scientific reports

Check for updates

OPEN

Additive modeling of zonal level crop production in Ethiopia

Yidnekachew Mare[1,3]✉, Temesgen Zewotir[2] & Denekew Bitew Belay[1,4]

Crop production is the main source of food security and income for smallholder private farmers in Sub-Saharan countries. To have a sustainable source of food security and economy, it is important to identify covariates that affect crop production linearly and nonlinearly. Annual agricultural sample survey data of eight Meher seasons, from 2012/13 to 2019/20, is used in this study with the main objective of identifying the set of covariates that have linear and nonlinear effects on crop production and estimating their effects using an additive mixed effects model. The minimum, mean, and maximum crop production across the country for the study period were 1.616, 8.693, and 147.843 quintals, respectively, and 50% of the farmers produced less than 6.95 quintals. The histogram, kernel density, and P-P plots suggested that log-transformed crop production is approximately normally distributed. From the competing models' summary statistics, information criteria values, and analysis of variance tests, relaxing the linearity assumption and including a random effect in the model has improved model performance, suggesting the additive mixed effects model best fits the data on hand. Gambella, SNNP, and Oromia regions have significantly different overall mean crop production than the reference in Dire Dawa town. Covariates like year, proportion of female farmers, household age, and UREA fertilizer used have a significant nonlinear effect, while covariates like proportion of educated farmers, area used, and proportion of farmers who received credit service have a significant linear effect on log crop production. The basic model assumptions are not violated, so the final additive mixed effects model can be used for prediction and inference purposes. We recommend farmers use more croplands, indigenous seeds, and UREA fertilizer; practice pure agriculture; and participate in local farmers associations. Policies regarding the participation of female and educated farmers, the implementation and effectiveness of credit services and extension programs, private investors to participate in crop production, and the provision of farm inputs to the elderly farmers should be revised.

Keywords Nonlinear effect, Smooth terms, Additive mixed models, Random effect, Crop production

Crop production is the main source of food security and income for the households of smallholder private farmers in Sub-Saharan countries[1–3]. In Ethiopia, the majority of Meher season's crop production is mainly contributed by smallholder private farmers, who play a huge role in food security and the economy of households and the country[4–6]. For the sustainability of food security and the economy, there should be well-organized and managed crop production at the farmer's level and county level to increase the productivity and efficiency of crop production using the available limited resources and inputs[7–15].

It is important to identify covariates affecting crop production under the given circumstances and then implement the findings to have a sustainable source of food security and economy[16–18]. Modeling the relationship between potential covariates and crop production can help identify the most significant covariates and make an appropriate decision[19–21]. Statistical linear and non-linear models are among the possible models for studying the relationships between the response variable and covariates; for the longitudinal dataset to address the correlations between repeated measures from the same subject (in our case, same subject will be each zone), mixed models are appropriate[22–26].

Linear mixed models, which are extensions of linear models, assume the expected values of the responses are linearly related to the covariates[25,27]. This assumption may not hold for all covariates in the model, i.e., crop production is not expected to increase or decrease infinitely with the changes in the covariates.

[1]Department of Statistics, College of Science, Bahir Dar University, P.O. Box 79, Bahir Dar, Ethiopia. [2]School of Mathematics, Statistics and Computer Science, College of Agriculture Engineering and Science, University of KwaZulu-Natal, Durban, South Africa. [3]Department of Statistics, College of Natural Sciences, Arba Minch University, Arba Minch, Ethiopia. [4]Department of Statistics, University of Pretoria, Pretoria, South Africa. ✉email: yidnekachew.mare08@gmail.com

Different studies have been conducted recently to model crop production in Ethiopia using various modeling approaches and covariates. These include an analog approach for seasonal prediction of crop yields[28], a crop simulation model to evaluate the influence of water management practices and multiple cropping systems on crop yields[29], a multi-crop model ensemble to assess the impact of climate change on the growth and yield of maize and wheat[30], a generalize method of moments and logit to investigate the effect of livestock size and adoption of mixed farming on crop yields[31], and a linear mixed model using longitudinal data to identify covariates linearly affecting crop production[32]. These studies have their own limitations, either on methodological approach, area and time coverage, or use of fewer covariates.

A previous study by the same authors using a linear mixed model showed that covariates like area used, farmers association production, UREA fertilizer used, indigenous seed used, farmer (household) age, crop damage percentage, improved seed used, year (Meher season), and proportion of farmers who practice pure agriculture were found to have significant linear effects on crop production. In this case, we are forcing these covariates to have a linear functional relationship with crop production. Due to this restriction, we are facing two major problems: the first one is that for significant covariates, the true underlying functional relationship may not be linear, and the other is that some important covariates are not found significant in explaining crop production[32]. So, to accommodate more covariates and relax the linearity functional assumption between the response variable and one or more of the covariates, we use additive mixed models[27,33–35].

