Computers and Electronics in Agriculture 236 (2025) 110450

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Quantifying error in fine-scale crop yield forecasts to guide data and algorithm improvements: case study of mango in Tamil Nadu, India

Louis Kouadio [a,b,*], Bhuvaneswari Kulanthaivel [c], Thanh Mai [a], Thong Nguyen-Huy [a], Qingxia (Jenny) Wang [a,d], Vivekananda M. Byrareddy [a], Shahbaz Mushtaq [a], Alagarswamy Senthil [e], Nathaniel K. Newlands [f], Vellingiri Geethalakshmi [g]

[a] Centre for Applied Climate Sciences, Institute for Life Sciences and the Environment, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[b] Africa Rice Center, Bouake 01, Côte d'Ivoire
[c] Centre for Agricultural and Rural development Studies, Tamil Nadu Agricultural University, Coimbatore 641003 Tamil Nadu, India
[d] School of Business, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[e] Department of Crop Physiology, Tamil Nadu Agricultural University, Coimbatore 641003 Tamil Nadu, India
[f] Summerland Research and Development Centre, Agriculture and Agri-Food Canada, Summerland, British Columbia V0H 1Z0, Canada
[g] Office of the Vice Chancellor, Tamil Nadu Agricultural University, Coimbatore 641003 Tamil Nadu, India

ARTICLE INFO

Keywords:
Mangifera indica
Climate variability
SPEI
Crop yield prediction

ABSTRACT

Accurate and timely prediction of mango yield is essential for optimizing resource management, market planning, and climate adaptation strategies. However, dealing with spatial variation of uncertainty and error in fine-scale (e.g., district) yield forecasts has yet to be fully explored. This study investigates a modelling approach that combines statistical methods including bootstrap robust least-angle regression, leave-one-out cross-validation, Bayesian-based spatial correlation analysis, and Markov chain Monte Carlo scheme, and machine learning (ML) (random forest technique) to enhance predictor selection, capture spatial trends, and generate probabilistic mango yield forecasts at the district scale in Tamil Nadu, India. Results showed that pre-flowering drought stress, temperature fluctuation and rainfall distribution, during flowering, fruit set, and fruit development, along with drought conditions in March and July, were dominant drivers of yield variability. Model evaluation revealed acceptable levels of errors in estimating mango yield, with root mean square error $\leq$ 2.0 t ha^{-1}, and mean absolute percentage error $\leq$ 30 % in 18 out of 31 districts. However, forecasting errors at three different lead times (two and one months prior to, and at start of harvest) varied spatially across districts, with lower errors in southern and north-western regions but higher errors in northern and central districts, reflecting the complexity of district-level forecasting under diverse environmental conditions. Agroclimatic variables alone might not be sufficient for accurate mango yield forecasts across Tamil Nadu. By integrating diverse data for model training and refining the ML-based forecast algorithm between fine-scale regions, this study can serve as a foundation for developing climate-resilient mango production strategies tailored to regional variability.

1. Introduction

Mangoes (*Mangifera indica*) are highly valued export commodities for many tropical and subtropical countries around the world. The mango sector contributes substantially to the overall agricultural economy, international trade, and foreign exchange earnings in these countries (FAOSTAT, 2023). In 2021, the global production of mangoes, guavas and mangosteens was ca. 57 million metric tonnes (Mt) (FAOSTAT, 2023). Asia was the top mango-producing region, accounting for ca. 42 Mt, followed by Africa with ca. 8.5 Mt, and the Americas with ca. 6.6 Mt. At the national scale, India was the largest producer with ca. 24.9 Mt, followed by China (ca. 3.9 Mt), Indonesia (ca. 3.5 Mt), Pakistan (ca. 2.6 Mt), and Mexico (c.a. 2.4 Mt) (FAOSTAT, 2023).

Recognizing the important contributions of the mango industry to regional and national economies, timely and accurate mango yield forecasts are crucial for the various stakeholders involved in the industry, as it enables farmers, investors, and financial institutions to make informed decisions regarding the optimization of resource

* Corresponding author.
E-mail address: louis.kouadio@unisq.edu.au (L. Kouadio).

https://doi.org/10.1016/j.compag.2025.110450
Received 11 January 2024; Received in revised form 16 April 2025; Accepted 21 April 2025
Available online 30 April 2025

allocation and risk management (Stone and Meinke, 2005). In addition, reliable mango yield forecasts help in workforce planning on farms, ensuring adequate labour and resources during the harvesting season. Mango cultivation in most producing countries is rainfed (Rojas-Sandoval and Acevedo-Rodríguez, 2022). With the increasing challenges linked to variable weather and climate patterns, water scarcity, and soil fertility depletion, mango yield forecasts are essential for adapting crop management practices to optimize resources while maintaining or increasing yield levels. For governmental agencies and other policy makers, accurate forecasts aid in formulating appropriate policies and interventions related to the mango industry and ensuring its sustainability and profitability.

In response to the need in forecasting fruit tree yield, various models incorporating multiple factors, such as climate variables, soil fertility, phenology, physiological processes, and remote sensing-derived information, have been developed to predict the yield and/or production of fruit trees (e.g., Anderson et al., 2021; He et al., 2022; Gao et al., 2024; Santos et al., 2024). These include process-based or crop models, statistical models and machine learning (ML)-based approaches. ML-based approaches are often employed to explore the relationships between a set of features and crop yield that achieve the highest level of accuracy and precision, particularly when the number of potential 'predictors' of a problem is large and dimensional reduction is required (van Klompenburg et al., 2020). Several studies have utilized ML-based approaches for mango yield prediction at various spatial scales (e.g., Fukuda et al., 2013; Ray et al., 2023; Torgbor et al., 2023). For example, random forest (RF) models based on rainfall and irrigation data have been developed to estimate mango yields under different water regimes (rainfed or irrigated) at the farm scale in northern Thailand (Fukuda et al., 2013). The models achieved varying levels of accuracy, with the Nash-Sutcliffe efficiency and Pearson's correlation coefficients ranging from 0.369 to 0.910 and 0.330 to 0.964, respectively, depending on the water regime (Fukuda et al., 2013). Similarly, Torgbor et al. (2023) evaluated the capability of six different ML algorithms including RF, support vector regression, ridge regression, least absolute shrinkage and selection operator regression, partial least square regression, and extreme gradient boosting, to predict mango yield at the orchard block and farm levels in Australia at 3-month lead time using weather variables and remote sensing-derived vegetation indices. The errors (normalized mean absolute error) of the best algorithm (i.e., RF) varied between 31 % and 52 % at the block level, and between 0.8 % and 43 % at the farm level (Torgbor et al., 2023). In-field machine vision techniques, high spatial resolution satellite or unmanned aerial vehicle imagery and deep learning techniques were used for the estimation of mango fruit load and yield at the orchard scale under various environmental conditions in Senegal (e.g., Sarron et al., 2018) and Australia (e.g., Koirala et al., 2019). For India, various models such as autoregressive integrated moving average models, exponential smoothing models, autoregressive neural networks-based models, and support vector regression-based models have been used to predict mango production at the state (Santosha et al., 2022) and country levels (Ray et al., 2016; Ray et al., 2023). However, predicting mango yield at a finer scale, e.g., district scale or sub-district scale, while dealing with spatial variation of uncertainty and error, remains an open research question.

The objective of this study was to assess the performance of a modelling approach that combines statistical methods including bootstrap robust least-angle regression, leave-one-out cross-validation, Bayesian-based spatial correlation analysis and Markov chain Monte Carlo scheme, and machine learning technique (i.e., RF) to enhance predictor selection, capture spatial trends, and generate probabilistic mango yield forecasts at the district scale. The approach used agroclimatic indicators as potential predictors. Mango growing districts in the state of Tamil Nadu, India (average district size varying between 1,300 km^2 and 12,000 km^2) were chosen as a case study. In this study, we utilize the term "yield estimates" to denote the quantification of mango yield potential based on logical consequences of model structure. This is distinct from "yield forecasting " which refers to a probabilistic projection of future mango yields after data are assimilated into the model (Newlands et al., 2014). The results of this study can serve as a basis for developing a tool for mango yield forecasting that would provide substantial benefits to the entire mango industry in India and other mango-producing countries worldwide.

2. Materials and methods

2.1. Study region

Tamil Nadu is among the top 10 leading states in mango production in India (Government of Tamil Nadu, 2022). Between 2017/2018 and 2021/2022 the total area under mango cultivation in Tamil Nadu was ca. 144,000 ha on average, and the total production varied between 482,000 tonnes and 640,000 tonnes (Government of Tamil Nadu, 2022). The major mango growing districts in Tamil Nadu are Krishnagiri, Dharmapuri, Vellore, Dindigul, Thiruvallur and Theni. Mango virtually can grow in all the districts of Tamil Nadu (Fig. 1) as it tolerates a wide range of climatic and soil conditions. Rainfall ranging between 890 mm year^{-1} and 1015 mm year^{-1}, along with average air temperatures of 24–30 °C are the optimum growing conditions. Red loamy and well-drained soils, with a pH of 5.5 to 7.5 are ideal for mango cultivation in Tamil Nadu. However, the state has a diverse topography. For example, the major portion of Tiruppur falls within the rain shadow region of the Western Ghats, making it prone to water scarcity, which can influence adversely mango growth conditions. Likewise, Namakkal features varied landscapes, with mountainous areas in the north and plains in the south. Mango orchards are typically cultivated under rainfed conditions in Tamil Nadu. *Alphonso*, *Totapuri*, *Banganapalli*, and *Neelum* are the main commercial mango varieties grown across the state.

2.2. Data

2.2.1. Yield data

The official total harvested areas and production data for mangoes at the district level in Tamil Nadu for the study period of 2008–2019 were acquired from publicly available databases maintained by the Government of Tamil Nadu Department of Economics and Statistics (https://des.tn.gov.in/). A description of the data collection methodology can be found in Government of India (2022). For each district and for each year, mango yield was calculated by dividing the total production by the total harvested area. Yield is considered as it allows for the assessment of productivity irrespective of orchard size and facilitates comparisons across districts. No production data were available for the district of Ariyalur in 2008; therefore no yield value was calculated for that year for this district. An overview of the mango yield distribution for the entire study period for each district, along with the coefficients of yield variations, is presented in Fig. 2.

