Additionally, the results may be sensitive to hyperparameters like window length and number of hidden units, requiring careful validation. Future studies could expand the baseline to cover the remaining areas of IMD, as there will be consistent treatment of regional variations in seasonality, thresholds, and event persistence under one evaluation protocol. For the recurrent models, optimization can be targeted to strengthen the models. An incremental learning setup can be implemented, where the model can be updated gradually over each year without retraining the model completely. This allows for better utilization of existing long historical records, and will enable us to evaluate and track, under the same leakage-safe protocol and without changing the definitions of heatwaves, how well the model performs over time (even if it adapts to the new conditions, retains skill on previous seasons and years, or requires a longer training time or more computation).

ACKNOWLEDGEMENT

The authors gratefully acknowledge the India Meteorological Department (IMD), Ministry of Earth Sciences, Government of India, for providing the gridded temperature datasets used in this study.

Author's certificate: The manuscript entitled "Performance Evaluation of Regional Heatwave Prediction Using Statistical and Deep Learning Models" is an original work of the authors and is not under consideration for publication elsewhere, either in whole or in part. The manuscript has been read and approved by all co-authors, and all authors agree to its submission for publication.

Conflict of Interests: The authors declare that there is no conflict of interest regarding the publication of this article.

Funding: This research was self-funded. No external funding was received for conducting this study.

Authors contribution: **Sulochana Devi:** Conceptualization, Methodology, Software Development, Formal Analysis, Writing-Original Draft; **Radhika Kotecha:** Supervision, Validation, Data Curation, Review and Editing

Disclaimer: The contents, opinions, and views expressed in the research article published in the Journal of Agrometeorology are the views of the authors and do not necessarily reflect the views of the organisations they belong to.

Publisher's Note: The periodical remains neutral concerning jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Aksoy, M. M., Mowla, M. N., Bilgili, M., Pinar, E., Durhasan, T., & Asadi, D. (2025). Forecasting near-surface air temperature via SARIMA and LSTM: A regional time-series study. *Journal of Atmospheric and Solar-Terrestrial Physics, 275*, 106604. https://doi.org/10.1016/j.jastp.2025.106604

Alsajri, F. A., Wijewardana, C., Irby, J. T., Bellaloui, N., Krutz, L. J., Golden, B., Gao, W., & Reddy, K. R. (2020). Developing functional relationships between temperature and soybean yield and seed quality. *Agronomy Journal, 112*(1), 194–204. https://doi.org/10.1002/agj2.20034

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Choudary, R. V., Johnvictor, A. C., & Sankar, P. N. (2025). Comparative analysis of machine learning approaches for heatwave event prediction in India. *Scientific Reports, 15*, 22431. https://doi.org/10.1038/s41598-025-04634-9

Cvijanovic, I., Mistry, M. N., Begg, J. D., Gasparrini, A., & Rodó, X. (2023). Importance of humidity for characterization and communication of dangerous heatwave conditions. *npj Climate and Atmospheric Science, 6*, 33. https://doi.org/10.1038/s41612-023-00346-x

Das, P. K., Podder, U., Das, R., Kamalakannan, C., Rao, G. S., Bandyopadhyay, S., & Raj, U. (2020). Quantification of

heat wave occurrences over the Indian region using long-term (1979–2017) daily gridded (0.5° × 0.5°) temperature data—a combined heat wave index approach. *Theoretical and Applied Climatology, 142*, 497–511. https://doi.org/10.1007/s00704-020-03329-7

Davis, L., Gertler, P., Jarvis, S., & Wolfram, C. (2021). Air conditioning and global inequality. *Global Environmental Change, 69*, 102299. https://doi.org/10.1016/j.gloenvcha.2021.102299

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531

Fischer, E. M., Sippel, S., & Knutti, R. (2021). Increasing probability of record-shattering climate extremes. *Nature Climate Change, 11*, 689–695. https://doi.org/10.1038/s41558-021-01092-9

Ganeshi, N. G., Mujumdar, M., Takaya, Y., Goswami, M. M., Singh, B. B., Krishnan, R., & Terao, T. (2023). Soil moisture revamps the temperature extremes in a warming climate over India. *npj Climate and Atmospheric Science, 6*, 12. https://doi.org/10.1038/s41612-023-00334-1

He, Z., Jiang, T., Jiang, Y., Luo, Q., Chen, S., Gong, K., He, L., Feng, H., Yu, Q., Tan, F., & He, J. (2022). Gated recurrent unit models outperform other machine learning models in prediction of minimum temperature in a greenhouse based on local weather data. *Computers and Electronics in Agriculture, 202*, 107416. https://doi.org/10.1016/j.compag.2022.107416

Heino, M., Kinnunen, P., Anderson, W., Ray, D. K., Puma, M. J., Varis, O., Siebert, S., & Kummu, M. (2023). Increased probability of hot and dry weather extremes during the growing season threatens global crop yields. *Scientific Reports, 13*, 3583. https://doi.org/10.1038/s41598-023-29378-2

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development, 15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Hou, J., Wang, Y., Zhou, J., & Tian, Q. (2022). Prediction of hourly air temperature based on CNN–LSTM. *Geomatics, Natural Hazards and Risk, 13*(1), 1962–1986. https://doi.org/10.1080/19475705.2022.2102942