The contribution of our study to the field is in its approach to use the additive mixed effects (fixed and random terms) modeling of the Ethiopian annual agricultural sample survey data with the main objective to identify the set of covariates that have linear and nonlinear effects on crop production and estimate their effects based on the final fitted additive mixed model.

Methodology

The study areas are ninety zones in Ethiopia, from which agricultural sample surveys are conducted by the Central Statistics Agency (CSA). These zones are from nine administrative regions (Tigray, Afar, Amhara, Oromia, Somali, Benishangul-Gumuz, Southern Nations Nationalities and Peoples, Gambella, and Harari) and one administrative town, Dire Dawa (Fig. 1).

Meher seasons' agricultural sample survey datasets for eight years, from 2012/13 to 2019/20, collected from smallholder private farmers from each zone by interviewing the selected agricultural holders and physically measuring their plots or parcels and crop cut experiments, are used and analyzed to attain the main objective of this study. To select the sample, a stratified, two-stage cluster sample design was implemented. Census Enumeration Areas (EAs) were taken to be the primary sampling units (PSUs), and agricultural households (HHs) were the secondary sampling units (SSUs). Meher season crop is any temporary crop harvested between the months of September and February, while Belg season crop is any temporary crop harvested between the months of March and August[36].

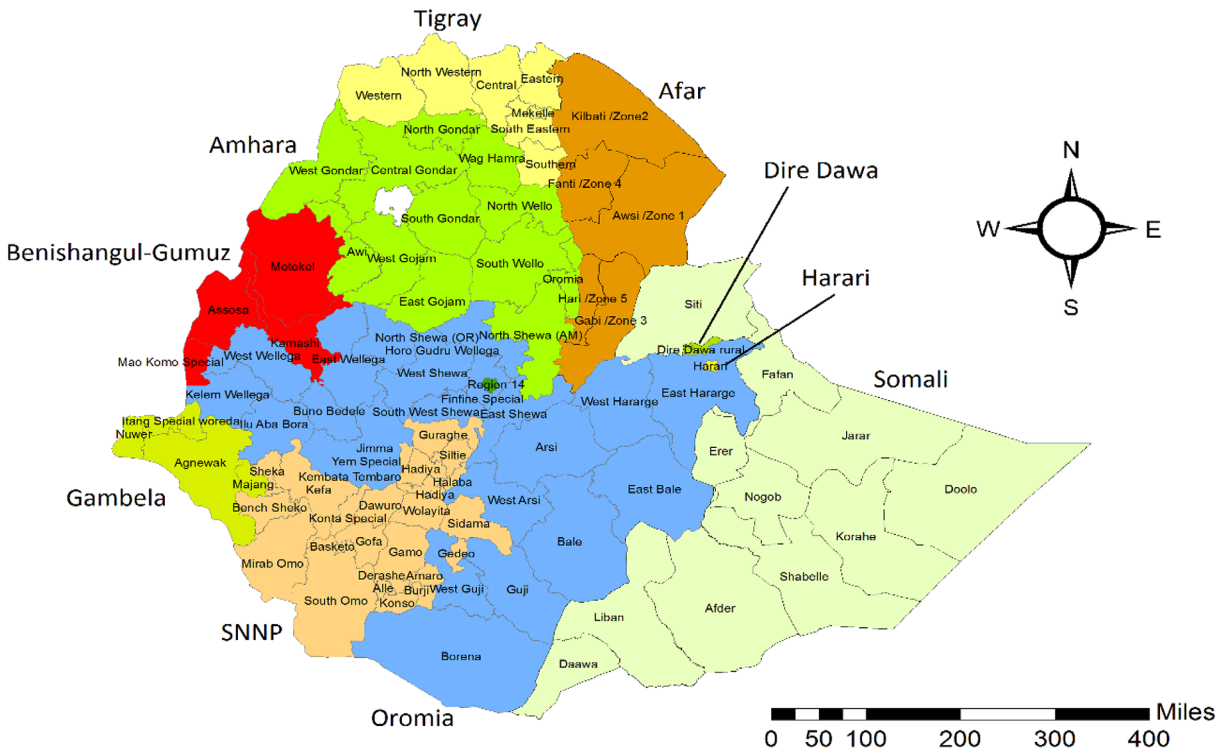

Fig. 1. Administrative regions and zones of Ethiopia.

For this study, the mean (weighted average per zone per house) of continuous variables and weighted percentages or proportions of levels of categorical variables are used, and the response variable, crop production, is log-transformed to make it normally distributed. The 'mgcv' package in R (4.2.3) software is used for all data analysis and model adequacy checking.