2.2.2. Climate data

Daily climate data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) for the 1991–2020 period were obtained from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/). Gridded data for rainfall (mm), minimum and maximum temperatures (°C), and solar radiation (MJ m^{-2}) at a 0.25° × 0.25° spatial resolution for the 2007–2019 period were used. The number of data points per district ranged between 2 (Nagapattinam and Perambalur) and 10 (Viluppuram) (Table S1). For a given district and for each climate variable, data for all grids in the district were averaged to obtain district-scale data. Monthly standardized precipitation-evaporation index (SPEI) values for the 2007–2019 period for each district were calculated using the 30-year ERA5 climate dataset (1991–2020) according to the method described in Vicente-Serrano et al. (2010). SPEI calculation were carried out using the package 'SPEI' (Beguería and Vicente-Serrano, 2023) in R (R Core Team, 2023). The

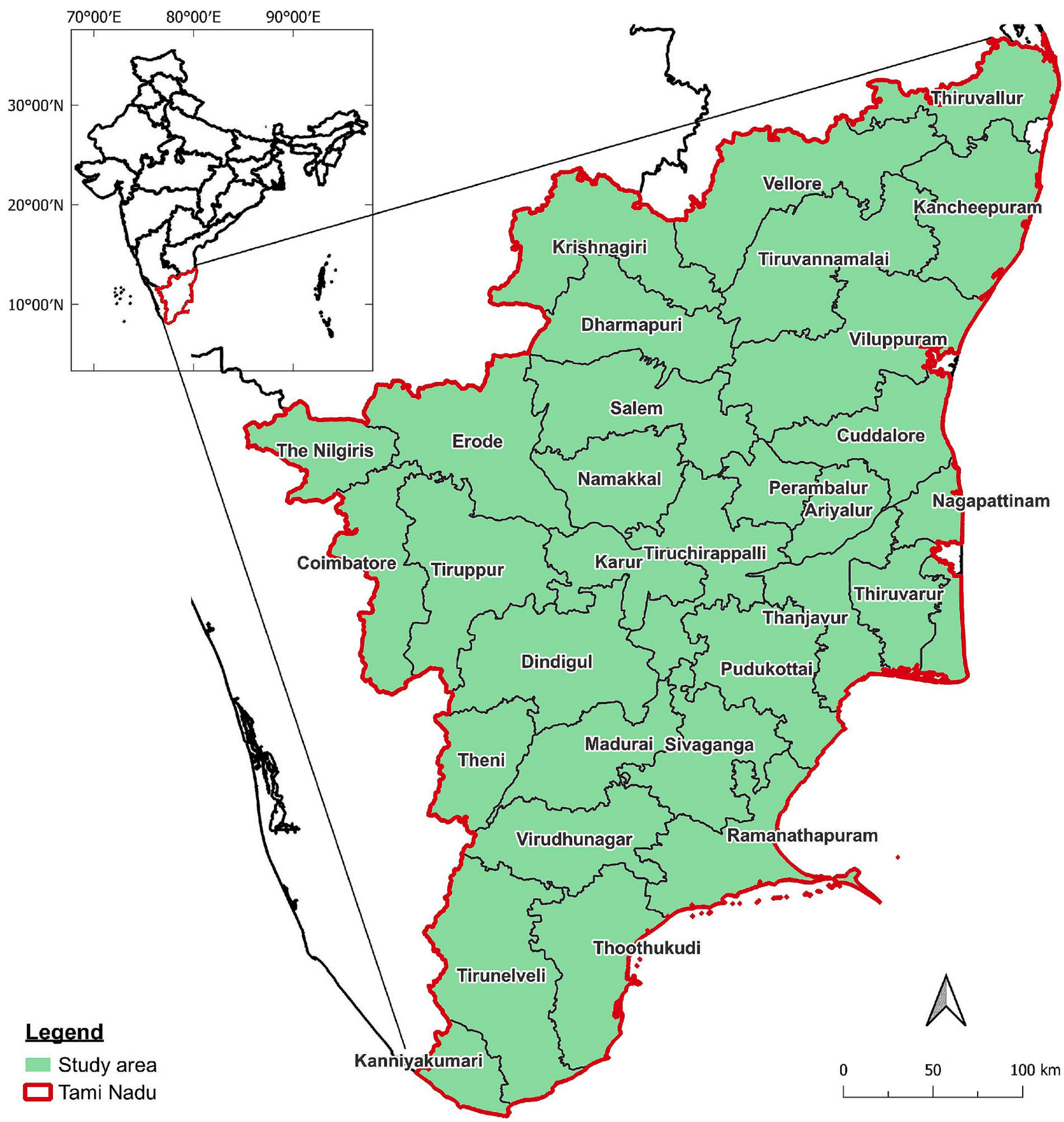

Fig. 1. Map of the study districts in Tamil Nadu, India.

SPEI is a drought index based on climatic data that can be used to determine the magnitude of drought conditions with respect to normal conditions (Vicente-Serrano et al., 2010). SPEI values ≥ 1.00 indicate wet weather conditions, with extreme wet conditions defined by values ≥ 2.00). Values ranging from -0.99 to 0.99 are indicative of near-normal conditions, whereas values ≤ -1.00 indicate drought conditions, with extreme drought conditions characterized by values ≤ -2.00 (Rhee and Im, 2017).

2.3. Data analysis

2.3.1. Agroclimatic variables

The mango season typically overlaps two calendar years in fruiting mango orchards in Tamil Nadu (Nithya Devi et al., 2019) (Table 1). In this modelling study, we focus on the months of October to December of year − 1 (hereafter referred to as Period 1 of the season) which encompasses vegetative rest (dormancy) and the beginning of flowering (bud burst), and the months of January to July of the year (hereafter referred to as Period 2 of the season). Period 2 encompasses flowering, fruit set and development and maturity (Table 1). For each mango season, the cumulative daily growing degree days (GDD; base temperature = 10 °C) (Hofman, 2011), daily rainfall and daily solar radiation were summed over Period 1 and included as potential predictors. Climate data were aggregated over Period 1 with the aim of focusing on mango yield response to climate variability during the phenological stages encompassing flowering to maturity. For each Period 2, daily GDD, rainfall, and solar radiation were temporally summed by month and included as potential predictors. The standard deviations (SD) of daily GDD, rainfall and solar radiation during each month of Period 2 (January–July) were also included to account for the variability of these meteorological factors. Monthly SPEI data for the two periods of the mango season were used as potential predictors to account for the effects of drought conditions on mango yield. Thus, in this study, 55 potential predictors were considered: Period 1 total rainfall, total solar radiation,

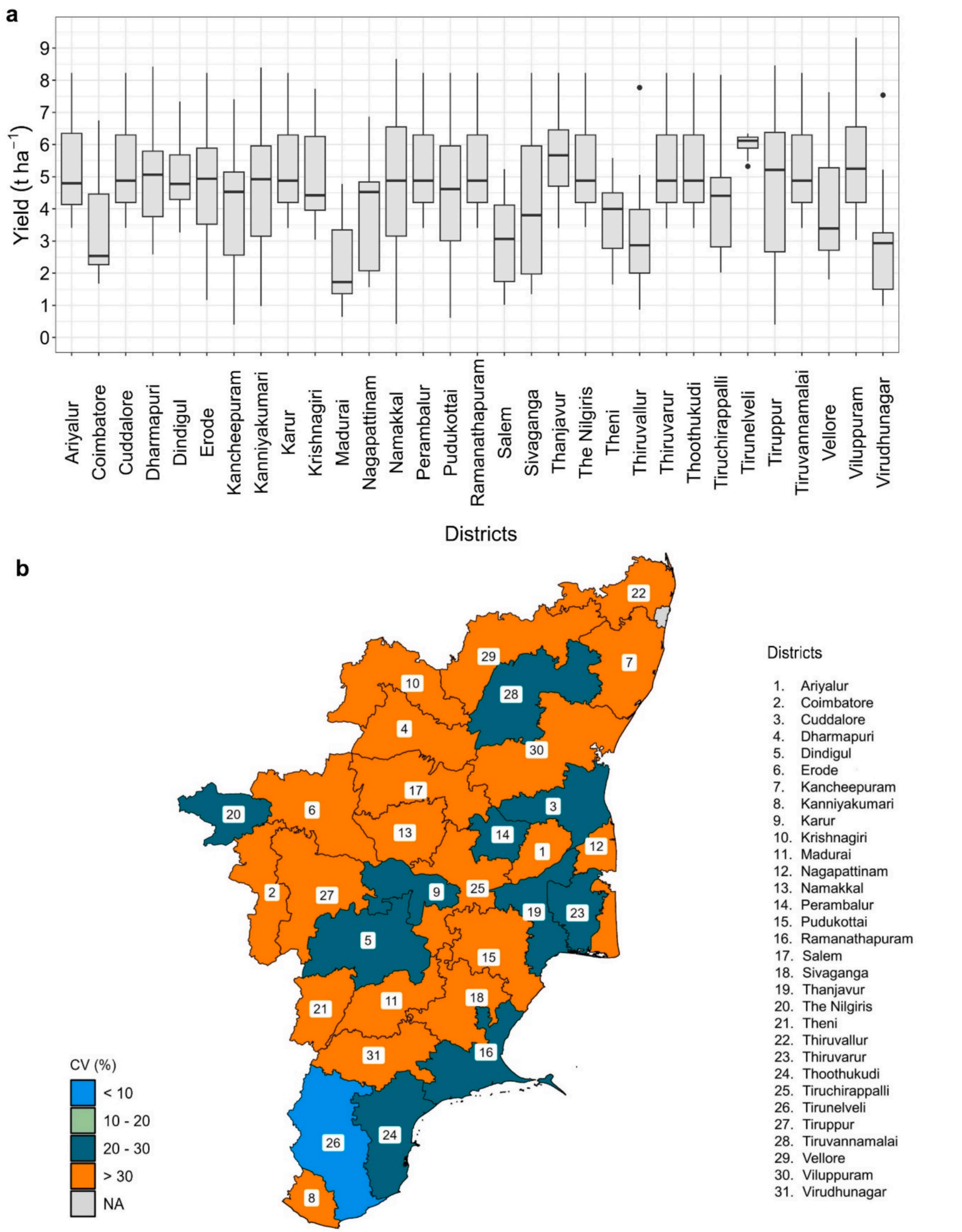

Fig. 2. The distributions of mango yields (**a**) and coefficients of variation (CV) of reported district-scale yields (**b**) in Tamil Nadu during 2008–2019. In a boxplot, the top and bottom of the box represent the 75th and 25th percentiles, respectively; the solid line indicates the median. The whiskers on the top and bottom represent the largest and smallest values within 1.5 times the interquartile range above the 3rd and 1st quartiles, respectively. Black circles are the outliers. NA: Chennai district where data was not available.

Table 1
The periods of different phenological stages for fruiting mango trees in Tamil Nadu.

Crop stage	Months	Period of mango season
Vegetative growth	August – September (Year − 1)	Period 0
Dormancy	October – November (Year − 1)	Period 1
Bud bursting	November – December (Year − 1)	Period 1
Flowering	January – February	Period 2
Fruit set	March	Period 2
Fruit development	April – May	Period 2
Maturity/Harvest	June – August	Period 2

Source: Nithya Devi et al. (2019).

cumulative GDD, and monthly SPEI, and Period 2 monthly total rainfall, monthly cumulative GDD, monthly total solar radiation, monthly SPEI, and SDs of daily GDD, rainfall and solar radiation.

Prior to the modelling, correlation analysis was performed for each district to explore the relationship between mango yield and individual agroclimatic variables. Normality tests were performed using the Shapiro-Wilk test. For variables with normally distributed data the Pearson correlation coefficient was used to quantify the degree of association; for non-normally distributed data, the Kendall's τ statistic was employed as the measure of association.

2.3.2. Mango yield models at the district scale

A comprehensive evaluation was conducted to identify the most significant predictors of mango yield at the district scale in Tamil Nadu. For each district, a multivariate regression equation (Eq. (1)) was used to model mango yield.

$$\widehat{Y}_{i,j} = \gamma_{i,0} + \gamma_{i,1} \times j + \sum_{l=1}^{n} \alpha_{i,j}^{(l)} x_{i,j}^{(l)} + \varepsilon_{i,j} \tag{1}$$

where $\widehat{Y}_{i,j}$ denotes the estimated or forecasted mango yield for district i in year j; $\Upsilon_{i,0}$ and $\Upsilon_{i,1}$ are the regression intercept and the technology trend coefficient, respectively; $x_{i,j}^{(l)}$ denotes the l predictor variables for i in year j; l could be any of the potential agroclimatic predictors; and n is the number of predictors (i.e., $n = 55$). $\alpha_{i,j}^{(l)}$ are the regression coefficients.; they are spatially and temporally varying (Newlands et al., 2014). In this study *Year* was included as predictor to incorporate the effect on yield of technology over time (i.e., increase in yield from improved management practices such as water conservation measures, soil fertility management, pruning technique), and the age of the orchards. The technology trend was assumed to be linear. The model uncertainty $\varepsilon_{i,j}$ is independent and normally distributed with mean zero and variance σ_i^2. $\Upsilon_{i,0}$ and $\Upsilon_{i,1}$ were used to detrend the yield data.

In the study there were 20 districts out of 31 with CV > 30 %, indicating very high variability in the reported mango yields (Fig. 2). Consequently, the log-transformed yield values were considered in the modelling approach.

