Probabilistic yield forecasting of robusta coffee at the farm scale using agroclimatic and remote sensing derived indices

Louis Kouadio[1,*], Vivekananda M. Byrareddy[1], Alidou Sawadogo[2], Nathaniel K. Newlands[3]

[1] Centre for Applied Climate Sciences, University of Southern Queensland, Toowoomba, Queensland, 4350, Australia

[2] Biosystems Engineering Department, Faculty of Agriculture, University of Uludag, Görükle Kampüsü, 16059 Nilüfer/Bursa, Turkey

[3] Summerland Research and Development Centre, Agriculture and Agri-Food Canada, Summerland, British Columbia, V0H 1Z0, Canada

* Correspondence: Email: louis.kouadio@usq.edu.au; Tel. +61 7 46311169

Abstract

Timely and reliable coffee yield forecasts using agroclimatic information are pivotal to the success of agricultural climate risk management throughout the coffee value chain. The capability of statistical models to forecast coffee yields at different lead times during the growing season at the farm scale was assessed. Using data collected during a 10-year period (2008-2017) from 558 farmers across the four major coffee-producing provinces in Vietnam (Dak Lak, Dak Nong, Gia Lai, and Lam Dong), the models were built through a robust statistical modelling approach involving Bayesian and machine learning methods. Overall, coffee yields were estimated with reasonable accuracies across the four study provinces based on agroclimate variables, satellite-derived actual evapotranspiration, and crop and farm management information. Median values of prediction mean absolute percentage error (MAPE) ranged generally from 8% to 13%, and median root mean square errors (RMSE) between 295 kg ha^{-1} and 429 kg ha^{-1}. For forecasts at four to one month before harvest, errors did not vary markedly when comparing the median MAPE and RMSE values. For farms in Dak Lak, Dak Nong, and Lam Dong, the median forecasting MAPE and RMSE varied between 13% and 16% and between 420 kg ha^{-1} and 456 kg ha^{-1}, respectively. Using readily and freely available data, the modelling approach explored in this study appears flexible for an application to a larger number of coffee farms across the Vietnamese coffee-producing regions. Moreover, the study can serve as basis for developing a coffee yield predicting forecasting system that will offer substantial benefits to the entire coffee industry through better supply chain management in coffee-producing countries worldwide.

Keywords: *Coffea canephora*, crop yield forecasting, remote sensing, climate risk management.

1. Introduction

Coffee bean production is currently dominated by two species worldwide, *Coffea arabica* L. , which represents approximately 60% of the global coffee production, and *C. canephora* Pierre ex A. Froehner ('robusta coffee') which accounts for the remaining 40% ((ICO, 2020)). In several American, Asian and African countries coffee farming represents an important source of incomes for smallholder farmers. As one of the most traded agricultural commodities globally, and given its sensitivity to environmental conditions such as changes in rainfall and temperature patterns during critical phenological stages, timely and reliable coffee production forecasts using agroclimatic information are of paramount interest for all the stakeholders of the coffee supply chain – from farmers to agribusinesses to governments and policy makers. When an impending outlook of likely drier than normal weather conditions is on the horizon, an improved insight into the likely impacts of these conditions on the final bean yield can help farmers plan for alternatives (e.g. planting a drought-tolerant short-season cash crop, improve their management practices) to cope with potential yield losses and ensure satisfactory levels of farm incomes. To agribusinesses that plan or sell ahead of the typical annual harvest, such knowledge could guide actions such as forward buying or selling in case of impeding outlook of likely wetter than normal weather conditions, along with strategic resource mobilization in the most insecure coffee-producing areas ((Stone and Meinke, 2005; Kouadio and Rahn, 2020)). Coffee yield forecasts may also help governments and policy makers to protect domestic agriculture and trade by offering early warning of potential disaster risks and crop loss impacts due to adverse weather events.

A wide range of biophysical, process-based models and statistical models have been used to generate crop yield estimates and forecasts at different temporal and spatial coverage and scale (e.g. (Supit, 1997; Genovese, 2001; Challinor et al., 2005; Potgieter et al., 2005; Wu et al., 2014; Chipanshi et al., 2015; Kouadio et al., 2018; Basso and Liu, 2019; Chen et al., 2019; van der Velde and Nisini, 2019; Schauberger et al., 2020; Kouadio et al., 2021)). Models differ by their level of complexity, ability to integrate diverse sources of available data, data requirements, and their deterministic versus probabilistic description of underlying biophysical processes (e.g. linearity or nonlinearity assumptions) ((Kouadio and Newlands, 2014; Newlands et al., 2014)). The complexity and parameterisation issues found in coffee crop growth simulation models constitute some of the major hurdles to their application in data-scarce environments ((van Oijen et al., 2010; Rahn et al., 2018; Ovalle-Rivera et al., 2020; Kouadio et al., 2021)); hence the use of robust statistical analytics as an alternative to predict and/or forecast crop yields. Following Luo et al. (2011) and Newlands et al. (2014), in this study we refer to crop yield prediction or estimates as the quantification of yield potential based on logical consequences of model structure, and yield forecasting approach as involving a probabilistic statement of future yield after data are assimilated into a model.

The accuracy gains of employing robust and reliable statistical analytics for coffee yield prediction or forecasting at the farm scale has yet to be fully investigated, so as to better gauge its potential usefulness and effectiveness to support stakeholder decision-making. The integration of data from various sources including satellite remote sensing sensors, climate stations, soil classification databases, and crop surveys, into a statistical modelling framework was successfully applied for spring wheat, barley and canola yield outlooks within the cropping season across the agricultural regions in Canada using a geospatial, statistical modelling tool, the Canadian Crop Yield Forecaster (CCYF) ((Newlands et al., 2014; Chipanshi et al., 2015)). Originally developed for in-season crop yield forecasts at the Canadian census agricultural region scale, the CCYF was further employed for crop yield forecasting at the ecodistrict

((Kouadio et al., 2014)) and township or rural municipality ((White et al., 2020)) scales; thus demonstrating its applicability at various spatial scales. However, the use of the CCYF modelling framework at finer spatial resolutions (e.g. farm scale) and/or for perennial crops has yet to be explored.

The main objective of this study was to investigate a robust statistical modelling approach to forecast coffee yield at the farm scale using data collected during a 10-year period (2008-2017) across the four major coffee-producing provinces in Vietnam (Dak Lak, Dak Nong, Gia Lai and Lam Dong). Specifically, the potential of agroclimate variables (cumulative growing degree day, rainfall, standardized precipitation-evaporation index), remote sensing satellite-derived actual evapotranspiration, and crop and management practices data (age of plant, irrigation amounts and fertilizer rates) for predicting *C. canephora* yield was first assessed using a methodology similar to the CCYF. Next, the capability of the statistical models to forecast coffee yield with different lead times (one to four months before harvest) was evaluated. Vietnam is the top producing country of *C. canephora* in the world; the total Vietnamese coffee production was on average 1.2 million metric tons over the 2010-2017 period ((FAO, 2018; ICO, 2020)). In addition to providing complementary information about the application of the CCYF framework at higher spatial scale and to perennial crops, the outcomes of the study can serve as basis for developing a crop yield forecasting tool for coffee that would offer substantial benefits to the entire coffee industry in Vietnam, or other coffee-producing countries, through better supply chain management.

2. Materials and Methods

2.1 Study region

C. canephora is dominant across the Central Highlands region in Vietnam, with the total annual production from the provinces Dak Lak, Dak Nong, Gia Lai, and Lam Dong ((GSOV, 2017; Byrareddy et al., 2019)). These provinces account altogether for more than 90% of the total national production ((GSOV, 2017)). The study region encompassed these four major coffee-producing provinces (Fig. 1). The climate in the Central Highlands is predominately humid tropical, with annual rainfall ranging from 1800 to 3000 mm on average, maximum daily temperatures normally above 24 °C and annual total solar radiation varying between 428 to 698 MJ m^{-2} ((Byrareddy et al., 2019)). There are two dominant soil types across the Central Highlands: reddish brown ferrasols and reddish yellow acrisols, with coffee trees being cultivated mostly on the former ((Tien, 2015; Tiemann et al., 2018)).

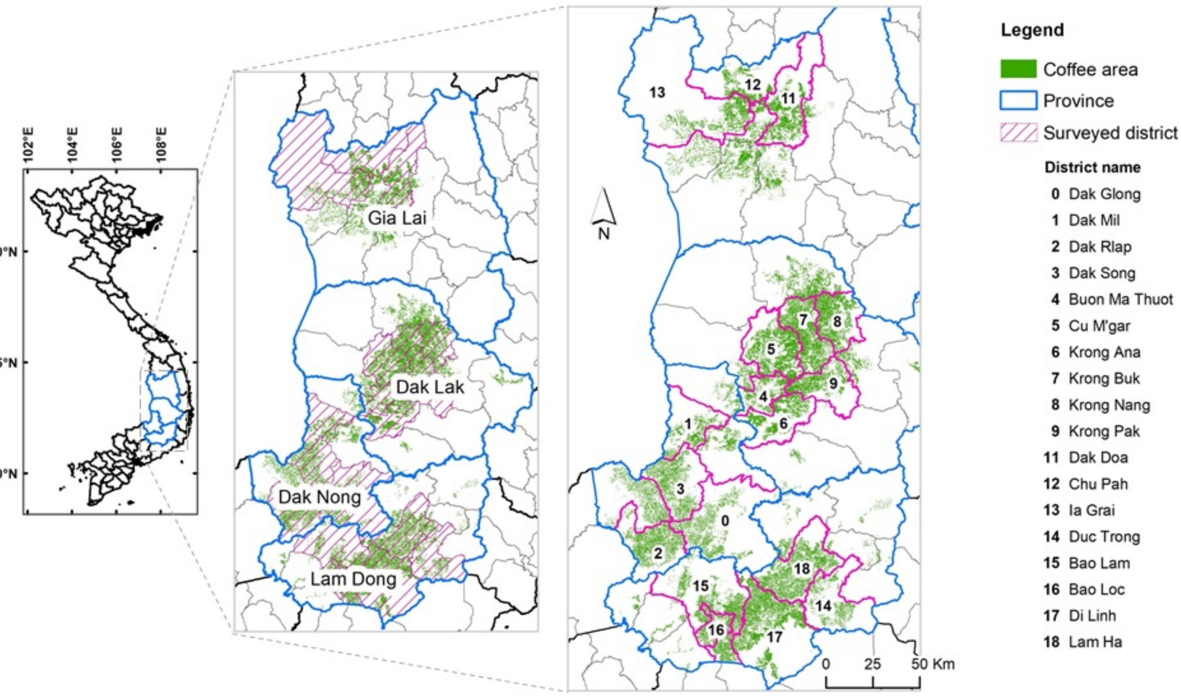

Fig. 1. Map of the four major coffee-producing provinces (Dak Lak, Dak Nong, Gia Lai and Lam Dong) in Vietnam. Districts across the coffee-producing provinces from which the 558

farms were surveyed during the period 2008-2017 are presented. The coffee crop mask was sourced from ECOM Vietnam (https://www.ecomtrading.com).