The independent variables or covariates used in this study are listed with their short name as follows: Meher Season (Year_Cnt), Proportion of female farmers (Female_Prop), Proportion of educated farmers (Educ_Prop), Proportion of farmers who practice pure agriculture (Field type) (PureAggr_Prop), Proportion of farmers who practice other (not mixed and pure) agriculture (OtherAggr_Prop), Proportion of farmers who hold only crop (CropHold_Prop), Proportion of farmers who hold both crop and livestock (BothHold_Prop), Proportion of farmers who own the farm privately (PrivateOwner_Prop), Proportion of farmers who get credit service (Credit_Prop), Proportion of farmers who received advisory service (Advice_Prop), Proportion of farmers who practice crop rotation (CropRotat_Prop), Proportion of farmers who involved in extension program (Extension_Prop), Proportion of farmers who used irrigation (Irrigation_Prop), Proportion of farmers who prevented their farmland from erosion (PrevErosion_Prop), Proportion of farmers who use broadcast method to sow seeds (BroadSow_Prop), Proportion of farmers who used improved seeds (kg) (ImpSeed_Prop), Proportion of farmers who reported crop damage (CropDamage_Prop), Proportion of farmers who took prevention measure for crop damage (MeasureDamage_Prop), Proportion of farmers who used any type of fertilizer (FertUse_Prop), Household size (HH_Size), Household age (HH_Age), Area (farmland) used (hectares) (Area), Development agent production (quintals) (DA_Product), Farmers association production (quintals) (FA_Product), Improved seed used (kg) (Imp_Seed), Indigenous seed used (kg) (Ind_Seed), UREA fertilizer only used (kg) (UREA_Fert), DAP fertilizer only used (kg) (DAP_Fert), Other fertilizer (except UREA and DAP) used (kg) (Other_Fert), Percent of damaged crop (Dam_Perc), Number of oxen in the household (Num_OXEN), and Number of fruit bearing trees (FruBear_Trees).

Generalized linear Models are used to model the linear functional relationships between the response variable and all the covariates in which the response variable is from the exponential family distributions[26,27,37]. Generalized additive models, which are extensions of generalized linear models, involve the sum of smooth functions of one or more of the covariates to explain the variability in the response variable[27,38].

Generalized additive mixed models (GAMMs) are an additive extension of generalized linear mixed models (GLMMs), i.e., random effects are added to the generalized additive models (GAM) to address the variability and dependency of observations on the same subject and allow non-linear functional relationships between the response variable and one or more of the covariates[27,34,35,38–40].

For the response variable from exponential family distributions, the generalized additive mixed model is formed by introducing smooth functions for one or more of the covariates into the generalized linear mixed model given by Eq. (1)[27].

$$g\left(\mu_i\right) = \boldsymbol{X}_i\beta + \sum_{k=1} f_k\left(x_{ki}^*\right) + \boldsymbol{Z}_i\gamma \tag{1}$$

where $\mu_i \equiv E(y_i)$ and.

$g(\cdot)$ is a monotonic differentiable link function.

y_i is independently distributed with some exponential family distribution.

$\boldsymbol{y}$ is a vector of the response variable ($n \times 1$).

$\boldsymbol{X}_i$ is the i^{th} row of the known design matrix ($\boldsymbol{X}_{\text{n}\times\text{c}}$) of fixed effects.

$\boldsymbol{\beta}$ is a vector of unknown fixed or constant effects ($\text{c} \times 1$).

$f_k(\cdot)$ is a smooth function of the covariate x_k^*

$\boldsymbol{Z}_i$ is the i^{th} row of the known design matrix ($\boldsymbol{Z}_{\text{n}\times\text{q}}$) of random effects.

$\boldsymbol{\gamma}$ is a vector of unknown random effects ($\text{q} \times 1$).

For the normally distributed response variable with the identity link function, the expression $g(\mu_i)$ in Eq. (1) will be replaced by y_i, and we will have the additive mixed model given by Eq. (2).

$$y_i = \boldsymbol{X}_i\beta + \sum_{k=1} f_k\left(x_{ki}^*\right) + \boldsymbol{Z}_i\gamma + \varepsilon_i \tag{2}$$

where $\boldsymbol{\varepsilon}_i$ is an unknown random error for the i^{th} observation ($n \times 1$).