For each district, the modelling approach involved two main stages (Fig. 3). Stage 1 involved an automatic ranking and selection of the best predictors, coupled to robust cross-validation to build each model. The automatic ranking and selection of leading potential predictors was achieved using a bootstrap robust least-angle regression scheme (B-RLARS) (Khan et al., 2007; Newlands et al., 2014). The B-RLARS is an enhancement of the least-angle regression algorithm that incorporates robustness and the bootstrap technique to improve reliability, particularly in cases where data may contain outliers or exhibit high variance (Khan et al., 2007; Newlands et al., 2014). It is suitable for high-dimensional datasets where the number of predictors is much larger than the number of observations, as it is in this study (12 years of observations, 55 potential predictors). Highly correlated predictors with correlation coefficient ≥ 0.70 were determined prior to the automatic ranking process; one variable of each pair of highly correlated variables was kept to avoid the risk of multicollinearity. Moreover, all predictors were assumed to follow a truncated normal distribution and were standardized as follows (Newlands et al., 2014): $\begin{cases} if MAD(x) = 0, x' = x - mean(x)/SD(x) \\ else, x' = x - median(x)/MAD(x) \end{cases}$, where MAD is the median absolute deviation, and SD is the standard deviation. To finalize the training of each yield model and remove any false (poor-performing) predictors, a robust cross-validation (i.e. k-fold cross-validation) was used. A maximum of four predictors was selected for each model to avoid model overfitting. The ability of the model to estimate mango yield was evaluated during this stage.

Stage 2 of the modelling approach focused on the forecasting of mango yield at different lead times (forecast dates) (Fig. 3). First, for each district, statistically neighbouring districts were identified through a Bayesian-based spatial correlation analysis (Bornn and Zidek, 2012) and the prior distributions of the best predictors (selected during Stage 1) were generated. The calibrated yield regression equation (that was built during Stage 1) was fitted jointly to the given district and its identified neighbours. The models fitted using these neighbouring districts were then cross-validated with the data from the given district. The top districts (a maximum of three districts in this study) were selected based on obtaining the lowest cross-validation error. Then, the prior distributions of the best predictors were generated using only information from the statistically selected neighbouring districts and that of the forecasting district. This approach proved effective in generating a more informative prior distribution for the model predictors by incorporating additional spatial covariance support, accounting for the residual spatial covariance among districts (Bornn and Zidek, 2012; Newlands et al., 2014).

Secondly, the prior distributions were used in a Markov chain Monte Carlo (MCMC) scheme (i.e., the Metropolis-Hastings algorithm) to estimate the posterior distributions of the predictors, along with data at the forecast time. The unobserved values of the selected predictors that are required to make the forecast were estimated using a RF technique (Breiman, 2001), coupled with a bootstrapping process. At each forecast time, the observed data after the forecast time were substituted with the model-generated data to simulate a real-time forecasting scenario. In initial analyses, multi-variate adaptive regression splines and RF algorithms were evaluated for their ability to predict these unobserved values. The random forest algorithm was chosen due to its superior computational efficiency and accuracy.

The third step in Stage 2 consisted in generating the probability distribution of mango yield forecasts using the model-generated data, combined with the available data at the forecast time generate. The main outputs of the forecasted yield probability distributions were the 10th percentile (representing the worst 10 % of the outcomes), the 50th percentile (representing the median of the outcomes) and the 90th percentile (representing the best 10 % of the outcomes). In this study, the median value of the forecasted yields was used. The forecast times selected in this study are described in section 2.3.3.

2.3.3. Lead times for mango yield forecasting

Three lead times were considered to forecast mango yield based on the start of harvest (SoH). Yield forecasts were generated on the first days of April, May, and June, which corresponded to two to one month (s) before SoH and at SoH, respectively. These months were selected based on the release schedule of official agricultural statistics in India. Crop area and production estimates are released at five points in time during a given cropping year: the first release is approximately at the end of the southwest monsoon season in September; the second is in February, the third in April-May, and the fourth in July-August, and the final estimates are released in February of the following agricultural crop year (Government of India, 2022). In this study, for each forecasting time, mango yields were forecasted using values from selected predictors corresponding to the months leading up to the forecast month, in conjunction with RF-based bootstrap estimates for the

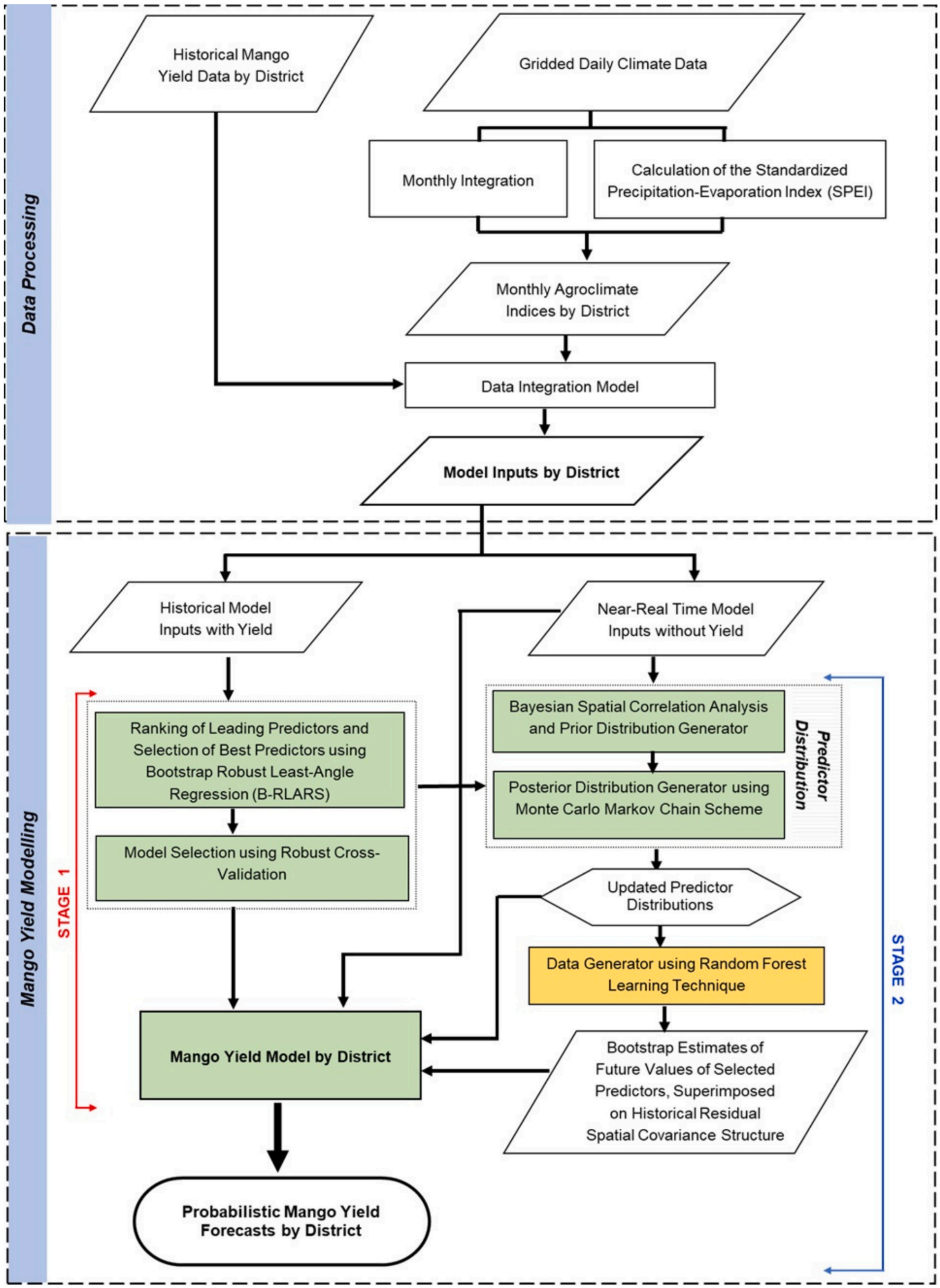

Fig. 3. A flowchart describing the modelling approach of mango yield at the district level in Tamil Nadu, India. Processes involving statistical methods are shaded in green and those involving a machine learning technique are shaded in orange.

remainder of the season (section 2.3.2). For example, if the selected predictors for a model pertain to January, February, March and June, the inputs for April forecasts would include the predictor values from January, February and March, along with the bootstrap estimates for June-related predictors. Note that for Period 1-related variables, no estimate would be generated if selected as it fell before the forecast months of April, May, and June.

2.3.4. Model performance evaluation

Leave-one-year-out-cross-validation (LOOCV) was used to evaluate the robustness of each model for estimating mango yield. In this process, for each district, a single year from the original 12-year dataset (11-year for Ariyalur) was selected as validation data, whereas the remaining observations were used as training data. This process was repeated iteratively until each year in the sample was used once as validation data. To assess the performance of the yield models after LOOCV, the Pearson coefficient of determination (R^2), the root mean square error (RMSE) and the mean absolute percentage error (MAPE) were used as statistical indicators. RMSE and the MAPE were also employed to assess the models' accuracy in forecasting mango yield.

All statistical analyses were performed using R version 4.3.1 (R Core Team, 2023) within the RStudio development environment (Posit Team, 2023). Geospatial analyses and mapping were performed using R and the Quantum Geographic Information System (QGIS; version 3.22.14) (https://qgis.org).

3. Results

3.1. Variability in climate data during the study period

Variations in climate conditions during the 2008–2019 across each mango-producing district in Tamil Nadu are presented in Figs. 4-6 and Supplementary Figs. S1-S4. Overall, there was a consistent trend in GDD between January and July across all districts. Indeed, monthly cumulative GDD decreased from January to February, followed by an increase from February to May, with a peak during this latter month in most cases; this was then followed by a decrease in June and July (Fig. 4). During 2008–2019 February had the lowest average GDD values, with values ranging from 700 °C.d to 800 °C.d (Fig. 4). Exception included The Nilgiris district where over the 2008–2019 period the average monthly cumulative GDD was below 800 °C.d. Daily GDD variability, expressed through the SDs for each month, appeared minimal (≤ 2°C·d) across all the districts, indicating relatively stable monthly thermal conditions between January and July (Fig. S1). The solar radiation (SRad) patterns, characterized by an increasing trend from January to March and a decreasing trend afterwards, remained similar across districts in Tamil Nadu during 2008–2019 (Fig. S2). Median monthly SRad values ranged from 19 MJ m^{-2} to 26 MJ m^{-2}, with March recording the