2.2 Input data

2.2.1 Farm and yield data

Farm level crop and management practices data including the age of trees, monthly irrigation amounts, annual fertilizer rates, and coffee bean yields were collected during the 2008-2017 period from 558 coffee farmers located across the four study provinces (Fig. 1). The total number of farmers surveyed each year was 180, 93, 120, and 165 in Dak Lak, Dak Nong, Gia Lai, and Lam Dong, respectively. The same farmers were surveyed each year. Detailed information about the methodology of data collection can be found in Byrareddy et al. (2019) and Byrareddy et al. (2020). In Vietnam, *C. canephora* was generally grown as an unshaded and clean-weeded monocrop at the surveyed farms, at plant density ranging from 1000 to 1100 plants ha^{-1}. Farm areas ranged from 0.1 to 11.2 ha, with 52% varying between 1 and 3 ha ((Byrareddy et al., 2019)). To be considered in the analysis, the total rates of all chemical fertilizers applied (i.e. blended NPK, urea, super phosphate, and potassium chloride) were expressed in total rate of each of the major nutrients nitrogen (N), phosphate (P_2O_5) and potassium (K_2O). Although the amounts of organic fertilizers (compost and lime) were recorded each year during the survey period, they were not included as potential predictors since no chemical analyses of these organic fertilizers were carried out to provide their detailed composition in key nutrients N, P_2O_5 and K_2O (such chemical analyses were out of the scope of the surveys).

2.2.2 Agroclimate data

Gridded daily data (0.5° × 0.5° spatial resolution) of rainfall, minimum and maximum temperatures for the 2008-2017 period, sourced from the NASA POWER (https://power.larc.nasa.gov/), were used. Data from the grid in which each farm falls were considered for that farm. NASA POWER data are satellite and model-based products which have been used in several studies dealing with crop growth and yield modelling at various spatial scales including farm scale (e.g. (Bai et al., 2010; van Bussel et al., 2011; van Wart et al., 2013)), providing reliable climate resource data over regions where surface measurements are sparse or non-existent (https://power.larc.nasa.gov/docs/methodology/). Daily GDD above a base temperature of 12 °C, calculated using daily minimum and maximum temperatures, and daily rainfall for the months January to September were temporally summed by month and included as potential predictors in the model.

Standardized precipitation-evaporation index (SPEI) data for the 2008-2017 period, with a 0.5° × 0.5° spatial resolution and a monthly time resolution, were also used as potential predictors to include the effect of drought conditions on coffee yield. SPEI data were sourced from the SPEIbase v2.6 database (https://digital.csic.es/handle/10261/202305; (Vicente-Serrano et al., 2010)). Similar to NASA POWER data, SPEI data from the grid in which each farm falls was considered for that farm. Because of the spatial resolutions of gridded climate and SPEI data, all farms falling in the same grid will have the same values of climate and SPEI variables.

2.2.3 Satellite derived actual evapotranspiration data

Satellites images from the sensors Landsat 5 Thematic Mapper (TM), Landsat 7 Enhanced Thematic Mapper plus (ETM+), and Landsat 8 Operational Land Imager (OLI) and Thermal

Infrared Sensor (TIRS) for the 2008-2017 period were used to derive the actual evapotranspiration (ET_a) data at each of the surveyed farms. Landsat imagery scenes for path 124 and rows 050-052 were used (Table S1). For farms in overlapping images, the image with low cloud cover was kept.

ET_a estimations were carried out using the python module for the Surface Energy Balance for Land model (PySEBAL; (Bastiaanssen et al., 1998; Hessels et al., 2017)). PySEBAL calculates the surface energy balance for the day that the satellite image was acquired, independently from the land use and based on information derived from the satellite images (i.e. NDVI, soil-adjusted vegetation index, soil emissivity, surface albedo, leaf area index, and surface temperature), weather (hourly and daily air temperature, wind speed, solar radiation and relative humidity) and digital elevation model data ((Bastiaanssen et al., 1998)). ET_a values at the date of satellite image acquisition are derived as residuals of the surface energy balance. In our study, following the approach in Trezza et al. (2018), daily ET_a values between successive satellite image acquisitions were estimated based on the construction of a crop coefficient curve for every pixel over the study area. Further information about the SEBAL model and procedures to interpolate daily results between image acquisition dates can be found in Bastiaanssen and Ali (2003), Zwart and Bastiaanssen (2007), and Trezza et al. (2018). In our study, daily ET_a data from January to September were temporally summed by month and considered as potential predictors in the model.

Prior to the analyses, the satellite images were pre-processed within PySEBAL. Pre-processing steps included atmospheric corrections, data resampling (i.e. data rescaling from 100 m to 30 m resolution), cloud mask creation, and handling of digital elevation model (DEM) data ((Hessels et al., 2017)). Landsat images were sourced from the U.S Geological Survey

(USGS)'s Earth Explorer (https://earthexplorer.usgs.gov/). Hourly and daily weather data were retrieved from http://www.soda-pro.com and https://power.larc.nasa.gov/. DEM data at a 30-m resolution were sourced from the U.S. National Aeronautics and Space Administration (NASA)'s Shuttle Radar Topography Mission (SRTM). Given the geographic position of the study area (Fig. 1) and the scarcity of clear-sky satellite images in such humid tropical conditions, a tolerance threshold of up to 37% of cloud cover was adopted during the selection of available Landsat images. The average number of days between available and usable satellite images varied between 8 and 30 days, which resulted in a minimum of one image per month for a given surveyed coffee farm (Table S1). The use of such a satellite image threshold was successfully tested to derive such evapotranspiration values ((Allen et al., 2007; Allen et al., 2014)). All satellite images processing and data handling were carried out using the QGIS software (v2.18.27; https://qgis.org/) and the software ArcGIS (v10.4; (ESRI, 2010)). QGIS was used in conjunction with PySEBAL to calculate ET_a data. ArcGIS was used for mapping.

2.3 The farm level coffee yield forecasting model

We used an approach similar to the CCYF ((Newlands et al., 2014; Chipanshi et al., 2015)), which is based on the Bayesian and artificial intelligence/machine learning (i.e. adapted random forest algorithm) methods, to develop the coffee yield forecasting models at the farm scale. Coffee yield at each farm was modelled as a multivariate regression equation as follows ((Newlands et al., 2014)):

$$\hat{y}_{i,j} = y_{i,0} + y_{i,1} + \sum_{l=2}^{n} \alpha_{i,j}^{(l)} x_{i,j}^{(l)} + \varepsilon_{s,i} \quad (1)$$

where $\hat{y}_{i,j}$ denotes the estimated or expected of coffee yield for a given farm i for year j; $y_{i,0}$ and $y_{i,1}$ are the regression intercept and the technology trend over time, respectively; $x_{i,j}^{(l)}$

denotes the l predictor variables for i at year j; l could be any of the potential predictors (monthly total rainfall, cumulative GDD, SPEI, satellite-derived ET_a, irrigation amounts, age of trees, and annual fertilizer rates), with n being the total number of predictors. $\alpha_{i,j}^{(l)}$ are the regression coefficients. The model uncertainty $\varepsilon_{s,i}$ is independent and normally distributed with mean zero and variance σ_i^2. $y_{i,0}$ and $y_{i,1}$ were used to detrend the yield data. The technology trend was assumed to be linear (i.e. historical increases in yield from genetics and improved crop management practices).

All potential predictors (age of tree, annual total fertilizer rate, monthly irrigation amounts from January to April monthly total rainfall, monthly cumulative GDD, monthly SPEI, and monthly satellite-derived ET_a) were assumed to have a truncated normal distribution ((Newlands et al., 2014; Chipanshi et al., 2015)) and were standardized before selection of the best predictors. The standardization was carried out as follows:

$\begin{cases} if\ MAD(x) = 0, x' = x - mean(x)/sd(x) \\ else, x' = x - median(x)/MAD(x) \end{cases}$, where MAD is the median absolute deviation.

During this model building step, highly correlated variables (correlation coefficient ≥ 0.75) were removed to avoid multicollinearity. Moreover, given the short time period of data (10 years), to avoid model overfitting, a maximum of four predictors was selected for each farm. An automatic ranking and selection of the best predictors was performed using a robust least-angle regression scheme (RLARS; (Efron et al., 2004; Khan et al., 2007)), coupled to robust cross-validation (RCV; (Khan, 2010)). The RCV (i.e. leave-one-year-out cross-validation) was used to finalize the training and calibration of each coffee yield model and stabilize such yield model by removing any false predictors selected from contaminated data ((Newlands et al., 2014)).