In this study, a random intercept additive mixed model is considered in which, log crop production is the response variable, region (fixed effect) and the rest covariates (smooth terms) are used as fixed terms having linear or nonlinear effects on log crop production while zone is used as a random term (grouping factor) and there are ninety columns (one column per zone as we are considering only random intercept) in the **Z** design matrix of random effects. For a total of 598 observations from ninety zones, the final model is given by Eq. (3).

$$y_{(598\times1)} = X_{(598\times10)}\beta_{(10\times1)} + \sum_{k=1}^{32} f_k\left(x_{ki}^*\right) + Z_{(598\times90)}\gamma_{(90\times1)} + \varepsilon_{(598\times1)} \tag{3}$$

In the above three equations, (1), (2), and (3), the centered twice-differentiable smooth function $f_k(\cdot)$ is the weighted sum of several k-basis functions defined over the covariate x_k^*, and this smoothing function can be for one or more covariates and/or factors in the model.

$$f\left(x_{ki}^*\right) = \sum_{k=0} A_k\left(x_k^*\right)\beta_k \tag{4}$$

where A_k and β_k in Eq. (4) are the k^{th} basis function and the corresponding weight or coefficient, respectively, for the given covariate x_k^*.

$$A_k\left(x_k^*\right) = x_k^{\alpha}\ for\ a = 0, 1, 2, 3, \tag{5}$$

If all the smooth functions $f_k(\cdot)$ in the above two equations, (4) and (5), are linear (i.e., only intercept and one slope), then the models reduce to a generalized linear mixed model and a linear mixed model, respectively.

Model assumptions

The assumptions for the additive mixed models are the same as for the linear mixed models, except for the linearity assumption between the response variable and one or more of the covariates and/or factors[27].

- The random effects and random errors are normally distributed with a mean of zero and covariance matrices of **G** and **R**, respectively, $\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$.

  where **G** is the variance-covariance matrix of random effects (random intercepts and/or one or more random slopes).

  R is the variance-covariance matrix of repeated measurements of the same subject over time, space or condition.

- Random effects and random errors are independent of each other, i.e., Cov $(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = \mathbf{0}$.

  → The vector of expected values is $E\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$ and

  → The variance covariance matrix is $Var\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}$

  → Therefore, we have $\begin{pmatrix}\gamma\\ \varepsilon\end{pmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathbf{G} & 0\\ 0 & \mathbf{R}\end{bmatrix}\right)$

- The expected mean values of the responses are linearly and additively related to one or more predictor variable(s).
- The smooth functions, $f_k(\cdot)$, used to model the functional relationships between covariates and response variable are smooth and continuous.
- Identifiability constraints (sum-to-zero) on the smooth functions is $\sum_k f_k\left(x_{ki}^*\right) = 0$.
- The variances of responses are Var $(\boldsymbol{y})$ = Var $(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}) = \mathbf{ZGZ}' + \mathbf{R}$, in which $\mathbf{ZGZ}'$ is the between-subjects (level-2) component and **R** is the within-subject (level- 1) component; let the sum equal to **V** then we can have $\boldsymbol{y} \sim N\left(\boldsymbol{X}\beta + \sum_{k=1} f_k\left(x_k^*\right), \mathbf{ZGZ}' + \mathbf{R}\right) \rightarrow \boldsymbol{y} \sim N\left(\boldsymbol{X}\beta + \sum_{k=1} f_k\left(x_k^*\right), V\right)$.

Parameters estimation

For the gaussian response variable, with the identity link function, restricted (residual) maximum likelihood (REML) is used to estimate the variance-covariance components as it gives unbiased estimation and best works with unbalanced dataset, and for the smooth terms, among the potential available techniques, thin plate penalized regression splines (appropriate for large dataset and covariates, its computational efficiency and avoid problems of knot placement) are used to determine the smooth curves from the data in such a way to balance the goodness of fit and wiggliness of the smooth curves[27,33,41].

Models comparison and diagnostics

For model comparison, we can use the usual summary statistics values (R^2-adj, Deviance Explained, and Residuals standard error, in which the first two are expected to be large enough while the third value is small for the model to be a good fit), information criteria values (Akaike Information Criteria (AIC) and/or Bayesian Information Criteria (BIC), a model with the smallest value best fits the data), and analysis of variance (ANOVA-test for comparing models based on the parameters estimated and significant F-test values suggest the importance of inclusion of additional parameters in the model specification)[27].

For model diagnostics, we can use residual and random effects plots (histograms, residual Quantile-Quantile (Q-Q) plots, observed versus fitted values, and residuals versus linear predictors). The histogram and Q-Q plot should show the agreement of the observed distribution with the theoretical normal distribution, as residuals are assumed to follow a normal distribution. Observed versus fitted values should show a strong relationship, and all values should lie on or very close to the diagonal straight line to check the accuracy of the fitted model. In the residuals versus linear predictors plot, the scatter of the residuals and predictor values should be evenly centered around zero, and the spread should be more or less within the horizontal band, i.e., without indication of any pattern of increasing or decreasing or spreading wide or narrow along the x-axis values, to ensure constant variance.