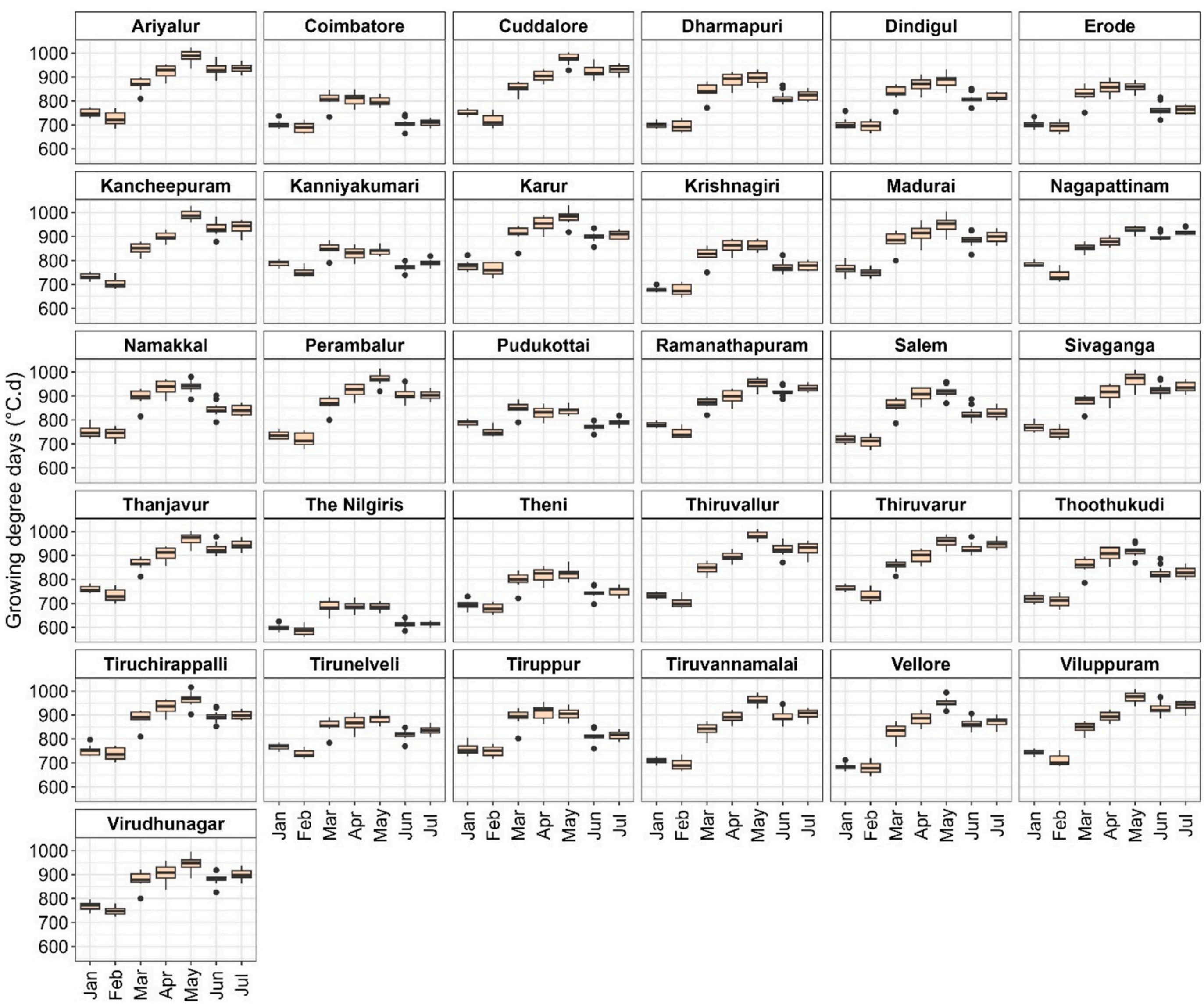

Fig. 4. Variations in the monthly cumulative growing degree days from January to July across mango-producing districts in Tamil Nadu during 2008–2019.

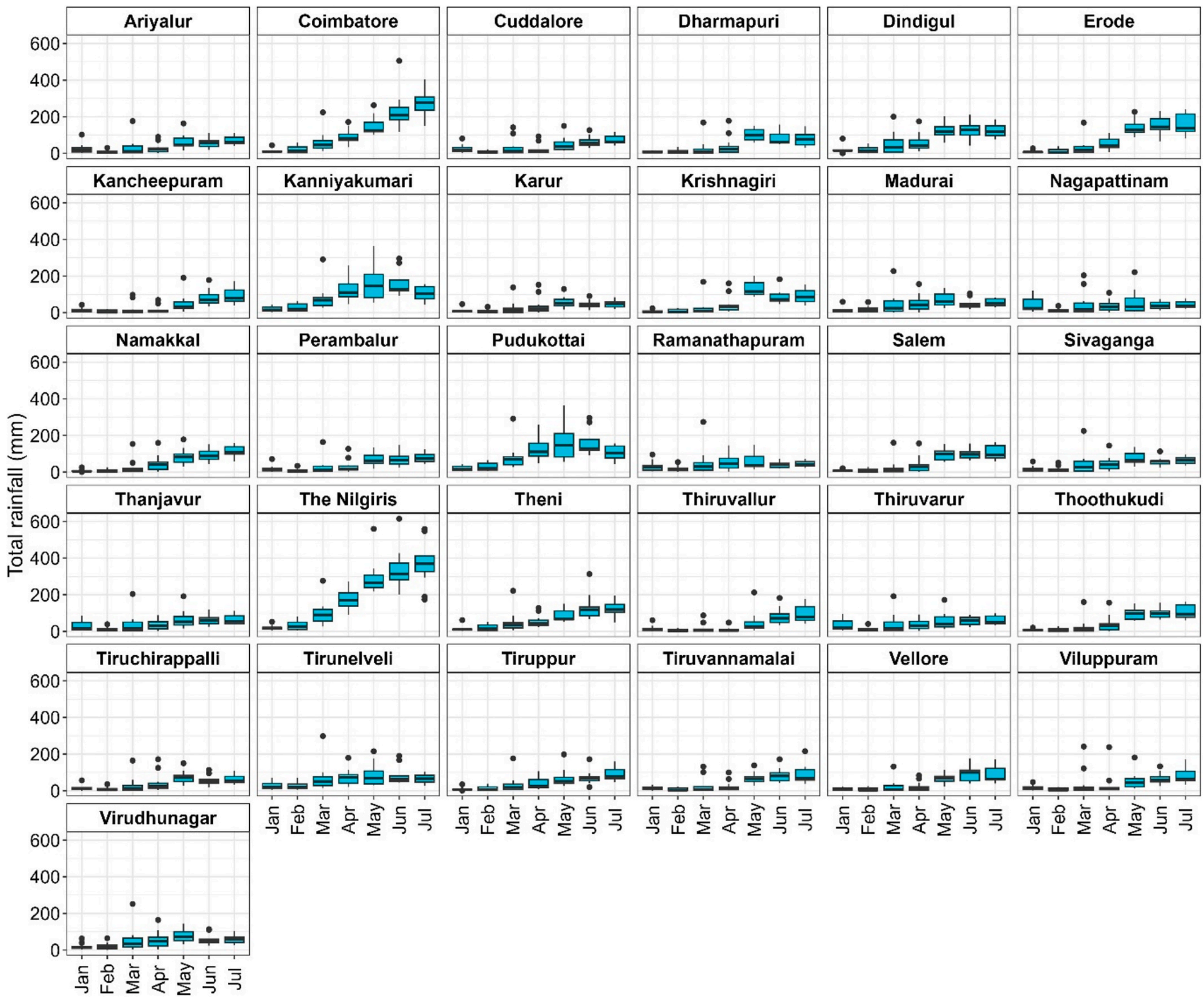

Fig. 5. Variations in the total monthly rainfall from January to July across mango-producing districts in Tamil Nadu during 2008–2019.

highest irradiance and July recording the lowest irradiance across most districts (Fig. S2). Variability in SRad during the 12-year period was low and did not exceed 0.3 MJ m^{-2} (Fig. S3).

Regarding rainfall, over the 2008–2019 period, the monthly total rainfall amount barely exceeded 200 mm in the majority of the mango-producing districts in Tamil Nadu (26 out of 31) (Fig. 5), January and February were the driest months (monthly rainfall < 50 mm in most cases. In districts such as Kancheepuram, Karur, Namakkal, Salem, Thiruvallur, Thoothukudi, Tiruvannamalai, and Viluppuram, this pattern persisted through March and April (Fig. 5). However, in Coimbatore and The Nilgiris, a marked increase in total rainfall between January and July was observed during the study period, compared to the pattern in other districts. For instance in The Nilgiris, the median monthly rainfall rose from < 5 mm in January to close to 400 mm in July (Fig. 5). In terms of variability, all the districts virtually exhibited low to moderate variability (SD ≤ 5 mm) between January and July each year of the study period (Fig. S4). Exceptions included The Nilgiris, and Coimbatore where the median rainfall standard deviations (SdRain) varied between 6 and 11 mm for the months of June and July, likely due to localized climatic events (e.g., heavy rainfall) (Fig. S4).

The analysis of drought conditions based on the SPEI indicated that most districts exhibited near normal conditions (SPEI values close to 0), with slight dryness observed in October-December and July (Fig. 6). A wider range of SPEI during the pre-flowering period, namely in November (relatively larger boxes or more outliers in Fig. 6) was also observed across the districts, suggesting higher variability in water availability during this month. In districts such as Coimbatore, Theni, Thiruvallur and Tiruppur, this pattern was also found in January/February and April/May (Fig. 6).

3.2. *Correlations between agroclimatic variables and mango yield*

Varying degrees of association between agroclimatic variables and mango yield were observed across districts (Fig. 7). The most frequently significant predictors ($p < 0.05$) included SdGDD_4 (GDD variability in April), SdGDD_6 (June), SdGDD_1 (January), SdGDD_7 (July), SRad_P1 (solar radiation for Period 1), SumGDD_2 (cumulative GDD for February), SumRain_5 (rainfall in May), SumRain_3 (March), SumSRad_4 (solar radiation in April), and SdRain_4 (rainfall variability in April). The effects of SPEI varied by month: SPEI_4 and SPEI_1 were generally positively correlated with mango yield in several districts, while SPEI_7, and to a lesser extent SPEI_6, negatively impacted yield (Fig. 7). Variability in heat accumulation (expressed as the standard deviation of daily GDD – SdGDD) generally had positive effects on

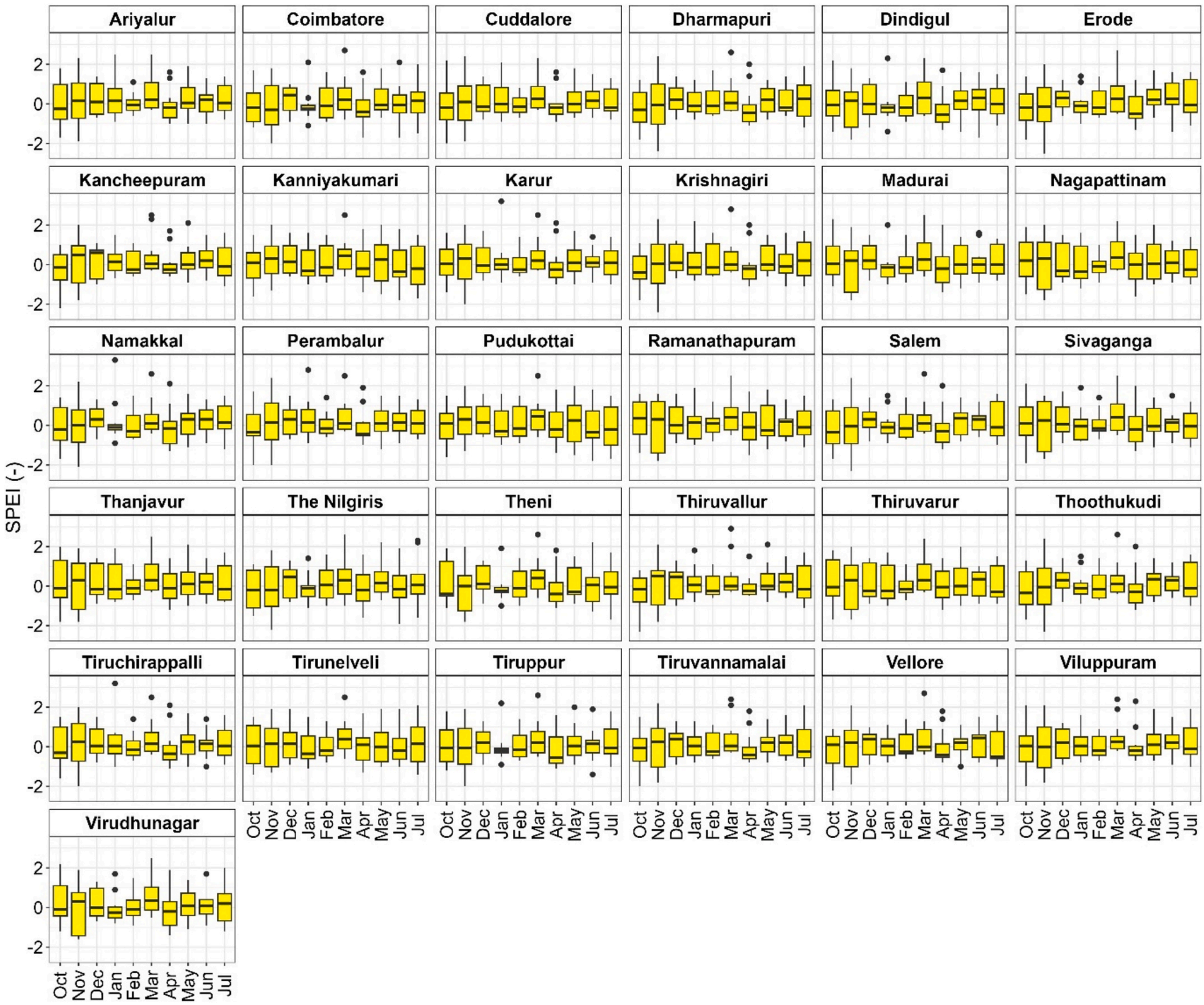

Fig. 6. Variations in the monthly standardized precipitation-evaporation index (SPEI) from October to July across mango-producing districts in Tamil Nadu during 2008–2019. SPEI values for October, November and December correspond to values for Period 1 of each mango season (year − 1).