To proceed with the sequential-based forecasting step, the Bayesian-based spatial correlation analysis, as described by Bornn and Zidek (2012), was first applied to analyse the residual spatial covariance between the modelling units (i.e. farm) and select statistically neighbouring coffee farms. Such a method was used to further increase each yield model's prediction power and stabilize its performance and obtain a more meaningful prior distribution of model predictors ((Newlands et al., 2014; Chipanshi et al., 2015)). Then, the prior distributions of selected predictors were generated using historical data for the statistically selected neighbouring farms and the forecasting farm. The posterior distributions of selected predictors were then obtained using the prior distributions and near-real time data at the time of forecast within the Markov-Chain Monte Carlo (MCMC) scheme ((Dowd, 2006)). Finally, the sequential-based forecasting of coffee yield was carried out using the random forests learning algorithm ((Breiman, 2001; Liaw and Wiener, 2002)), coupled to a bootstrapping process. The estimated variables, along with those available at the time of forecasting, were then used as inputs into the model to forecast the yield probability distribution for each coffee farm. The main outputs of forecasted yield probability distributions included the 10^{th} percentile (worst 10%), the 50^{th} percentile (median) and the 90^{th} percentile (best 10%). Similar probability measures have been used in e.g. Potgieter et al. (2003); Newlands et al. (2014); and Chipanshi et al. (2015). A schematic representation of the data processing and modelling approach is shown in Fig. 2. Detailed information about all the statistical procedures involved within the CCYF can be found in Newlands et al. (2014) and Chipanshi et al. (2015).

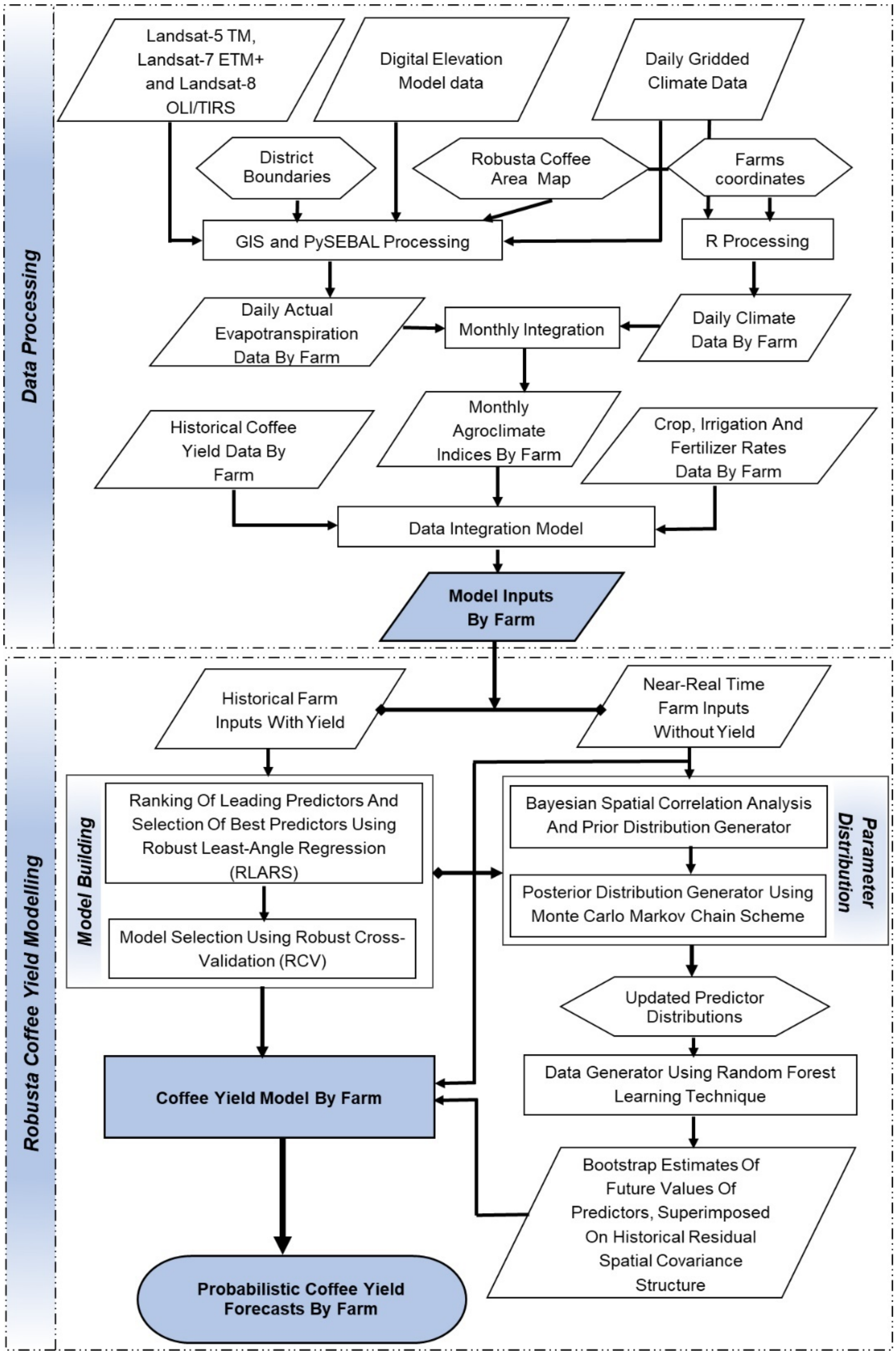

Fig. 2. Flowchart describing the data processing and modelling approach for predicting and forecasting coffee yields at the farm scale (adapted from (Chipanshi et al., 2015)). Abbreviations: TM: thematic mapper; ETM+: enhanced thematic mapper plus; OLI: Operational Land Imager; TIRS: thermal infrared sensor; GIS: geographic information system; PySEBAL: python module for the surface energy balance model.

2.4 Model performance assessment

For each of the 558 coffee farms, monthly data for rainfall, cumulative GDD, SPEI, satellite-derived ET_a, and irrigation, along with the annual total fertilizer rate, and the age of trees for the period 2008 to 2017 were used to develop the yield model. *Year* was included as additional predictor in all modelling cases to incorporate the technology effect over time on crop yield. Given the short farm data period (10 years) a leave-one-year-out cross-validation (LOOCV) was used to assess the robustness of each of the models to estimate coffee yield. Through the LOOCV, for a given farm, a single year observation from the original 10-year record was used as a sample of validation data, and the remaining observations as the training data. This was repeated iteratively until every year in the sample was used once as validation data. With regard to coffee yield forecasting, four different lead times were considered in the study. With *C. canephora* harvest occurring generally during October to December in Vietnam ((Byrareddy et al., 2019; Byrareddy et al., 2020)), yield forecasts were generated on the first dates of June, July, August, and September. For each of the lead times, coffee yields were forecast using the observed data from January till the last day of the month preceding that lead time (for predictors belonging to that period) and the bootstrap estimates generated within the modelling process (for predictors belonging to the remainder of the growing season). If the annual total fertilizer rate or the age of trees or the monthly irrigation amounts were selected as model predictors, then no estimates were generated for yield forecast for these variables, since irrigation data spanned the months of January to April and the age of trees or fertilizer rate had unique value in each year.

Predicted and forecast coffee yields were compared against observed yields to assess the performance of the yield models. A combination of performance metrics are often required to assess the overall model performance since one metric only emphasizes a certain aspect of the error characteristics ((Chai and Draxler, 2014)). Four statistical indicators were used in the study. They were the Pearson correlation coefficient (r), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and the Nash-Sutcliffe Efficiency index (NSE; (Nash and Sutcliffe, 1970)). Their respective equations are as follows:

$$r = \frac{\sum_{i=1}^{N}(Y_i^m - \bar{Y}^m)(Y_i^p - \bar{Y}^p)}{\sqrt{\sum_{i=1}^{N}(Y_i{}^m - \bar{Y}^m)^2}\sqrt{\sum_{i=1}^{N}(Y_i{}^p - \bar{Y}^p)^2}} \qquad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i^m - Y_i^p)^2} \qquad (3)$$

$$MAPE = 100 \times \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i{}^m - Y_i{}^p}{Y_i{}^m}\right| \qquad (4)$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N}(Y_i^m - Y_i^p)^2}{\sum_{i=1}^{N}(Y_i^m - \bar{Y}^m)^2}\right] \qquad (5)$$

where N is the number of sample years; Y_i^m is the *i*th observed value; $\bar{Y}^m$ is the mean observed value; and $Y_i{}^p$ is the *i*th predicted value.

The RMSE gives the weighted variations in errors (residual) between the predicted and observed yields. The MAPE is an accuracy measure of the forecast quality, suitable for comparing model performance among different spatial units given the likely differences in their average historical yields ((Chipanshi et al., 2015)). The RMSE and the MAPE have both the same interpretation: their value decreases as the prediction (or forecast) improves. The NSE can be a useful index for evaluating model efficiency, in addition to the RMSE and MAPE ((Murphy, 1993; Krause et al., 2005)). The closer the NSE is to 1, the more skilful the model forecast is. Negative NSE values are indicative of less skilful model forecast; the forecast value in such case can be replaced by the historical average yield ((Murphy, 1993)).

3. Results

3.1 Observed coffee yield variations

An overview of *C. canephora* yield distribution per province and the yield coefficient of variation (CV) are presented (Fig. 3). In all provinces the coffee yields were on average lower during 2008-2010 than during the 2011-2017 period (Fig. 3A). Although in provinces like Dak Lak, Dak Nong and Gia Lai, the overall yield distribution seemed to exhibit a biannual variation, analyses of the interannual yield variability showed that there was no substantial difference between yields from one year to another for a given farmer, which was illustrated through the strong ($r^2 \geq 0.80$) and statistically significant linear annual trends ($P < 0.05$; Fig. 3C). At the level of each surveyed farm, yield CVs were most often below 20%, regardless of the province (Fig. 3B).