0% (Min)	25% (Q1)	Mean	50% (Median)	75% (Q3)	100% (Max)	Sd
1.616	4.716	8.693	6.950	10.828	147.843	8.262

Table 1. Summary of crop production from all zones for the study period.

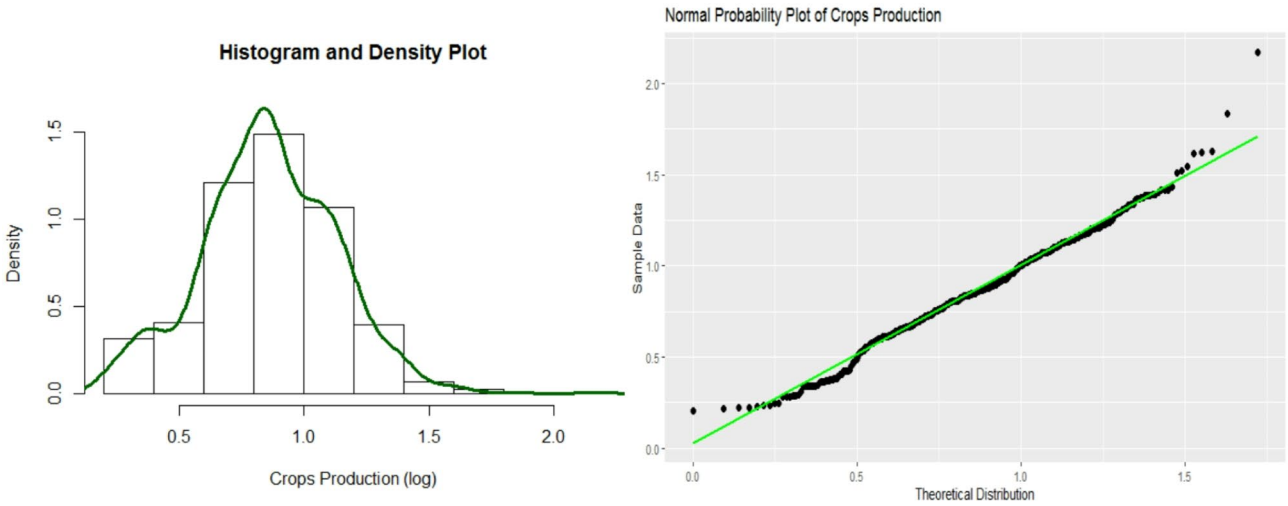

Fig. 2. Histogram and density plot and normal probability plot of log crop production.

Finally, for the assumption of normal distribution of the random effects, a Q-Q plot of random effects is plotted against the theoretical Gaussian quintiles, and to guarantee this assumption, the plots should lie on or very close to the diagonal straight line[27].

Results and discussion

Table 1; Fig. 2 given below show the summary statistics and histogram and normal probability plots of crop production and log crop production, respectively, based on the total observations.

The overall mean and standard deviation of crop production across all zones for the study period (i.e., summaries of aggregated crop production per zone per household) are 8.693 and 8.262 quintals (1 quintal equals 100 kg), respectively; the minimum and maximum productions observed are 1.616 and 147.843 quintals, respectively; the first, second, and third quartiles of crop production are 4.716, 6.950, and 10.828 quintals, respectively; and we can see half of the farmers produced less than the overall mean crop production.

From the histogram, density plot, and normal probability plots, we can see the log transformation of crop production is following a normal distribution, with only very few observations at both extremes indicating outliers.

The assumption that there is a linear relationship between covariates and the log of crop production is assessed graphically and is not justified, at least for some of the covariates considered in the model. Figure 3, given below, shows the non-linear relationship (departures from linearity) for the selected covariates (indigenous seed used (kg), UREA fertilizer used (kg), and proportion of farmers who used any type of fertilizer) with the log transformed crop production without the effect of other covariates.

Figure 4 given below shows the box plot of log transformed crop production for each zone, and from the plot, we can see repeated observations from the same zone are more similar to each other than observations from different zones, indicating dependency (correlation) within the zone and variation between zones, i.e., by-zone variability of crop production.

Three different models were fitted for all covariates to compare and select the one that best fits the data, and the summary results from these fitted models are displayed under Tables 2 and 3.