mango yield in April, May and June, while high monthly values in GDD, rainfall, and solar radiation were often linked to yield reductions. For instance, higher cumulative GDD in April and July (SumGDD_4 and SumGDD_7), total rainfall in July (SumRain_7), and solar radiation in January and April (SumSRad_1 and SumSRad_4) often had detrimental effects on yield. The associations with variables such as SumGDD_2, SdGDD_4, and SdGDD_1 highlight that both absolute temperature sums and their variability during some months were critical factors influencing mango yield across the study regions.

3.3. Predictors explaining mango yield interannual variability

The selection of the best agroclimatic predictors of mango yield through a B-RLARS varied according to the district (Table 2, Fig. S5). The most frequently selected predictors included July, December, and March SPEI (SPEI_7, SPEI_12, SPEI_3, respectively), total solar radiation for January and June (SumSRad_1 and SumSRad_6), standard deviations of daily GDD in May, July, April, June, and February, (SdGDD_5, SdGDD_7, SdGDD_4, SdGDD_6, and SdGDD_2), total monthly rainfall for July and April (SumRain_7 and SumRain_4), and cumulative monthly GDD for May and January (SumGDD_5 and SumGDD_1) (Table 2). Variability in mango yield at the district scale was primarily driven by temperature fluctuations (expressed as standard deviations) and drought conditions (captured by the SPEI), particularly before and during bud burst (SPEI_12), and during the subsequent phenological stages (flowering, fruit set, and fruit development). The least frequently selected predictors influencing mango yield across districts included GDD_P1 and SumRain_6 (each selected only once), followed by SumRain_1 and SPEI_2 (twice), and SumRain_3, SdSRad_1, and SdRain_1 (thrice). These predictors had minimal impact compared to others, indicating limited relevance in explaining yield variations. Factors like SdRain_5, SdSRad_3, SumRain_5, and Rain_P1, with slightly higher frequencies, also contributed less consistently to mango yield across districts in Tamil Nadu (Fig. S5).

The frequency with which the most influential agroclimatic predictors were associated with either positive or negative effects on mango yield across models and districts is presented in Fig. 8. Overall, rainfall-related variables (monthly totals and variability in daily values) were predominantly linked to negative impacts on mango yield in the districts where they were identified as key predictors (Fig. 8a and e). In contrast, temperature-related predictors (cumulative GDD and variability in daily GDD) exhibited both positive and negative associations depending on

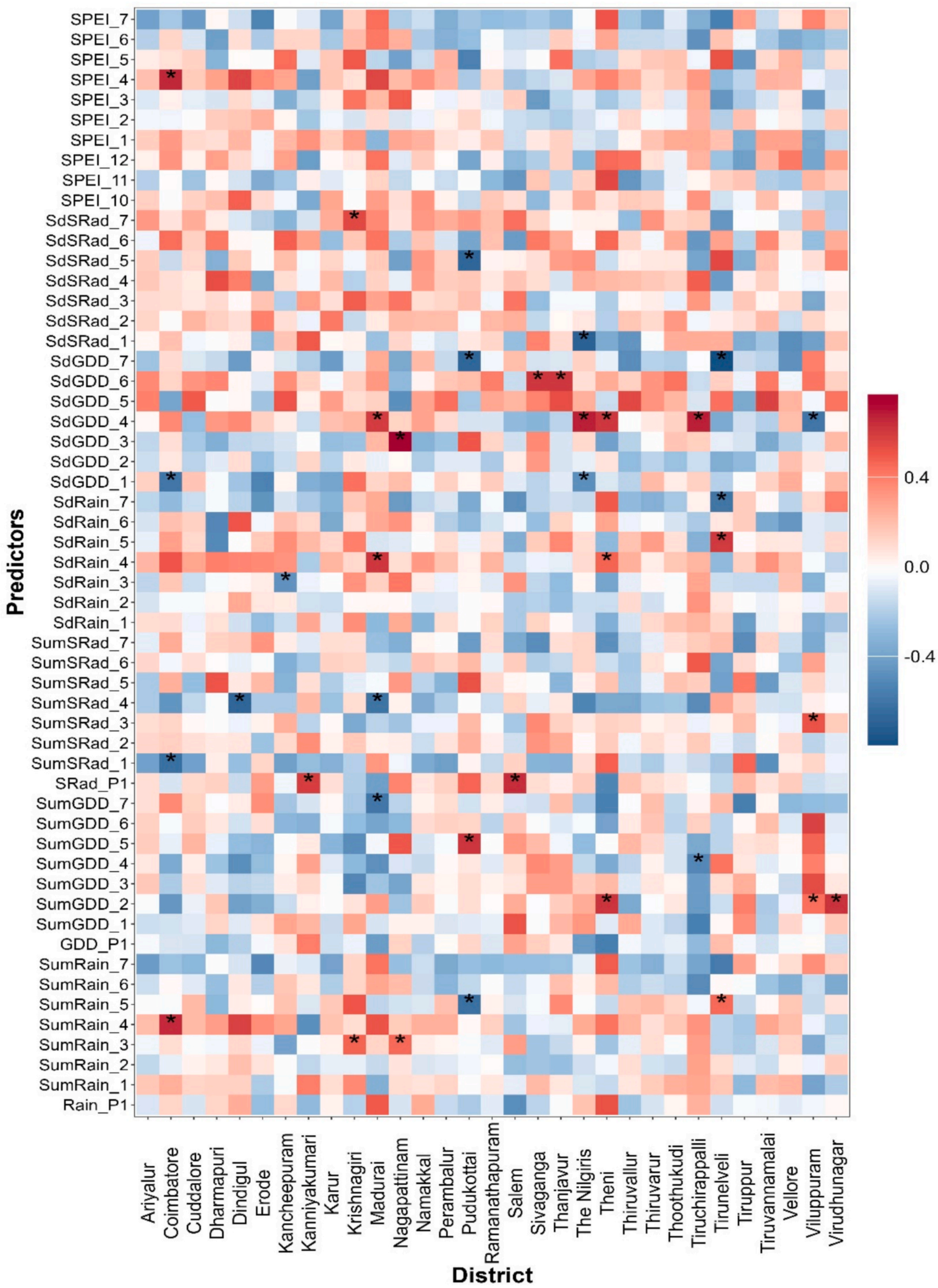

Fig. 7. Correlation coefficient analysis of all agroclimatic predictors and mango yields for the period of 2008–2019. Positive correlations are represented by red shades, while blue shades indicate negative correlations, with the intensity of colour reflecting the strength of the relationship. Statistically significant correlations ($p < 0.05$) are marked by an asterisk (*).

the month considered (Fig. 8b and f). The SPEI predictors were generally associated with negative effects across districts (Fig. 8d). Solar radiation and its variability played a more nuanced role in mango yield (Fig. 8c and g). Among the top best predictors (Table 2), SdGDD_5, SumGDD_5, SumRain_4, SdGDD_6, and SumGDD_1 most often impacted positively on mango yield across the districts where they were selected. For the remaining top best predictors, a negative effect on mango yield was generally found (Fig. 8).

3.4. Performance of yield models at the district level

The yield models performances were evaluated using LOOCV over the 12-year study period. The spatial distribution of average LOOCV R^2, RMSE and MAPE values is presented in Fig. 9. In 18 out of 31 districts, the models explained between 20 % and 70 % of the variability in district-level mango yields (Fig. 9a). Average RMSE values reached up to 2.0 t ha^{-1} in 19 districts (Fig. 9b). In terms of MAPE, these districts achieved good to acceptable accuracy, with MAPE ≤ 30 % (Fig. 9c). Theni (south-west) was the only district among the 19 to have a MAPE

Table 2
Top agroclimatic variables selected as best predictors for mango yield models at the district scale.

Rank	Predictors	Definition
1	SPEI_7	July standardized precipitation-evaporation index (SPEI)
2	SumSRad_1	Total solar radiation for January
2	SdGDD_5	Standard deviation of daily GDD in May
2	SPEI_12	December SPEI
3	SumRain_7	Total rainfall for July
4	SumGDD_5	Cumulative GDD for May
4	SdGDD_7	Standard deviation of daily GDD in July
5	SumRain_4	Total rainfall for April
5	SdGDD_4	Standard deviation of daily GDD in April
5	SdGDD_6	Standard deviation of daily GDD in June
5	SumGDD_1	Cumulative GDD for January
6	SdGDD_2	Standard deviation of daily GDD in February
6	SumSRad_6	Total solar radiation for June
6	SPEI_3	March SPEI

above 30 %. Conversely, in districts such as Kanniyakumari (south), Tiruppur (west), and Namakkal (central-west), Ariyalur (central-east), and Kancheepuram (north-east), notable RMSE values exceeding 3.0 t ha^{-1} were found (Fig. 9b). These districts, along with others with lower R^2, often exhibited higher MAPE (> 30 %), indicating poor model performance. Poor-performing districts were primarily located in the western, central-eastern and north-eastern parts of the state (Fig. 9c), highlighting regions where model refinement and additional data collection could enhance accuracy.

3.5. Evaluation of the model performance in forecasting mango yields at different lead times

The spatial and temporal variability in the skills of yield models used to forecast mango yield at the district scale was evaluated for three lead times: April (two months before the start of harvest, SoH), May (one month before SoH), and June (SoH). Analysis of the mean yield deviations at these lead times (Fig. 10) revealed relatively small deviations from observed mango yields for most districts, as indicated by narrow interquartile ranges (IQRs) tightly clustered around 0 % (Fig. 10). Tiruppur consistently exhibited the highest deviations across all three months. Other districts with higher yield deviations, evidenced by wider IQRs and/or more outliers, included Erode, Kancheepuram, Namakkal, Pudukkottai, and Thiruvallur, suggesting poor model performance in these areas (Fig. 10).

In April and May, southern and coastal districts had the lowest RMSE ($\leq$ 2.0 t ha^{-1}) (Fig. 11a and b). This trend persisted in June for most coastal and southern regions, whereas northern districts experienced higher errors (RMSE > 4.0 t ha^{-1}). Regarding MAPE, most districts showed moderate to high errors (> 20 %) across the three forecasting months. Similar to RMSE, southern and coastal districts demonstrated superior performance in predicting mango yield two months prior to and up to the SoH (Fig. 11). In Tirunelveli (south) mango yield was predicted with a RMSE of 0.53 t ha^{-1} (< 1.0 t ha^{-1}) and a MAPE of 7 % during this period (Fig. 11). Other districts with consistently low RMSE ($\leq$ 2.0 t ha^{-1}) and acceptable MAPE ($\leq$ 30 %) across months included Dharmapuri and Krishnagiri (north-west) (Fig. 11).