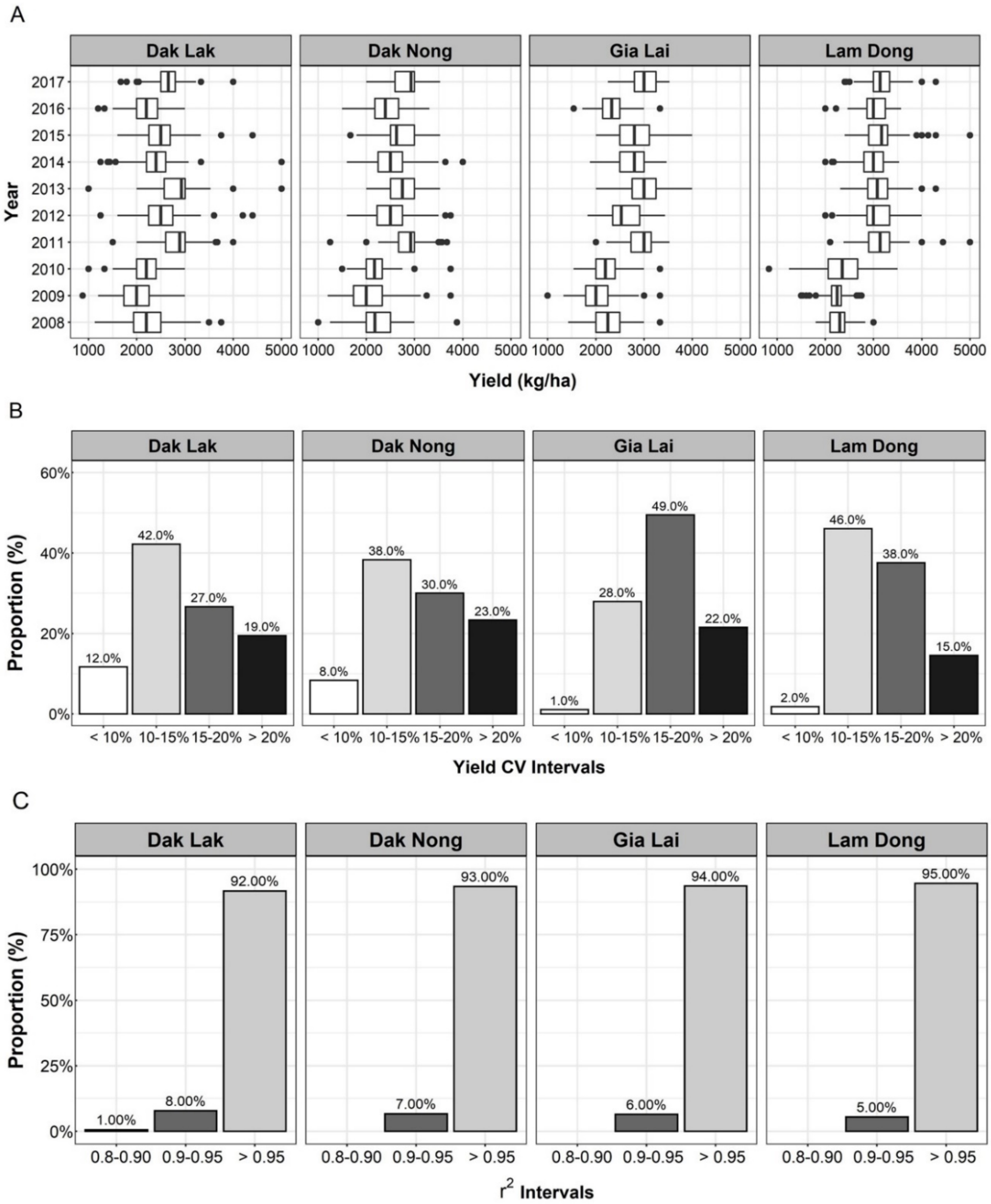

Fig. 3. Yield variations of *C. canephora* during the 2008-2017 period across the study provinces in Vietnam. (**A**) boxplots of annual coffee yields; (**B**) distribution of yield coefficients of variation (CV); and (**C**) distribution of the coefficient of determination (r^2) obtained after the annual yield trend analysis (*Yield = f(year)*). In (**A**) all data per province were pooled. In (**B**) and (**C**) the CV and trend analysis results were determined for each of the surveyed farms in a given province. All the model estimates in (C) were statistically significant

at $\alpha = 0.0001$. In a boxplot, the top and bottom of the box represent the 75^{th} and 25^{th} percentiles; the solid line indicates the median. The whiskers on the top and bottom represent the largest and smallest values within 1.5 times interquartile range above the 3^{rd} and 1^{st} quantiles, respectively. Black circles are the outliers.

3.2 Predictors explaining coffee yield interannual variability

A total of 42 potential predictors were evaluated to predict coffee yield at the farm scale across the four major coffee-producing provinces in Vietnam. Irrespective of the province, predictors such as the monthly irrigation of March and April, and the annual fertilizer rate were not selected. This can be explained by the lack of predictive ability of these variables or the statistical selection process of leading explanatory variables in which predictors with zero or near zero variance were systematically discarded. For the remainders of the potential predictors, the majority was selected to explain the variability in coffee yield at each of the surveyed farms, though in varying proportions (Fig. 4). For all the surveyed farms, predictors related to months of January to May, namely those for rainfall, GDD and ET_a, were among those with relatively highest proportions (up to 15%; Fig. 4). However, specificities in terms of variables selected were found in some provinces. This is the case for coffee farms in Dak Nong and Lam Dong. For farms in in the former province, only the age of trees (Age_plant), GDD for June (GDD_6), SPEI for January (SPEI_1), and monthly irrigation for February (IR_2) were not selected in any of the models. In Lam Dong the predictive ability of SPEI values was not dominant for most months; only SPEI for March (SPEI_3) and May (SPEI_5) were among the selected predictors (Fig. 4). For the surveyed farms in that province, variations in SPEI seemed not to impact coffee yields.

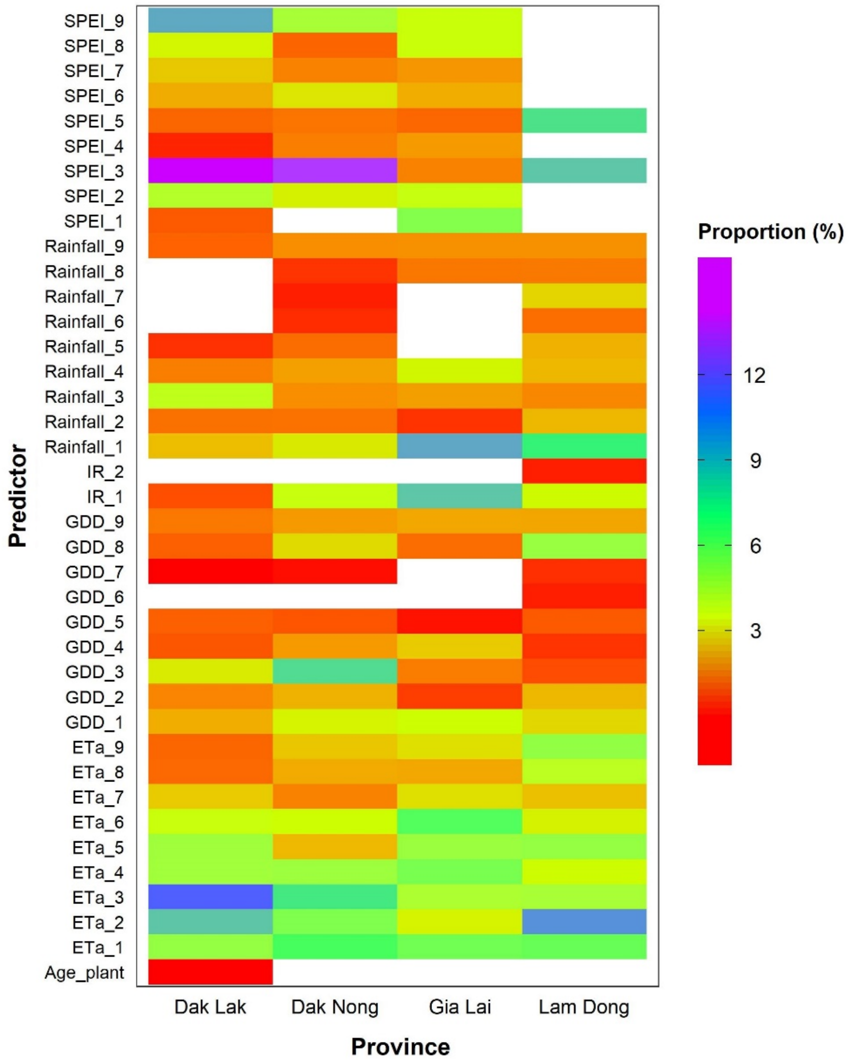

Fig. 4. Proportions of predictors selected during the model building step. Predictors selected for each of the models in each province were pooled to calculate the proportions. ET_a, *GDD, SPEI*, and *IR* refer to actual evapotranspiration, cumulative growing degree days, standardized precipitation-evaporation index, and irrigation amount, respectively. The numbers 1 to 9 correspond to months January to September, respectively. For a given predictor, blank areas indicate the non-selection of the variable in all the models built because of its zero predictive ability.