Table 2 shows the comparison of simple model with the more complex model in terms of parameters (predictors) and we can see that linear fixed effects model is better than intercept only model ($F = 141.5835$, $df = 41$, p-value $= 0.000$), additive fixed effects model is better than linear fixed effects model ($F = 13.0218$, $df = 51.079$, p-value $= 0.000$) and finally additive mixed effects model is better that additive fixed effects model ($F = 7.0986$, $df = 52.745$, p-value $= 0.000$) in explaining the variability in crop production; Table 3 shows the summary results of these three fitted models and we can see additive mixed effects model has the highest adjusted R-Sq. (0.918) and Deviance Explained (93.80%) values and the smallest Res.SE (1.78), Akaike Information Criterion (AIC = − 1130.5588) and REML (1469.7) values. Both tables suggest that including more predictors, allowing predictors to have nonlinear effects, and considering random effects in the model gives substantially improved performance of the fitted model[42,43].

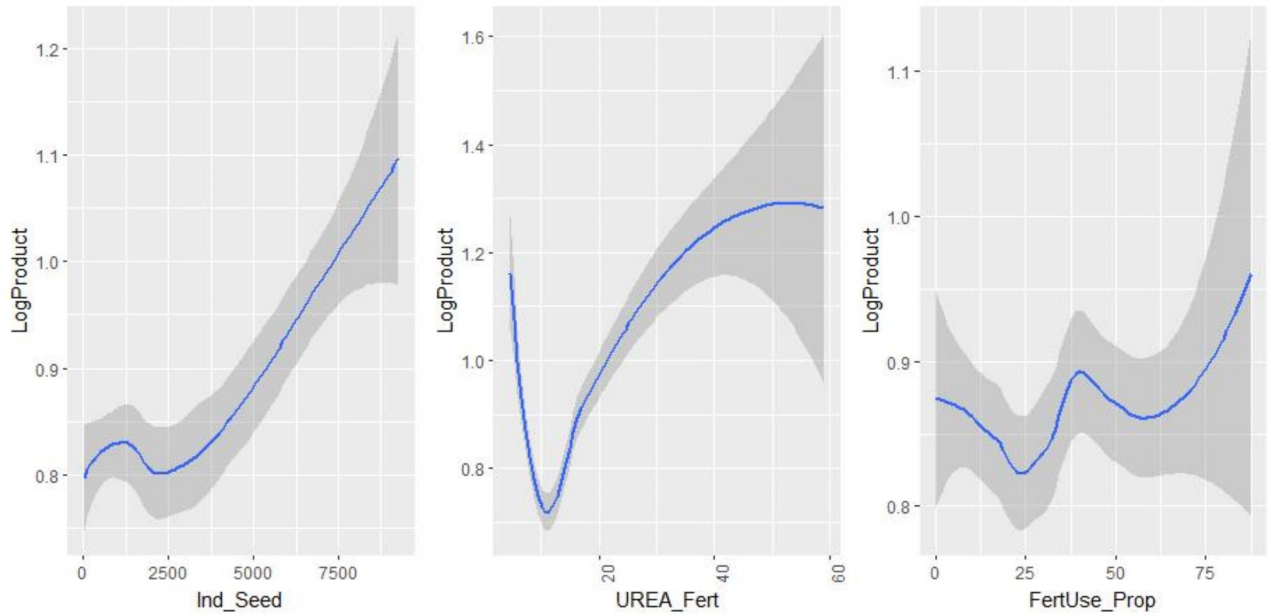

Fig. 3. Smooth plots for selected covariates against the log crop production.

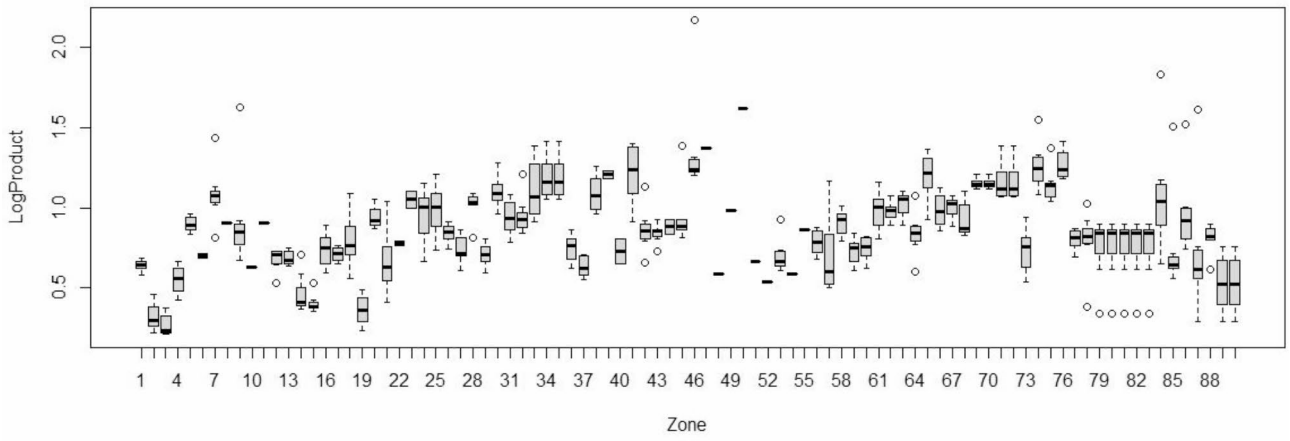

Fig. 4. Box plot of the log crop production for each zone.