4. Discussion

Multivariate regression models based on agroclimatic indicators, including monthly cumulative GDD, rainfall, and solar radiation, and SPEI, were developed to forecast mango yields at the district scale in Tamil Nadu, India. Our modelling approach combined statistical methods (B-RLARS, LOOCV, Bayesian-based spatial correlation analysis, and MCMC scheme) and ML-based modelling (i.e., RF), offering a robust and complementary framework for yield forecasting. The statistical methods used in the approach ensured thorough variable selection, identification of spatial correlations and regional patterns in yield responses, and uncertainty quantification. Meanwhile, RF modelling facilitated a stable prediction of the unobserved values of selected predictors required for the forecasts.

4.1. The need for comprehensive understanding of the climate drivers affecting mango yield

The most influential variables influencing mango yield at the district level in Tamil Nadu included SPEI for July, December, and March, total

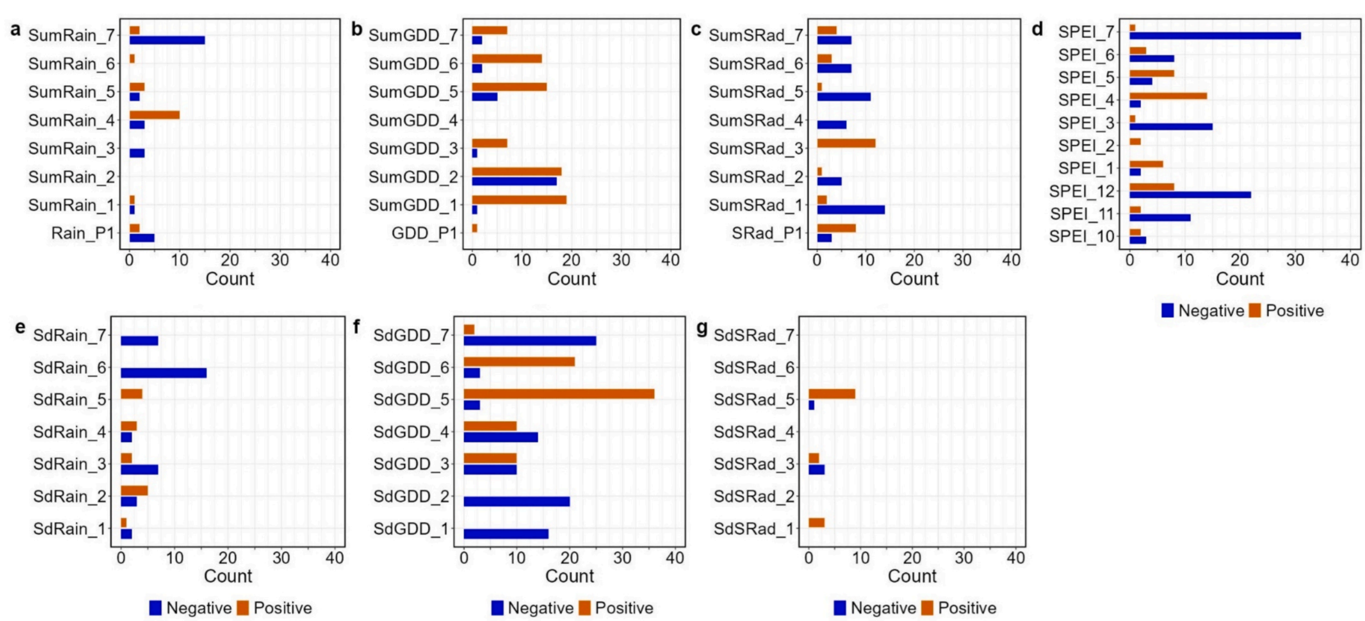

Fig. 8. Frequency distribution of the effects of most influential agroclimatic predictors on mango yield across models and districts in Tamil Nadu. Positive (orange) and negative (blue) effect as indicated by the sign of the coefficient estimate are depicted for (**a**), (**b**), (**c**): Period 1 total and Period 2 monthly totals for rainfall, growing degree days (GDD), and solar radiation, respectively; (**d**) monthly SPEI; (**e**), (**f**), (**g**): Period 2 standard deviations of daily rainfall, GDD, and solar radiation, respectively.

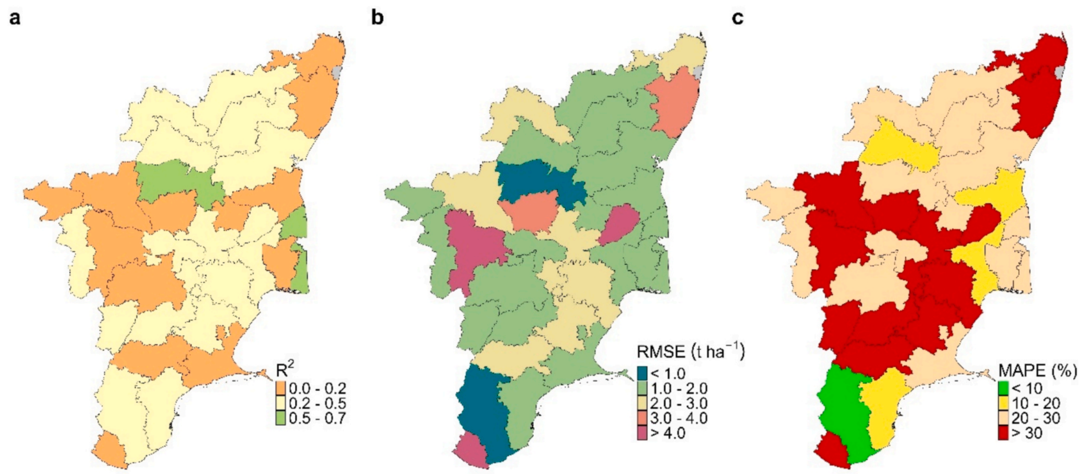

Fig. 9. Performance of mango yield models at the district scale in Tamil Nadu, India. Leave-one-year-out-cross-validation average (**a**) Pearson coefficient of determination (R^2), (**b**) root mean square error (RMSE), and (**c**) mean absolute percentage error (MAPE).

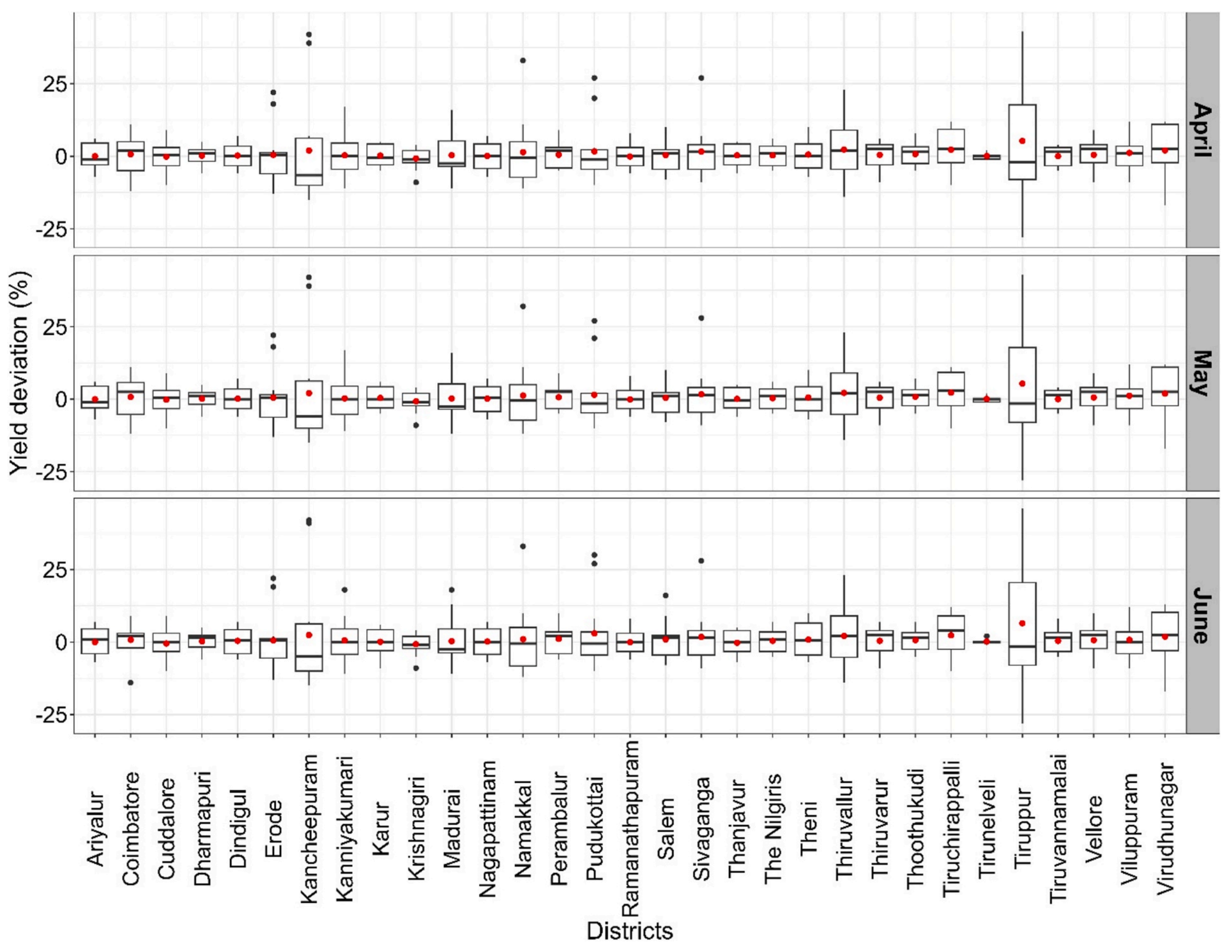

Fig. 10. Variability in mango yield deviations across districts for the three lead months April, May, and June. Yield deviations were calculated as the differences between the predicted and observed yield, expressed as a percentage of the observed yield, for each lead month over the 12-year study period.

solar radiation for January and June, standard deviations of daily GDD for May, July, April, June, and February, total monthly rainfall for July and April, and cumulative monthly GDD for May and January. The period from January to June typically encompasses the flowering, fruit set and fruit development stages of mangoes across the study area (Nithya Devi et al., 2019). Our results suggest that weather conditions during these periods remain critical in determining mango yield. The selection of SPEI variables for December (SPEI_12) and March (SPEI_3) underscores the importance of water availability during bud burst and fruit set. These two variables mostly had negative effect on mango yield across districts, indicating that drought stress during flowering and fruit set could be detrimental to mango yield. This aligns with findings by Carr (2014) and Zuazo et al. (2021), where water stress during bud burst, flowering and fruit growth stages impacted mango yields as it can