3.3 Performance of models in predicting coffee yield at the farm scale

Yield estimation errors (RMSE and MAPE), along with Pearson correlation coefficients, derived from the LOOCV, are presented in Fig. 5. Overall, the median RMSE values did not vary substantially over the 10-year period across each of the study provinces, indicating similar model performance between years. Models for farms in Lam Dong were among those with good performance indicators when comparing prediction errors and Pearson correlation coefficients altogether (Fig. 5; Table 1). Median RMSE values for farms in Lam Dong varied between 295 and 337 kg ha^{-1}; median MAPEs were most often below 15%; and r values were most often above 0.5 (Fig. 5). On the other hand, models for farms in Gia Lai were those with relatively higher prediction errors, compared to those in other provinces. For farms in that province half of the models resulted in 12% to 13% MAPE and RMSE ranging from 382 to 429 kg ha^{-1} over the 10-year period (Table 1; Fig. 5). For farms in Dak Lak and Dak Nong the models generally achieved similar estimation errors over 2008 to 2017, with median MAPE values varying between 9% and 12%, and median RMSEs ranging from 306 to 363 kg ha^{-1} (Table 1). Compared to the yield CV observed at most coffee farms (Fig. 3B), median MAPEs were generally in the same range or lower. For example, for farms in Lam Dong, yield CV was below 15% for 48% of the surveyed farms (Fig. 3B), whereas MAPEs were ≤ 15% for 75% of the farms (Table 1). Such results indicate acceptable performance of models for estimating coffee yield at the farm scale across the study provinces in Vietnam. Nevertheless, the presence of MAPE values well above 20%, and as high as 63% (e.g. in 2017 in Dak Lak) (Fig. 5) indicates that there were models which did not perform well. In such cases the interannual variability of the selected predictors did not explain that of coffee yields. Given the short data period, the maximum number of predictors in each model was limited to three. This threshold can be revised as data become available over years while avoiding any overfitting, which can result in better yield estimations.

Table 1. Median values of the Pearson correlation coefficients (r), root mean square errors (RMSE), and mean absolute errors (MAPE) obtained after the leave-one-year-out cross-validation process of coffee yield models. Numbers in brackets are the 75th percentiles. Boxplots showing the full distributions of these statistics are presented in Fig. 5.

	Year	r	RMSE (kg ha-1)	MAPE (%)		Year	r	RMSE (kg ha-1)	MAPE (%)
Dak Lak	2008	0.443 (0.620)	347 (447)	11 (16)	Dak Nong	2008	0.592 (0.750)	323 (464)	11 (16)
	2009	0.430 (0.597)	356 (448)	12 (15)		2009	0.586 (0.709)	349 (501)	11 (17)
	2010	0.507 (0.680)	329 (453)	11 (15)		2010	0.558 (0.709)	345 (500)	12 (15)
	2011	0.493 (0.668)	307 (408)	9 (12)		2011	0.637 (0.758)	312 (461)	9 (14)
	2012	0.504 (0.679)	312 (438)	11 (14)		2012	0.631 (0.730)	326 (436)	10 (14)
	2013	0.467 (0.638)	334 (424)	10 (14)		2013	0.614 (0.740)	313 (454)	10 (14)
	2014	0.475 (0.642)	332 (418)	11 (13)		2014	0.613 (0.737)	326 (428)	10 (14)
	2015	0.533 (0.663)	313 (409)	10 (13)		2015	0.611 (0.753)	329 (462)	10 (15)
	2016	0.505 (0.653)	336 (447)	10 (14)		2016	0.620 (0.747)	314 (481)	10 (15)
	2017	0.486 (0.694)	355 (453)	11 (15)		2017	0.592 (0.728)	363 (486)	12 (16)

	Year	r	RMSE (kg ha-1)	MAPE (%)		Year	r	RMSE (kg ha-1)	MAPE (%)
Gia Lai	2008	0.509 (0.691)	391 (534)	13 (17)	Lam Dong	2008	0.660 (0.753)	313 (445)	9 (13)
	2009	0.465 (0.598)	399 (498)	12 (17)		2009	0.655 (0.742)	337 (430)	10 (12)
	2010	0.497 (0.657)	393 (472)	13 (15)		2010	0.656 (0.755)	335 (428)	10 (13)
	2011	0.514 (0.682)	394 (469)	12 (15)		2011	0.723 (0.799)	303 (402)	8 (10)
	2012	0.525 (0.658)	382 (443)	12 (14)		2012	0.685 (0.780)	329 (457)	9 (11)
	2013	0.527 (0.674)	388 (494)	12 (15)		2013	0.697 (0.786)	305 (421)	9 (12)
	2014	0.535 (0.685)	390 (493)	12 (16)		2014	0.714 (0.804)	295 (393)	9 (11)
	2015	0.558 (0.726)	395 (471)	13 (16)		2015	0.731 (0.799)	306 (394)	9 (12)
	2016	0.488 (0.639)	422 (520)	13 (16)		2016	0.710 (0.778)	315 (434)	9 (12)
	2017	0.537 (0.712)	429 (545)	13 (16)		2017	0.699 (0.798)	336 (469)	10 (12)

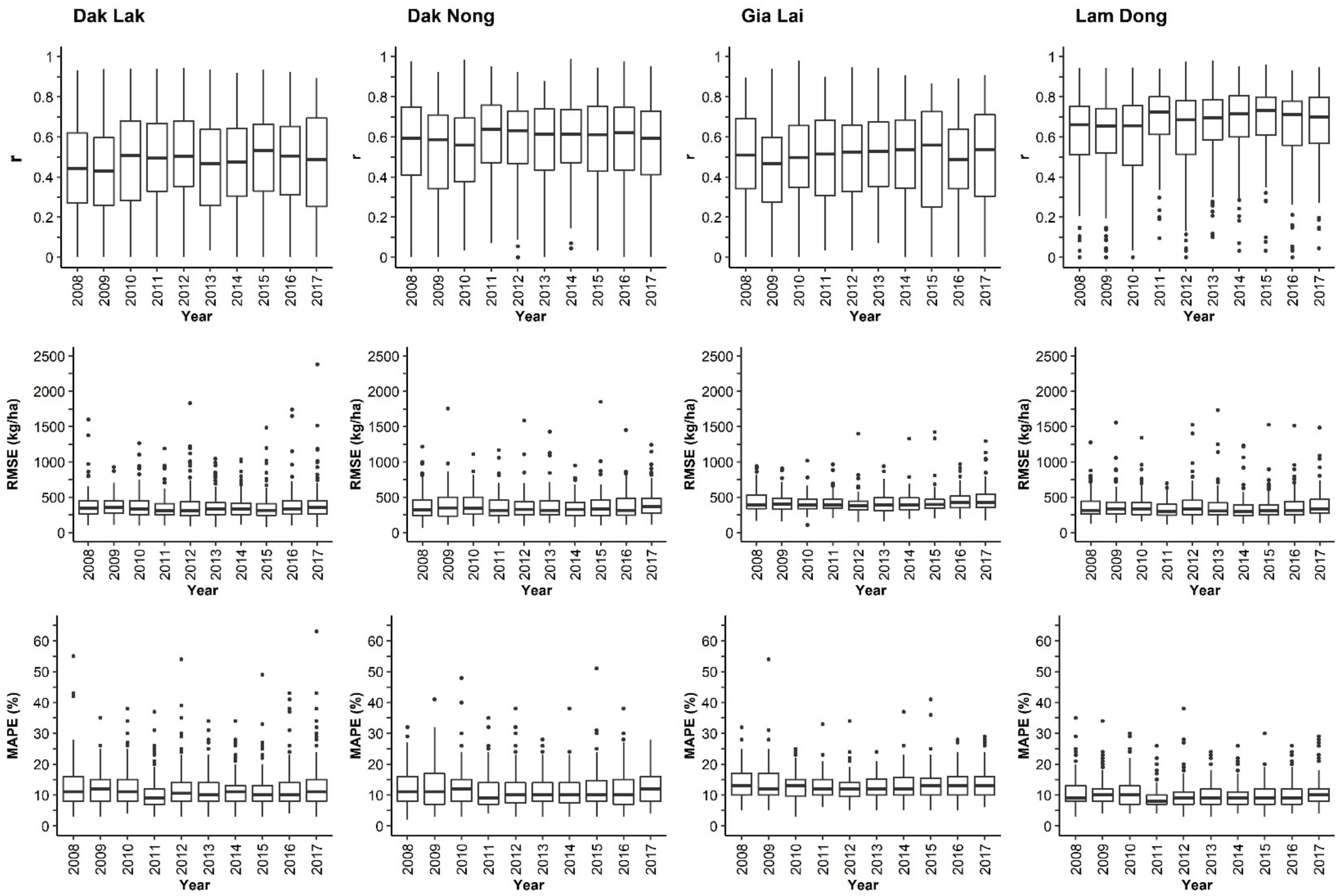

Fig. 5. Boxplots of leave-one-year-out cross-validation (LOOCV) Pearson coefficient of correlation (r), mean absolute percentage error (MAPE), and root mean square error (RMSE) of the farm level coffee yield models. In a boxplot, the top and bottom of the box represent the 75th and 25th

percentiles; the solid line indicates the median. The whiskers on the top and bottom represent the largest and smallest values within 1.5 times interquartile range above the 3rd and 1st quantiles, respectively. Black circles are the outliers.

3.4 Model forecasting skill according to the lead time

The skills of each of the statistical models for forecasting coffee yield at four to one months before harvest were evaluated. The full distribution (i.e. boxplots) of the statistical indicators r, RMSE, and MAPE is presented in Fig. S1. Here, only the median values of the three indicators are presented (Fig. 6). For farms in a given province, the forecasting skills did not vary substantially between the four lead times when comparing the median MAPE and RMSE values, even though there were slight increases in errors as the lead time decreased (Fig. 6). Overall, models for farms in Lam Dong were those with better forecast skills during the study period (i.e. lowest median MAPEs and RMSEs). For these models, median MAPEs were 13% for forecasts at four to two-month lead, and 14% at one-month lead; the median RMSEs increased slightly from around 420 to 456 kg ha^{-1} from four-month to one-month lead (Fig. 6). Models for farms in Dak Lak and Dak Nong yielded both in similar patterns in terms of median RMSEs and MAPEs, with respective ranges of median RMSEs and MAPEs being 420 to 447 kg ha^{-1}, and 14% to 16% (Fig. 6). The relatively highest forecast errors were found for farm-scale models in Gia Lai: median MAPEs were around 18%-19% and median RMSEs varied between 530 and 560 kg ha^{-1} (Fig. 6).