Fitted model	Res.DF	RSS	DF	Sum of Sq	F	Pr(> F)
Linear fixed effects	556	4737.8	41	18436.6	141.5835	< 2.2e- 16***
Additive fixed effects	504.92	2625.3	51.079	2112.5	13.0218	< 2.2e- 16***
Additive mixed effects	452.18	1436.1	52.745	1189.2	7.0986	< 2.2e- 16***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Table 2. ANOVA table for comparing the fitted models.

Based on the above figures and model summary tables, we can justify that among the competing models, the additive mixed effects model for a Gaussian response variable is appropriate to fit the data and identify the effect of covariates on log crop production by letting region have a linear or parametric effect while all the rest covariates have a linear or non-linear or nonparametric effect, and let the data itself identify the underling relationship between these covariates and log crop production by including the by-zone random effect (random intercept model).

Figure 5, given below, shows the summed effect of the covariate proportion of farmers who used any type of fertilizer from the additive fixed effects (left panel) and additive mixed effects (right panel) models. From the

Fitted model	*R*-sq. (adj)	Res.SE	Deviance explained	REML	Scale Est.	DF	AIC
Linear fixed effects	0.780	2.92	79.6%	1700.5	8.5213	43	− 655.4206
Additive fixed effects	0.866	2.28	88.7%	1533.8	5.1994	111.2244	− 872.0211
Additive mixed effects	0.918	1.78	93.80%	1469.7	3.176	162.3259	− 1130.5588

Table 3. AIC and model summary values of the fitted models.

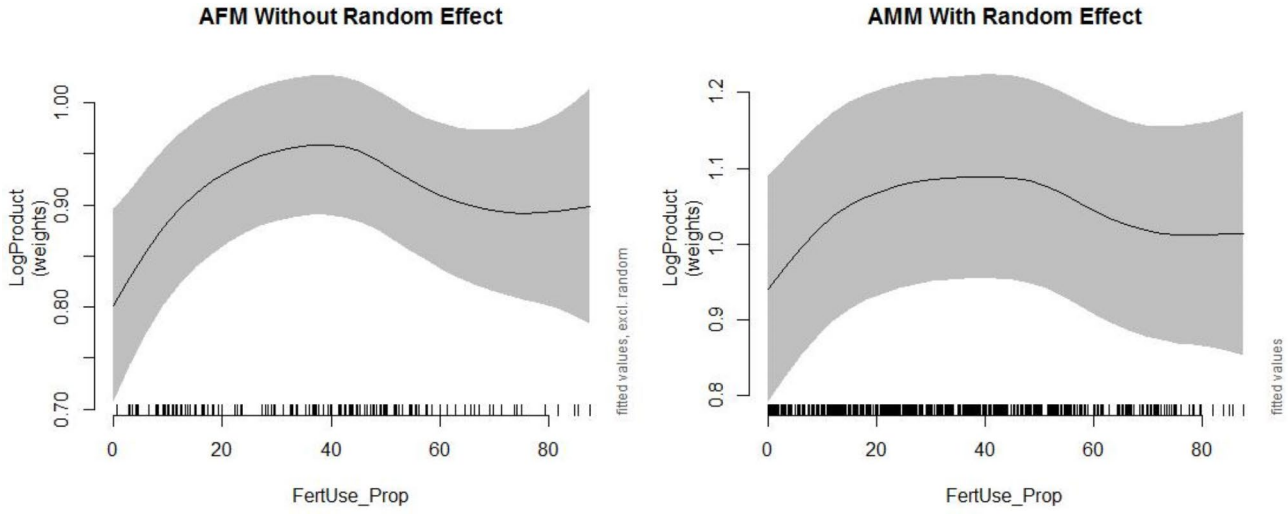

Fig. 5. Smooths showing the summed effect without (left) and with (right) random effect.

two panels, we can see that the effect of the covariate (right panel) on the fitted value of log crop production has increased throughout its range values for a given values of other covariates in each case and this justifies the importance of proper inclusion of the random effects (zone) in the model.