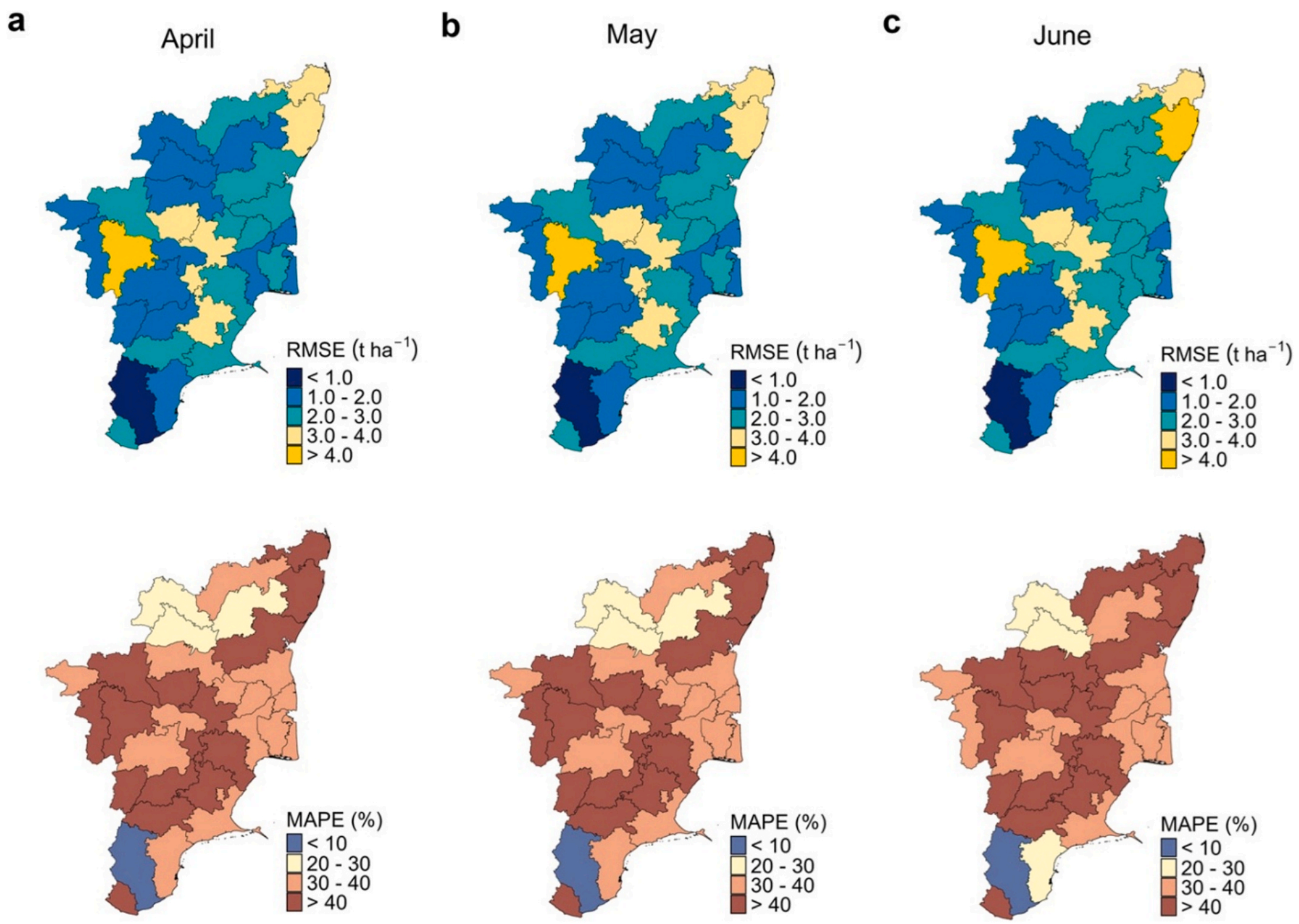

Fig. 11. The spatial distribution of forecasts errors, root mean square error (RMSE; upper row) and mean absolute percentage error (MAPE; lower row), at different lead times. (**a**) 1st of April (2 months before start of harvest (SoH)), (**b**) 1st of May (1 month before SoH), and (**c**) 1st of June (SoH). For a given month, the models' performances were calculated based on the predicted mango yields for that month over the 12-year period (11 years for Ariyalur).

delay the development of vegetative buds and can affect fruit retention and size. However, the response degree varies depending on the environmental conditions, i.e., tropical or sub-tropical regions (Zuazo et al., 2021). Indeed, in tropical regions, a short period of water stress helps trigger flowering, whereas in subtropical areas, it plays a minor role, as low winter temperatures naturally limit growth and floral induction (Zuazo et al., 2021). A negative effect of drought conditions during July (SPEI_7) on mango yield was found through the modelling. Water stress during maturity can lead to smaller fruit due to limited translocation of photosynthates to the fruit, or increase fruit drops (Liu et al., 2023). Over the study period 2008–2019 most districts did not experience severe drought conditions in July (Fig. 6). Thus, the negative effect on mango yield of drought conditions in that month suggests that this influence was exacerbated in association with other factors.

Total solar radiation in January and June (SumSRad_1 and SumSRad_6) and the cumulative GDD for May and January (SumGDD_5 and SumGDD_1) reflect the critical role of energy inputs during flowering and fruit development, corroborating observations from previous research (e.g., Geetha et al., 2016; Clonan et al., 2021). Indeed, it has been shown that temperature plays a major role in flowering in mangoes: from triggering flowering under cool temperatures (low nighttime temperatures in the 15–20 °C range), inhibiting it under high daytime temperatures (range of 32–35 °C), or through its impact on the production of hermaphrodite flowers under temperatures varying between 15 °C and 17 °C, depending on the varieties (Clonan et al., 2021). Temperature variability, represented by the standard deviations of GDD (e.g., SdGDD_1, SdGDD_2, SdGDD_4, SdGDD_5), emerged as an important factor for managing temperature-related risks in mango in Tamil Nadu. While previous research has focused primarily on temperature impacts on mango yield (e.g., Clonan et al., 2021; Rojas-Sandoval and Acevedo-Rodríguez, 2022), the variability in heat accumulation may have a notable influence as well, as shown in this study. Indeed, while adequate heat accumulation is essential for fruit development and maturation, excess heat or erratic temperature fluctuations—especially during flowering and fruit set— can be detrimental to yield at harvest. Strategies such as adjusting pruning and flowering induction schedules, growing heat-tolerant varieties can be adopted to cope with the adverse impact of temperature variability on mango yield.

Rainfall predictors, such as SumRain_7 and SumRain_4, exhibited a negative and positive effect on mango yield, respectively (Fig. 8a), consistent with e.g., Rojas-Sandoval and Acevedo-Rodríguez (2022), who reported variable influence of rainfall on mango yield depending on timing and intensity. Erratic rainfall patterns during fruit set or late in the season tended to have a predominantly negative influence across districts, suggesting that they may have disrupted water balance and impair fruit development. This is particularly important in Tamil Nadu, where mango orchards are predominantly rainfed. The use of deficit irrigation during periods of inadequate rainfall – when economically feasible – may help sustain yield levels (Zuazo et al., 2021). Characterizing the main climate drivers of mango yield is crucial for developing adaptive measures to maintain or increase yield levels and ensure the sustainability and profitability of farming activities. The identification of the best predictors of mango yield in this study can serve as a foundation for future research to develop on-farm adaptation strategies to manage climate risks and improve mango yield.

4.2. Performance of yield models and challenges

The spatial variability in model performance revealed distinct patterns (Figs. 8 and 9). Poor-performing districts, predominantly located in the western, central-eastern, and north-eastern regions, had lower R^2 and higher MAPE values. These areas likely faced more complex interactions between climatic factors, topography and management practices. In forecasting yield, acceptable model errors (RMSE $\leq$ 2.0 t ha^{-1} and MAPE $\leq$ 20 %) were observed in four districts (Fig. 11). For 15 out of 31 districts, the forecasting errors (MAPE) consistently ranged from 30 % to 40 % across the lead months. Studies on mango yield prediction using ML-based techniques have shown a range of prediction errors. For instance, the normalized mean absolute error for the best-performing model for mango yield estimation at the farm scale reached up to 43 % (with a minimum of 0.8 %), and up to 52 % (with a minimum of 31 %) at the orchard block scale (Torgbor et al., 2023). Our results exhibited comparable model performance. However, the subpar performance of yield models in most districts may stem from several factors, including high interannual yield variability, a limited pool of potential predictors, or the ML algorithm utilized.

The spatial patterns of model errors closely followed the coefficient of variation (CV) of reported yields (Fig. 2b). Good model performance was found in district with low yield CVs. Conversely, models generally performed poorly in districts with high yield CVs (> 30 %), even after utilizing log-transformed yields to reduce variability. The limited historical data on mango orchards (only 12 years in this study), coupled with data collection limitations, may have impeded the development of accurate predictive models. Mango yield variability across districts is influenced by numerous factors such as climate, soil composition and fertility, geographic location, local agricultural practices, the mango variety, tree growth disparities, fruit production, nutrient requirements, and resistance to pests and diseases (Rojas-Sandoval and Acevedo-Rodríguez, 2022). These factors can cause notable fluctuations in growth patterns and fruiting across orchards and districts, leading to substantial year-to-year yield variability.

Additionally, the complex interplay of endogenous factors, including tree age, health, and variety, adds another layer of challenge in developing accurate yield models. An alternative approach could involve conducting analyses at a finer spatial administrative scale (e.g., sub-district level) and then aggregating the results at the district level. This would involve using weight-averaging of all block-level predicted or forecasted yields based on harvested areas and selecting the median or average as the district-level yield (Chipanshi et al., 2015). Incorporating weekly or fortnightly climatic data instead of monthly data can help to better capture rapid fluctuations in weather conditions.

As data become available, the data period would get longer, which will ensure accurate yield predictions without risking overfitting the models. Data collection could be improved using scalable and cost-effective tools for monitoring mango production. These include freely available satellite imagery (e.g., Sentinel-2, Landsat), low-cost unmanned aerial systems, coupled with photogrammetry tools to obtain high-resolution orchard level data, partnering with local initiatives or agribusinesses that already have data applications for farmer outreach.

4.3. The challenges in forecasting mango yield at the district scale

In this study, we employed a modelling approach similar to that described in previous research, e.g., Newlands et al. (2014), Chipanshi et al. (2015), and Kouadio et al. (2021), which has demonstrated strong performance across various contexts, including different crops and countries. For example, Kouadio et al. (2021) assessed this approach for estimating and forecasting robusta coffee yield at the farm scale in Vietnam. The MAPE of the models varied between 8 % and 13 % for the estimates and between 13 % and 19 % for the forecasts (Kouadio et al., 2021). Although the models' performances in this study performed weaker than those for robusta coffee, the findings provide valuable insights into the replicability of this approach for perennial crops like mango.

The analysis of forecast results revealed an increase in the proportion of districts with poor model performance in June (Fig. 11), suggesting that late-season climatic variability likely play a crucial role. Incorporating predictors such as flowering-to-harvest temperature or rainfall variability, and/or using weekly or fortnightly climatic data instead of monthly data to better capture rapid weather fluctuations could enhance model accuracy. Fine-tuning region-specific models to account for unique agroclimatic conditions and phenological stages may reduce errors and improve prediction reliability. However, relying solely on agroclimatic variables may not suffice for forecasting mango yield at the district level in Tamil Nadu. In this study, only a limited number of districts produced acceptable model performance. This limitation could hinder the adoption of these models by industries and other stakeholders for mango yield forecasting in the region. Therefore, there is a need for further research to improve the performance of the models. Potential directions for future research may include: (i) incorporating additional data such as management practices (e.g., pruning and canopy management, fertilizer application, pests and diseases control), socioeconomic factors (e.g., farmer demographics, labour availability), soil data (e.g., soil types, water retention capacity, pH, salinity) to capture the effects of soil conditions on mango yield, and remote sensing derived information and/or vegetation indices to track mango orchard health dynamically; (ii) implementing ensemble ML or deep learning models to capture non-linear relationships and interactions among variables; (iii) engaging with mango processing industries and exporters to access additional datasets to improve the model building process and validate model outputs with their historical records.

Another avenue for model improvement might be developing process-based models capable of representing mango growth and development processes under varying environmental conditions and agricultural set-ups which can be used at such larger spatial scales (Léchaudel et al., 2005; Shahhosseini et al., 2021), and integrating them with ML-based models (e.g., von Bloh et al., 2024). Process-based models generally account for the interactions between crop, environment and management practices and are typically applied at the tree or plot or farm levels. For example, Léchaudel et al. (2005) developed a process-based model that accounted for the main processes of fruit growth such as photosynthesis, fruit demand, fruit respiration, and storage and mobilization of leaf and stem reserves, to simulate the response of mango fruit growth to weather at the plot scale. The model was subsequently extended to an architectural development model that characterized mango growth and developmental processes of inflorescence and fruiting at the tree scale across successive growth cycles (Boudon et al., 2020). Such models could be tested and integrated with ML models to leverage the strength of each.