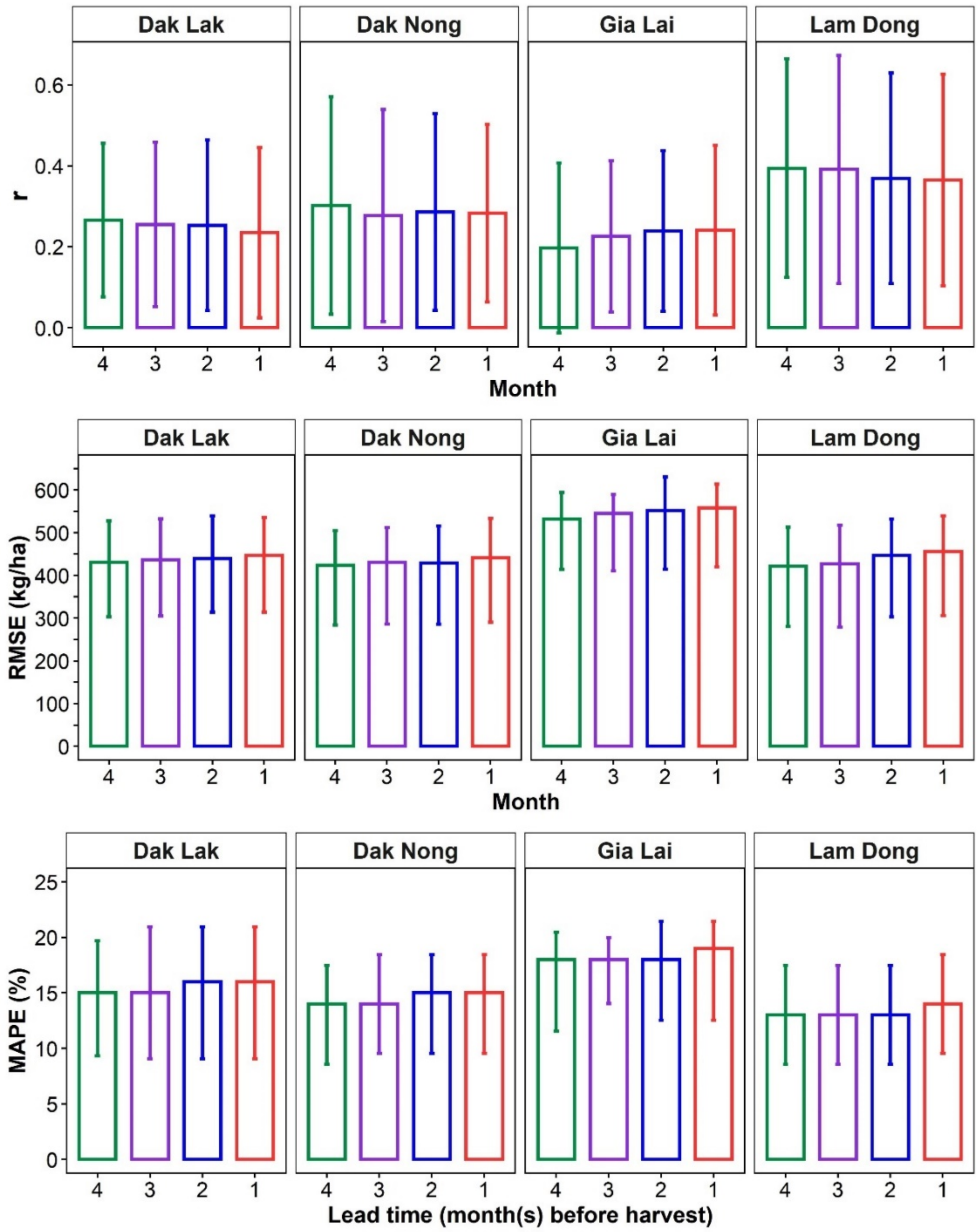

Fig. 6. Performance of the statistical models used to forecast coffee yield at the farm scale at four different lead times (one to four months before harvest) in the study provinces in Vietnam. The median values for Pearson correlation coefficients (r), root mean square error (RMSE), and mean absolute percentage error (MAPE) are presented. Error bars indicate the minimum and maximum median absolute errors.

We further analysed the distribution of NSE values for all the models according to their forecasting skills at the defined lead times. The probability of having models with negative NSE was high ($F(x) \geq 0.75$) and fairly consistent between lead times for farms in in Dak Lak, Gia Lai, and Dak Nong (Fig. 7), indicating that for the majority of models the historical average coffee yield might be suitable compared to the forecast, irrespective of the lead month. However, for farms in Lam Dong, the distribution of NSE values showed that positive NSE were obtained for coffee yield forecasts in half of the surveyed coffee farms, namely at four and three-month lead times ($F(x) = 0.50$; Fig. 7). The proportions of models with negative NSE suggest that for most of the surveyed farms, at least for those in Dak Lak, Dak Nong and Gia Lai, additional data input and modelling considerations (see section 4) is needed to improve the forecasting skills.

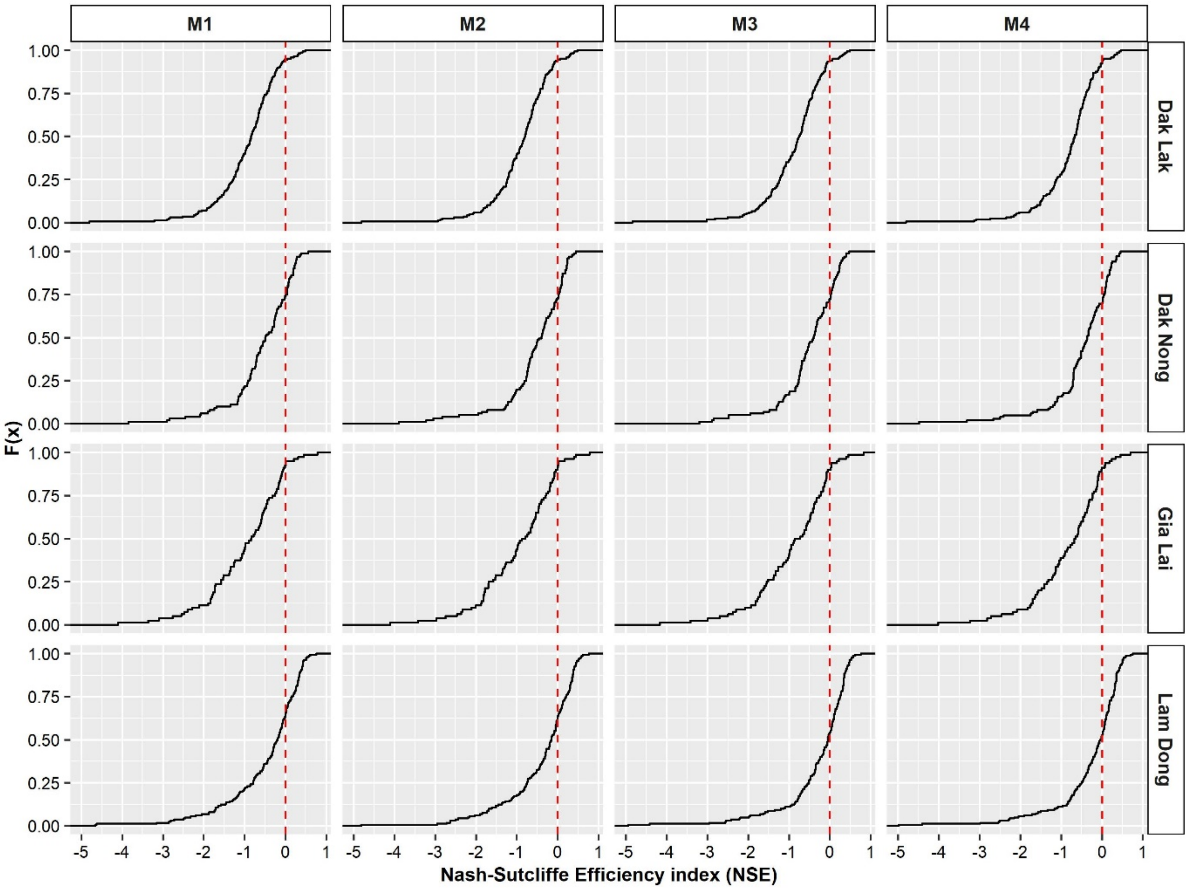

Fig. 7. Empirical cumulative density function of the Nash-Sutcliffe Efficiency index (NSE) of the farm-scale coffee yield models according to the forecast lead time. The closer the NSE is to 1, the more skilful the model forecast is. Negative NSE values are indicative of least skilful model forecast. The total number of surveyed farmers was 180, 120, 93, and 165 in Dak Lak, Dak Nong, Gia Lai, and Lam Dong, respectively. Coffee harvest starts typically in October each year and spans three months.

4. Discussion

4.1. Comparison to previous studies

We investigated the capability of statistical models to forecast coffee yield at different lead times during the growing season at the farm scale across the four major coffee-producing provinces in Vietnam (Dak Lak, Dak Nong, Gia Lai, and Lam Dong). The models were also evaluated according to their performance in estimating the final yield. The statistical modelling approach explored in this study was similar to that of the CCYF ((Newlands et al., 2014; Chipanshi et al., 2015)), which was successfully applied for yield estimation and forecasting of cereal and oilseed crops at various spatial scales in Canada ((Kouadio et al., 2014; Chipanshi et al., 2015; White et al., 2020)). To the best of the authors' knowledge, applying such an approach at a finer spatial scale (farm scale) and for a perennial crop has never been reported in the literature. Thus, the investigation of such a statistical modelling approach for coffee yield estimation and forecasting in Vietnamese coffee farm systems constitutes an original contribution of this research.