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

India Meteorological Department (IMD). (n.d.). Chapter 2: Cold and heat wave indices and methodology [PDF]. Ministry of Earth Sciences, Government of India. https://mausam.imd.gov.in/responsive/pdf_viewer_css/met2/Chapter%20-2/Chapter%20-2.pdf (Accessed February 1, 2026)

Kan, C., Vieira Passos, M., Destouni, G., Barquet, K., Ferreira, C. S. S., & Kalantari, Z. (2025). Seasonal heatwave forecasting with explainable machine learning and remote sensing data. *Stochastic Environmental Research and Risk Assessment, 39*, 3333–3352. https://doi.org/10.1007/s00477-025-03020-1

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Li, P., Li, Y., Qiu, S., Bai, Y., Gong, X., & Wang, R. (2023). Regional heatwave prediction using graph neural network and weather station data. *Geophysical Research Letters, 50*(7), e2023GL103405. https://doi.org/10.1029/2023GL103405

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303. https://doi.org/10.1093/biomet/65.2.297

Lorenz, R., Jaeger, E. B., & Seneviratne, S. I. (2010). Persistence of heat waves and its link to soil moisture memory. *Geophysical Research Letters, 37*, L09703. https://doi.org/10.1029/2010GL042764

Lu, Y.-C., & Romps, D. M. (2022). Extending the heat index. *Journal of Applied Meteorology and Climatology, 61*(10), 1367–1383. https://doi.org/10.1175/JAMC-D-22-0021.1

Majeed, S., Rana, I. A., Mubarik, M. S., Atif, R. M., Yang, S.-H., Chung, G., Jia, Y., Du, X., Hinze, L., & Azhar, M. T. (2021). Heat stress in cotton: A review on predicted and unpredicted growth-yield anomalies and mitigating breeding strategies. *Agronomy, 11*(9), 1825. https://doi.org/10.3390/agronomy11091825

Mandal, R., Joseph, S., Sahai, A. K., Phani, R., Dey, A., Chattopadhyay, R., & Pattanaik, D. R. (2019). Real-time extended range prediction of heat waves over India. *Scientific Reports, 9*, 9008. https://doi.org/10.1038/s41598-019-45430-6

Mandal, R., Joseph, S., Waje, S., Chaudhary, A., Dey, A., Kalshetti, M., & Sahai, A. K. (2025). Heat waves in India: Patterns, associations, and subseasonal prediction skill. *Climate Dynamics, 63*, Article 42. https://doi.org/10.1007/s00382-024-07539-x

Nandi, S., Patel, P., & Swain, S. (2022). IMDLIB: A python library for IMD gridded data (Version v2) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.7205414

Nannewar, R. G., Kanitkar, T., & Srikanth, R. (2023). Role of agrometeorological advisory services in enhancing food security and reducing vulnerability to climate change. *Weather, Climate, and Society, 15*(4), 1013–1027. https://doi.org/10.1175/WCAS-D-22-0130.1

Perkins-Kirkpatrick, S. E., & Lewis, S. C. (2020). Increasing trends in regional heatwaves. *Nature Communications, 11*, 3357. https://doi.org/10.1038/s41467-020-16970-7

Ratnam, J. V., Behera, S. K., Ratna, S. B., Rajeevan, M., & Yamagata, T. (2016). Anatomy of Indian heatwaves. *Scientific Reports, 6*, 24395. https://doi.org/10.1038/srep24395

Ratnam, J. V., Behera, S. K., Nonaka, M., Martineau, P., & Patil, K. R. (2023). Predicting maximum temperatures over India 10-days ahead using machine learning models. *Scientific Reports, 13*, 17208. https://doi.org/10.1038/s41598-023-44286-1

Rathore, L. S., Ghosh, K., & Singh, K. K. (2025). Evolution of agromet advisory services in India. *MAUSAM, 76*(1), 231–256. https://doi.org/10.54302/mausam.v76i1.6486

Rohini, P., & Rajeevan, M. (2023). An analysis of prediction skill of heat waves over India using TIGGE ensemble forecasts. *Earth and Space Science, 10*(3), e2020EA001545. https://doi.org/10.1029/2020EA001545

Satyanarayana, G. Ch., & Bhaskar Rao, D. V. (2020). Phenology of heat waves over India. *Atmospheric Research, 245*, 105078. https://doi.org/10.1016/j.atmosres.2020.105078

Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., Teuling, A.J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews, 99*(3–4), 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004

Somanathan, E., Somanathan, R., Sudarshan, A., & Tewari, M. (2021). The impact of temperature on productivity and labor supply: Evidence from Indian manufacturing. *Journal of Political Economy, 129*(6), 1797–1827. https://doi.org/10.1086/713733

Srivastava, A. K., Rajeevan, M., & Kshirsagar, S. R. (2009). Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. *Atmospheric Science Letters, 10*(4), 249–254. https://doi.org/10.1002/asl.232

Srivastava, A., Mohapatra, M., & Kumar, N. (2022). Hot weather hazard analysis over India. *Scientific Reports, 12*, 19768. https://doi.org/10.1038/s41598-022-24065-0

Talukder, A. S. M. H. M., McDonald, G. K., & Gill, G. S. (2014). Effect of short-term heat stress prior to flowering and early grain set on the grain yield of wheat. *Field Crops Research, 160*, 54–63. https://doi.org/10.1016/j.fcr.2014.01.013