5. Conclusions

We developed a district-scale mango yield forecasting approach combining statistical methods—including bootstrap robust least-angle regression, leave-one-year-out cross-validation, Bayesian-based spatial correlation analysis, and Markov Chain Monte Carlo scheme, and—a random forest technique, using Tamil Nadu, India as a case study. Unlike previous studies that primarily focused on farm-level assessments or models at the state/national scale, our research bridges this gap by providing district-level predictions, a critical yet underexplored scale for regional agricultural planning and market decision-making. The results showed that temperature fluctuations, rainfall distribution, pre-flowering drought stress, and drought conditions (expressed through the SPEI) in March and July were among the strongest determinants of mango yield variation. The study also revealed notable spatial disparities in model performance, with southern and north-western districts exhibiting lower forecasting errors compared to northern and central regions. Despite the models' ability to capture key drivers of yield, forecasting errors remained above 30 % in the majority of districts, underscoring the challenges of district-scale yield prediction due to environmental heterogeneity and data limitations. These findings emphasize the need to integrate additional explanatory variables, such as orchard management practices, soil properties, and remote sensing-derived indicators, to improve model accuracy and robustness. While our investigation demonstrates the potential of agroclimatic indicators for mango yield forecasting, it also highlights key limitations, particularly data availability. A more comprehensive understanding of the climate – mango yield relationships, as identified in this study, is essential for refining the modelling approach through adjustment in the ML-based forecast algorithm between districts. By addressing these gaps, our research lays the foundation for more sophisticated, data-driven forecasting approaches at finer spatial scales. Future advancements in machine learning, combined with improved data accessibility and training, will be crucial in enhancing forecast reliability and supporting climate-resilient mango production strategies.

CRediT authorship contribution statement

Louis Kouadio: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Bhuvaneswari Kulanthaivel:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Thanh Mai:** Writing – review & editing, Visualization, Methodology. **Thong Nguyen-Huy:** Writing – review & editing, Investigation. **Qingxia (Jenny) Wang:** Writing – review & editing. **Vivekananda M. Byrareddy:** Writing – review & editing. **Shahbaz Mushtaq:** Writing – review & editing, Project administration, Funding acquisition. **Alagarswamy Senthil:** Writing – review & editing, Project administration, Funding acquisition. **Nathaniel K. Newlands:** Writing – review & editing, Software, Methodology, Conceptualization. **Vellingiri Geethalakshmi:** Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was financially supported by the Australian Government Department of Industry, Science and Resources through the Australia-India Strategic Research Fund (AISRF) Round 12 (grant number DISER_AIRXII000075). NKN was supported by funding from the Canadian Sustainable Agriculture Partnership (CSAP) Program (Agriculture and Agri-Food Canada). We thank Dr S. Mohan Kumar from Tamil Nadu Agricultural University, India, for assisting in climate data curation.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2025.110450.

Data availability

The date used in this study can be accessed at https://doi.org/10.5281/zenodo.15208083.

References

Anderson, N.T., Walsh, K.B., Wulfsohn, D., 2021. Technologies for forecasting tree fruit load and harvest timing—From ground, sky and time. Agronomy 11, 1409. https://doi.org/10.3390/agronomy11071409.

Beguería, S., Vicente-Serrano, S., 2023. SPEI: calculation of the standardized precipitation-evapotranspiration index. R Package Version 1 (8), 1. https://CRAN.R-project.org/package=SPEI.

Bornn, L., Zidek, J.V., 2012. Efficient stabilization of crop yield prediction in the Canadian Prairies. Agric. for. Meteorol. 152, 223–232. https://doi.org/10.1016/j.agrformet.2011.09.013.

Boudon, F., Persello, S., Jestin, A., Briand, A.-S., Grechi, I., Fernique, P., et al., 2020. V-Mango: a functional–structural model of mango tree growth, development and fruit production. Ann. Bot. 126, 745–763. https://doi.org/10.1093/aob/mcaa089.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Carr, M.K.V., 2014. The water relations and irrigation requirements of mango (*Mangifera indica* L.): a review. Exp. Agric. 50, 1–23. https://doi.org/10.1017/S0014479713000288.

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., et al., 2015. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric. for. Meteorol. 206, 137–150. https://doi.org/10.1016/j.agrformet.2015.03.007.

Clonan, M., McConchie, C., Hall, M., Hearnden, M., Olesen, T., Sarkhosh, A., 2021. Effects of ambient temperatures on floral initiation in Australian mango (*Mangifera indica* L.) selections. Sci. Hortic. 276, 109767. https://doi.org/10.1016/j.scienta.2020.109767.

FAOSTAT, 2023. Crops and livestock products. Food and Agriculture Organization of the United Nations (FAO). Rome, Italy.

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., Müller, J., 2013. Random Forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. Agric. Water Manag. 116, 142–150. https://doi.org/10.1016/j.agwat.2012.07.003.

Gao, C., Jiang, H., Liu, X., Li, H., Wu, Z., Sun, X., et al., 2024. Improved binocular localization of kiwifruit in orchard based on fruit and calyx detection using YOLOv5x for robotic picking. Comput. Electron. Agric. 217, 108621. https://doi.org/10.1016/j.compag.2024.108621.

Geetha, G.A., Shivashankara, K.S., Reddy, Y.T.N., 2016. Varietal variations in temperature response for hermaphrodite flower production and fruit set in mango (*Mangifera indica* L). South Afr. J. Bot. 106, 196–203. https://doi.org/10.1016/j.sajb.2016.07.011.

Government of India, 2022. Agricultural statistics at a glance 2021. Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Government of India. Available at https://eands.dacnet.nic.in/PDF/Agricultural%20Statistics%20at%20a%20Glance%20-%202021%20(English%20version).pdf. (Accessed 12 December 2023).

Government of Tamil Nadu, 2022. Season and crop report 2021-22. FASLI 1431. Report No. 6/2023. Department of Economics and Statistics, Government of Tamil Nadu. Available at https://des.tn.gov.in/en/node/346. (Accessed 12 December 2023).

He, L., Fang, W., Zhao, G., Wu, Z., Fu, L., Li, R., et al., 2022. Fruit yield prediction and estimation in orchards: a state-of-the-art comprehensive review for both direct and indirect methods. Comput. Electron. Agric. 195, 106812. https://doi.org/10.1016/j.compag.2022.106812.

P. Hofman Hofman, P., 2011. Development of best practice pre and postharvest protocols for production of Calypso mango: Phase 2. Project Number: MG06005. Horticulture Australia Ltd, Sydney, NSW, Australia. Available online: https://www.horticulture.com.au/globalassets/hort-innovation/historic-reports/development-of-best-practice-pre-and-post-harvest-protocols-for-production-of-calypso-mango—phase-2-mg06005.pdf. (Accessed 6 October 2023).

Khan, J.A., Van Aelst, S., Zamar, R.H., 2007. Robust linear model selection based on least angle regression. J. Am. Stat. Assoc. 102, 1289–1299. https://doi.org/10.1198/016214507000000950.

Koirala, A., Walsh, K.B., Wang, Z., McCarthy, C., 2019. Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of 'MangoYOLO'. Precis. Agric. 20, 1107–1135. https://doi.org/10.1007/s11119-019-09642-0.

Kouadio, L., Byrareddy, V.M., Sawadogo, A., Newlands, N.K., 2021. Probabilistic yield forecasting of robusta coffee at the farm scale using agroclimatic and remote sensing derived indices. Agric. for. Meteorol. 306, 108449. https://doi.org/10.1016/j.agrformet.2021.108449.

Léchaudel, M., Génard, M., Lescourret, F., Urban, L., Jannoyer, M., 2005. Modeling effects of weather and source–sink relationships on mango fruit growth. Tree Physiol. 25, 583–597. https://doi.org/10.1093/treephys/25.5.583.

Liu, X., Gao, T., Liu, C., Mao, K., Gong, X., Li, C., et al., 2023. Fruit crops combating drought: physiological responses and regulatory pathways. Plant Physiol. 192, 1768–1784. https://doi.org/10.1093/plphys/kiad202.

Newlands, N.K., Zamar, D.S., Kouadio, L.A., Zhang, Y., Chipanshi, A., Potgieter, A., et al., 2014. An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. Front. Environ. Sci. 2. https://doi.org/10.3389/fenvs.2014.00017.

Nithya Devi, A., Arulmozhiyan, R., Balamohan, T.N., 2019. Induction of off-season flowering in mango through farmer's participatory research in Krishnagiri district under NAIP. J. Pharmacogn. Phytochem. 8, 2466–2471.

Posit Team, 2023. RStudio: Integrated development environment for R. Posit Software, PBC, Boston, MA. http://www.posit.co/.

R Core Team, 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.

Ray, S., Bhattacharyya, B., Dasyam, R., 2016. An empirical investigation of ARIMA and GARCH models in forecasting for horticultural fruits in India. Green Farming Int. J. 7, 233–236.

Ray, S., Mishra, P., Ayad, H., Kumari, P., Sharma, R., Kumari, B., et al., 2023. Prediction of fruit production in India: an econometric approach. J. Hortic. Res. 31, 25–34. https://doi.org/10.2478/johr-2023-0005.

Rhee, J., Im, J., 2017. Meteorological drought forecasting for ungauged areas based on machine learning: using long-range climate forecast and remote sensing data. Agric. for. Meteorol. 237–238, 105–122. https://doi.org/10.1016/j.agrformet.2017.02.011.

Rojas-Sandoval, J., Acevedo-Rodríguez, P., 2022. Mangifera indica (mango). CABI Compendium, CABI.

Santos, T.T., de Souza, K.X.S., Camargo Neto, J., Koenigkan, L.V., Moreira, A.S., Ternes, S., 2024. Multiple orange detection and tracking with 3-D fruit relocalization and neural-net based yield regression in commercial sweet orange orchards. Comput. Electron. Agric. 224, 109199. https://doi.org/10.1016/j.compag.2024.109199.

Santosha, R., Vijayakumar, S., Nirmala, B., Gayathri, C., 2022. Prediction of mango production using machine intelligence techniques: a case study from Karnataka, India. Acta Sci. Agric. 6, 16–22.

Sarron, J., Malézieux, É., Sané, C.A.B., Faye, É., 2018. Mango yield mapping at the orchard scale based on tree structure and land cover assessed by UAV. Remote Sens. 10, 1900. https://doi.org/10.3390/rs10121900.

Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci. Rep. 11, 1606. https://doi.org/10.1038/s41598-020-80820-1.

Stone, R.C., Meinke, H., 2005. Operational seasonal forecasting of crop performance. Philos. Trans. Roy. Soc. B Biol. Sci. 360, 2109–2124. https://doi.org/10.1098/rstb.2005.1753.

Torgbor, B.A., Rahman, M.M., Brinkhoff, J., Sinha, P., Robson, A., 2023. Integrating remote sensing and weather variables for mango yield prediction using a machine learning approach. Remote Sens. 15, 3075. https://doi.org/10.3390/rs15123075.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agric. 177, 105709. https://doi.org/10.1016/j.compag.2020.105709.

Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.I., 2010. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J. Climate 23, 1696–1718. https://doi.org/10.1175/2009JCLI2909.1.

von Bloh, M., Lobell, D., Asseng, S., 2024. Knowledge informed hybrid machine learning in agricultural yield prediction. Comput. Electron. Agric. 227, 109606. https://doi.org/10.1016/j.compag.2024.109606.

Zuazo, V.H.D., García-Tejero, I.F., Rodríguez, B.C., Tarifa, D.F., Ruiz, B.G., Sacristán, P. C., 2021. Deficit irrigation strategies for subtropical mango farming. A review. Agron. Sustain. Dev. 41, 13. https://doi.org/10.1007/s13593-021-00671-6.