In our study, for all the surveyed coffee farms across the four provinces, the median estimation MAPEs ranged from 8% to 13%, with corresponding median RMSEs varying between 295 kg ha^{-1} and 429 kg ha^{-1}, indicating good and acceptable models' performance. Such performance is indicatively comparable to that of the CCYF-based models developed for cereal and oilseed crops at the ecodistrict scale ((Kouadio et al., 2014)) and the census agricultural region scale ((Chipanshi et al., 2015; Chipanshi et al., 2019)). In a study dealing with coffee yield prediction in Lam Dong, in which different machine learning (ML)-based modelling approaches including random forest (RF) and extreme learning machine (ELM) were used to estimate *C. canephora* yield at the farm scale, the prediction errors (RMSE) of the optimal RF and ELM models were 561 kg ha^{-1} and 497 kg ha^{-1}, respectively ((Kouadio et al., 2018)). These were relatively higher to those obtained for the majority of farms in this study for the same province, where for 75% of farm-scale models in Lam Dong, RMSEs were as high as 469 kg ha^{-1} (Table 1). Applying the ML-based approach developed in Kouadio et al. (2018) across all the coffee-producing provinces remains challenging given the lack of the data used (i.e. soil fertility data) for many coffee farms and the logistic and human resources to mobilize in order to collect such data every year, which make such a collection laborious and costly. Thus, the approach presented in this study adds valuable information to the challenging topic of reliable and accurate coffee yield prediction in Vietnam since the modelling approach, based on readily and freely available data, emerges as a cost-efficient approach applicable to a large number of coffee farmers to predict coffee yield at the farm scale.

4.2. Impacts of the coffee yield predictors used

When comparing the ranges of errors from one lead time to another, similar ranges of errors (RMSE and MAPE) between forecasts at four to two-month leads in Gia Lai and Lam Dong were observed; in Dak Lak and Dak Nong such similarities were between forecasts at four to three-month and two to one-month leads (Fig. 6). The similarity of forecast errors between the lead times can be explained by the statistical nature of the best-fit selected predictors and their relative importance (i.e. the proportion of observed variance explained by the predictor). Similar errors between forecasts at four to two-month leads indicate that either the leading predictors in terms of relative importance belong to months preceding these forecast months or all the predictors belong to months up to the fourth month before harvest. As for the statistical nature of the modelling approach, some best-fit selected predictors may not be related agronomically to the yield (i.e. variations of the selected predictor have little to no known effect on coffee yield). This can be the case, for example, of September rainfall. In Vietnam, September generally corresponds to the maturation/ripening phase in *C. canephora*, which is a less sensitive phase compared to the blossoming and fruit setting phases ((Kath et al., 2020; Kath et al., 2021; Kouadio et al., 2021)). Water deficits during blossoming or fruit setting, which will negatively affect coffee bean size, would impact more on yield than water deficits during maturation. Using additional variables with agronomic meaning (i.e. derived from process-based, biophysical models) as potential predictors can help to address such limitations, and potentially improve models' forecast skills. Notwithstanding, the capability of models to provide satisfactory forecasts at four-month lead suggests that they can be helpful for providing an outlook of future yield and production at the farm scale across the study coffee-producing provinces in Vietnam.

Year was included as additional predictor in all modelling cases to account for the technology trend due to genetic, nutrient/water use efficiency, harvesting and other technological improvements. Indeed, over the years 2013-2017, several coffee farmers in province such as Lam Dong have been adopting the grafting technique as management practice to increase the number of productive branches, resulting in consistent and relatively high coffee yields under good rainfall and irrigation conditions ((Byrareddy et al., 2020)). Moreover, although the biennial bearing effect on coffee yield in *C. canephora* is well-documented (e.g. (DaMatta, 2004; DaMatta et al., 2007)), such effect was virtually inexistent at the surveyed farms in Vietnam because of irrigation practices ((Byrareddy et al., 2020)) and chemical fertilizers use (generally at higher rates than recommended) ((Byrareddy et al., 2019)).

4.3. Limitations of the study and future directions

Despite the encouraging performance of the statistical models developed, some limitations were found in this study that indicates a need for further research. Fertilizer rates were not selected as yield predictor during the modelling process in all the cases, although nutrients are determinant in the formation of coffee berries, thereby the final coffee bean yield ((DaMatta et al., 2007)). The use of data including wider ranges of fertilizer rates could result in different statistically selected predictors and different model performance. We intended to model coffee yield at the farm scale using agroclimate and remotely sensed variables with agronomic meaning. Quantifying the impact of fertilizer rates on *C. canephora* yield in the study provinces can be investigated further using process-based biophysical models. The modelled relationships can then be integrated within the statistical modelling approach presented in this study to improve the accuracy of coffee yield forecasts.

Another limitation of the study is related to the satellite-derived ET_a data. Using optical remote sensing data in tropical regions such as the coffee-producing provinces in Vietnam was challenging due to cloudiness issue that might have affected the accuracy of the satellite-derived ET_a values. It was beyond the scope of the study to carry out a detailed validation process of such values at each of the surveyed farms. Nonetheless, data from satellites such as those from the Copernicus Sentinel-2 mission which can provide more usable data (i.e. data with reduced cloud proportion), can be explored in future to estimate ET_a, as data become available. Given the flexibility of the modelling approach, other satellite remote sensing data and indices, process-based biophysical outputs, and deep-learning algorithms could be explored to improve the performance of the farm-scale models.

The shortness of data period (10 years) might have impacted on the sequential-based forecasting process within the modelling approach. Although the climate and remote sensing data are available for longer periods, the use of long-period data was limited by the availability of farm data across the study provinces. As data become available over years, the forecasting capability of the models developed can be improved through the better estimation of prior and posterior distributions of potential predictors and refinement of the models' design. Another improvement pathway can involve the inclusion of correlation effects across years, which were not considered. These effects and additional correlation terms could, in the future, be added to the model to improve prediction power.

5. Conclusions

The capability of models built from a robust statistical modelling approach for estimating and forecasting coffee yield at the farm scale across the four major coffee-producing provinces in Vietnam was evaluated. Statistical models were developed using various data including agroclimate (cumulative GDD, SPEI, and rainfall), satellite-derived ET_a, and farm and crop management (age of trees, irrigation, and fertilizer) data. Results showed that the models were efficient in estimating *C. canephora* yield in most farms, with median MAPE and RMSE values ranging from 8% to 13%, and 295 kg ha^{-1} to 429 kg ha^{-1}, respectively. The forecasting skill of the various models did not vary substantially between four to one-month before harvest when comparing the median MAPE and RMSE values, with only slight increases in errors measured as the lead time decreased. However, the proportions of skilful models to forecast coffee yield at the defined lead times, expressed through NSE values, suggest that further research for improvement is needed. Such improvements could be to consider additional potential predictors (e.g. derived from process-based biophysical models) or using a different approach for the temporal aggregation of predictors. Nonetheless, the results presented show sufficiently high yield forecast accuracy, especially given the limited training data period (10 years). The study also ties in valuable, complementary information regarding the application of similar approach as that of the CCYF to perennial crops and at a higher spatial scale. As such, the study provides important analytical findings to improve coffee yield predicting/forecasting for the coffee industry in Vietnam, with potential application also in other coffee-producing countries.

Acknowledgments

We gratefully acknowledge the valuable support of ECOM's field surveys team in collecting and processing field data, as well as all the coffee farmers who participated in the interviews and generously provided information about their production systems. We acknowledge the funding received from the German Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety (BMUB) and the World Meteorological Organisation (WMO) through the DeRISK project. The Prediction of Worldwide Energy Resource (POWER) Project is funded through the NASA Applied Sciences Program within the Earth Science Division of the Science Mission Directorate. The data obtained through the POWER web services were made possible with collaboration from the NASA LaRC Sciences Data Center (ASDC).

CRediT authorship contribution statement

Louis Kouadio: Conceptualization, Methodology, Software, Data Curation, Validation, Formal analysis, Writing - Original Draft. **Vivekananda Byrareddy**: Investigation, Data Curation, Writing - Review & Editing. **Alidou Sawadogo**: Investigation, Data Curation, Writing - Review & Editing. **Nathaniel K. Newlands**: Conceptualization, Methodology, Software, Writing - Review & Editing.

References

Allen, R., Trezza, R., Tasumi, M., Kjaersgaard, J., 2014. Mapping Evapotranspiration at High Resolution using Internalized Calibration Applications. Manual for Landsat Satellite Imagery, Version 3.0.

Allen, R.G., Tasumi, M., Morse, A., Trezza, R., Wright, J.L., Bastiaanssen, W., et al., 2007. Satellite-Based Energy Balance for Mapping Evapotranspiration with Internalized Calibration (METRIC)—Applications. J. Irrig. Drain. Eng. 133, 395-406. https://doi.org/10.1061/(ASCE)0733-9437(2007)133:4(395)

Bai, J., Chen, X., Dobermann, A., Yang, H., Cassman, K.G., Zhang, F., 2010. Evaluation of NASA satellite- and model-derived weather data for simulation of maize yield potential in China. Agron. J. 102, 9-16. https://doi.org/10.2134/agronj2009.0085

Basso, B., Liu, L., 2019. Seasonal crop yield forecast: Methods, applications, and accuracies. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 201-255.

Bastiaanssen, W.G.M., Menenti, M., Feddes, R.A., Holtslag, A.A.M., 1998. A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. J. Hydrol. 212-213, 198-212. https://doi.org/10.1016/S0022-1694(98)00253-4

Bastiaanssen, W.G.M., Ali, S., 2003. A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan. Agric. Ecosyst. Environ. 94, 321-340. https://doi.org/10.1016/S0167-8809(02)00034-8

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

Breiman, L., 2001. Random Forests. Mach. Learning 45, 5-32. https://doi.org/10.1023/A:1010933404324

Byrareddy, V., Kouadio, L., Mushtaq, S., Stone, R., 2019. Sustainable production of Robusta coffee under a changing climate: A 10-year monitoring of fertilizer management in coffee farms in Vietnam and Indonesia. Agronomy 9, 499. http://dx.doi.org/10.3390/agronomy9090499

Byrareddy, V., Kouadio, L., Kath, J., Mushtaq, S., Rafiei, V., Scobie, M., Stone, R., 2020. Win-win: Improved irrigation management saves water and increases yield for robusta coffee farms in Vietnam. Agric. Water Manag. 241, 106350. https://doi.org/10.1016/j.agwat.2020.106350

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Challinor, A.J., Slingo, J.M., Wheeler, T.R., Doblas–Reyes, F.J., 2005. Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. Tellus A 57, 498-512. https://doi.org/10.1111/j.1600-0870.2005.00126.x

Chen, K., O'Leary, R.A., Evans, F.H., 2019. A simple and parsimonious generalised additive model for predicting wheat yield in a decision support tool. Agric. Syst. 173, 140-150. https://doi.org/10.1016/j.agsy.2019.02.009

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. http://dx.doi.org/10.1016/j.agrformet.2015.03.007

Chipanshi, A.C., Qi, D., Zhang, Y., Lin, H., Newlands, N.K., 2019. Prediction of canola and spring wheat yield based on the Canadian Meteorological Centre's monthly forecasting system. Climate Res. 77, 155-165. https://doi.org/10.3354/cr01547

DaMatta, F.M., 2004. Ecophysiological constraints on the production of shaded and unshaded coffee: a review. Field Crops Res. 86, 99-114. http://dx.doi.org/10.1016/j.fcr.2003.09.001

DaMatta, F.M., Ronchi, C.P., Maestri, M., Barros, R.S., 2007. Ecophysiology of coffee growth and production. Braz. J. Plant Physiol. 19, 485-510. http://dx.doi.org/10.1590/S1677-04202007000400014

Dowd, M., 2006. A sequential Monte Carlo approach for marine ecological prediction. Environmetrics 17, 435-455. https://doi.org/10.1002/env.780

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407-499. https://doi.org/10.1214/009053604000000067

ESRI, 2010. Arcgis desktop: Release 10, Technical report, Environmental Systems Research Institute 843 Redlands, CA.

FAO, 2018. FAOSTAT. Crops, National Production. FAO, Rome, Italy.

Genovese, G.P., 2001. Introduction to the MARS Crop Yield Forecasting System (MCYFS). Meeting on 4 and 5 October 2001, Luxembourg. Space Applications Institute, Joint Research Centre of the European Commission, Ispra , Italy , pp 15.

GSOV, 2017. Statistical Yearbook of Vietnam 2017. Statistical documentation and service centre, General Statistics Office of Vietnam (GSOV), Hanoi, Vietnam. https://www.gso.gov.vn/default_en.aspx?tabid=515&idmid=5&ItemID=18941 (accessed on 20 November 2020).

Hessels, T., van Opstal, J., Trambauer, P., Bastiaanssen, W.G.M., Faouzi, M., Mohamed, Y., Er-Raji, A., 2017. pySEBAL version 3.3.7. 2017. https://pypi.org/project/SEBAL/.

ICO, 2020. Historical data on the global coffee trade. The International Coffee Organization. Available at http://www.ico.org/new_historical.asp (accessed on 20 November 2020).

Kath, J., Byrareddy, V.M., Craparo, A., Nguyen-Huy, T., Mushtaq, S., Cao, L., Bossolasco, L., 2020. Not so robust: Robusta coffee production is highly sensitive to temperature. Glob. Change Biol. 26, 3677-3688. http://dx.doi.org/10.1111/gcb.15097

Kath, J., Mittahalli Byrareddy, V., Mushtaq, S., Craparo, A., Porcel, M., 2021. Temperature and rainfall impacts on robusta coffee bean characteristics. Climate Risk Manag. 32, 100281. https://doi.org/10.1016/j.crm.2021.100281

Khan, J., Aelst, S. V., and Zamar, R., 2010. Fast robust estimation of prediction error based on resampling. Comput. Stat. Data Anal. 54, 3121–3130. https://doi.org/10.1016/j.csda.2010.01.031

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

Kouadio, L., Newlands, N., Davidson, A., Zhang, Y., Chipanshi, A., 2014. Assessing the performance of MODIS NDVI and EVI for seasonal crop yield forecasting at the ecodistrict scale. Remote Sens. 6, 10193-10214. https://doi.org/10.3390/rs61010193

Kouadio, L., Newlands, N.K., 2014. Data hungry models in a food hungry world - an interdisciplinary challenge bridged by statistics. In: Lawless, J.F. (Ed.), Statistics in Action: A Canadian Outlook. CRC Press, Taylor and Francis Group, New York, U.S.A, pp. 371-385.

Kouadio, L., Deo, R.C., Byrareddy, V., Adamowski, J.F., Mushtaq, S., Nguyen, V.P., 2018. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. Comp. Electron. Agric. 155, 324-338. https://doi.org/10.1016/j.compag.2018.10.014

Kouadio, L., Rahn, E., 2020. Agricultural climate risk management and global food security: Recent progress in South-East Asia. In: Lyubchich, V., Gel, Y.R., Kilbourne, H., Miller, T.J., Newlands, K.N., Smith, A.B. (Eds.), Evaluating Climate Change Impacts. Chapman and Hall/CRC, Boca Raton, FL, USA, pp. 347-362.

Kouadio, L., Tixier, P., Byrareddy, V., Marcussen, T., Mushtaq, S., Rapidel, B., Stone, R., 2021. Performance of a process-based model for predicting robusta coffee yield at the regional scale in Vietnam. Ecol. Model. 443, 109469. https://doi.org/10.1016/j.ecolmodel.2021.109469

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89-97. https://doi.org/10.5194/adgeo-5-89-2005

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R news, 2, 18-22.

Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., et al., 2011. Ecological forecasting and data assimilation in a data-rich era. Ecol. Appl. 21, 1429-1442. http://dx.doi.org/10.1890/09-1275.1

Murphy, A.H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecast. 8, 281-293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10, 282-290. https://doi.org/10.1016/0022-1694(70)90255-6

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

Ovalle-Rivera, O., Van Oijen, M., Läderach, P., Roupsard, O., de Melo Virginio Filho, E., Barrios, M., Rapidel, B., 2020. Assessing the accuracy and robustness of a process-based model for coffee agroforestry systems in Central America. Agrofor. Syst. 94, 2033-2051. https://doi.org/10.1007/s10457-020-00521-6

Potgieter, A.B., Everingham, Y.L., Hammer, G.L., 2003. On measuring quality of a probabilistic commodity forecast for a system that incorporates seasonal climate forecasts. Int. J. Climatol. 23, 1195-1210. http://dx.doi.org/10.1002/joc.932

Potgieter, A.B., Hammer, G.L., Doherty, A., de Voil, P., 2005. A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia. Agric. For. Meteorol. 132, 143-153. https://doi.org/10.1016/j.agrformet.2005.07.009

Rahn, E., Vaast, P., Läderach, P., van Asten, P., Jassogne, L., Ghazoul, J., 2018. Exploring adaptation strategies of coffee production to climate change using a process-based model. Ecol. Model. 371, 76-89. https://doi.org/10.1016/j.ecolmodel.2018.01.009

Schauberger, B., Jägermeyr, J., Gornott, C., 2020. A systematic review of local to regional yield forecasting approaches and frequently used data resources. Eur. J. Agron. 120, 126153. https://doi.org/10.1016/j.eja.2020.126153

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

Supit, I., 1997. Predicting national wheat yields using a crop simulation and trend models. Agric. For. Meteorol. 88, 199-214. https://doi.org/10.1016/S0168-1923(97)00037-3

Tiemann, T., Aye, T.M., Dung, N.D., Tien, T.M., Fisher, M., de Paulo, E.N., Oberthür, T., 2018. Crop nutrition for Vietnamese Robusta coffee. Better Crops 102, 20-23. http://www.ipni.net/publication/bettercrops.nsf/article/BC102320

Tien, T.M., 2015. Effects of annual potassium dosage on the yield and quality of Coffea robusta in Vietnam. Research Findings: e-ifc. International Potash Institute, Switzerland.

Trezza, R., Allen, R.G., Kilic, A., Ratcliffe, I., Tasumi, M., 2018. Influence of Landsat revisit frequency on time-integration of evapotranspiration for agricultural water management. In: Bucur, D. (Ed.), Advanced Evapotranspiration Methods and Applications. IntechOpen, London, United Kingdom.

van Bussel, L.G.J., Ewert, F., Leffelaar, P.A., 2011. Effects of data aggregation on simulations of crop phenology. Agric. Ecosyst. Environ. 142, 75-84. http://dx.doi.org/10.1016/j.agee.2010.03.019

van der Velde, M., Nisini, L., 2019. Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015. Agric. Syst. 168, 203-212. https://doi.org/10.1016/j.agsy.2018.06.009

van Oijen, M., Dauzat, J., Harmand, J.-M., Lawson, G., Vaast, P., 2010. Coffee agroforestry systems in Central America: II. Development of a simple process-based model and preliminary results. Agrofor. Syst. 80, 361-378. http://dx.doi.org/10.1007/s10457-010-9291-1

van Wart, J., Grassini, P., Cassman, K.G., 2013. Impact of derived global weather data on simulated crop yields. Glob. Change Biol. 19, 3822-3834. http://dx.doi.org/10.1111/gcb.12302

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

White, J., Berg, A.A., Champagne, C., Zhang, Y., Chipanshi, A., Daneshfar, B., 2020. Improving crop yield forecasts with satellite-based soil moisture estimates: An example for township level canola yield forecasts over the Canadian Prairies. Int. J. Appl. Earth Obs. Geoinf. 89, 102092. https://doi.org/10.1016/j.jag.2020.102092

Wu, B., Meng, J., Li, Q., Yan, N., Du, X., Zhang, M., 2014. Remote sensing-based global crop monitoring: experiences with China's CropWatch system. Int. J. Digit. Earth 7, 113-137. https://doi.org/10.1080/17538947.2013.821185

Zwart, S.J., Bastiaanssen, W.G.M., 2007. SEBAL for detecting spatial variation of water productivity and scope for improvement in eight irrigated wheat systems. Agric. Water Manag. 89, 287-296. https://doi.org/10.1016/j.agwat.2007.02.002