As there are multiple covariates, a thin plate regression spline smoothing function is used to identify the wiggliness (based on the weighted sum of basis functions, and there should be a sufficiently large number of basis functions for the non-linear relationship to be guaranteed) of the fitted curve for each covariate under the non-parametric part. To estimate variance-covariance parameters, a restricted maximum likelihood (REML) estimation method is used, and the results are presented under Table 4 and discussed as follows:

The upper part of Table 4 shows the parametric estimates of the intercept and the factor Region. From the p-values we can see Gambella, Southern Nations Nationalities and Peoples (SNNP) and Oromia regions have higher overall mean of log of crop production compared to the reference Dire Dawa town at less than 5% level of significance given all the smooth terms and random effects in the model.

The lower part of the table shows the smooth terms which assumed to have a nonlinear flexible effect on log crop production and from the effective degrees of freedom (edf) and corresponding p-values we can decide the wiggliness and significance of these terms, respectively, for each covariate considered in the non-parametric part of the model.

At the bottom of the table, we have the summary statistics which tell us how the model best fits the data in explaining log crop production. The sum of all the parametric, non-parametric, and random terms results in the prediction of the fitted model on the log crop production.

Figure 6a–d, given below, show the directions and significance of partial effects of selected smooth terms in the additive mixed effects model. Partial effects are the isolated contributions of one particular predictor to the model predictions and are centered around a zero-horizontal line (where there is no partial effect of the predictor, so log crop production will be at its overall mean). Partial effects represent deviations from the group or overall mean defined by regions when other covariates are held constant at their most typical value, i.e., the partial effect of a term in the model specification is the contribution of that term to the model predictions indicated by the difference from the overall mean.

Also, we can see rug plots on the x-axis of each panel showing the number of data points, and the 95% confidence bands get wider where there are no or few data points, as we cannot be sure about the relationship based on no or few data points.

Based on the summary results in Table 4 and the partial effects plots in Fig. 6a and b, we can see covariates year (log crop production decreases until 2015/16 (due to the pre and post effects of the El Nino in the country during 2015/16) and starts to increase onwards[32,44–46]), proportion of female farmers, proportion of farmers who practice other (not mixed and pure) agriculture, proportion of farmers who use broadcast method to sow seeds, proportion of farmers who used any type of fertilizer (log crop production increases with increase in

Parametric coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.62663	0.10262	6.106	2.2e- 09***
RegionAfar	0.06801	0.12214	0.557	0.577949
RegionAmhara	0.08891	0.11309	0.786	0.432169
RegionBensh/Gumuz	0.10830	0.12594	0.860	0.390264
RegionGambella	0.50234	0.12833	3.915	0.000105***
RegionHarari	0.24646	0.15029	1.640	0.101715
RegionOromia	0.21970	0.11026	1.993	0.046903*
RegionSNNP	0.35986	0.11018	3.266	0.001173**
RegionSomali	0.20088	0.11282	1.780	0.075669.
RegionTigray	0.13974	0.12382	1.129	0.259707
Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Approximate significance of smooth terms:				
	Effective degrees of freedom(edf)	Ref.df	F	p-value
s(Year_Cnt)	3.636	3.908	15.993	< 2e- 16***
s(Female_Prop)	3.073	3.798	3.409	0.01158*
s(Educ_Prop)	1	1	9.602	0.00206**
s(PureAggr_Prop)	1	1	24.107	1.51e- 06***
s(OtherAggr_Prop)	3.64	4.56	3.582	0.00710**
s(CropHold_Prop)	1.076	1.143	4.177	0.02465*
s(BothHold_Prop)	1	1	0.502	0.47896
s(PrivateOwner_Prop)	1.856	2.356	2.303	0.07827.
s(Credit_Prop)	1	1	4.047	0.04485*
s(Advice_Prop)	3.467	4.361	1.875	0.10692
s(CropRotat_Prop)	2.256	2.825	1.251	0.34837
s(Extension_Prop)	2.269	2.878	1.054	0.44575
s(Irrigation_Prop)	1	1	0.01	0.92239
s(PrevErosion_Prop)	1	1	0.156	0.69317
s(BroadSow_Prop)	6.178	7.289	2.901	0.00530**
s(ImpSeed_Prop)	4.345	5.377	1.616	0.11801
s(CropDamage_Prop)	2.382	3.033	1.5	0.2138
s(MeasureDamage_Prop)	1	1	0.087	0.76872
s(FertUse_Prop)	4.769	5.819	6.118	5.26e- 06***
s(HH_Size)	1	1	0.5	0.47993
s(HH_Age)	4.846	5.995	3.113	0.00518**
s(Area)	1	1	49.553	< 2e- 16***
s(DA_Product)	1	1	1.924	0.1661
s(FA_Product)	3.906	4.764	34.252	< 2e- 16***
s(Imp_Seed)	1	1	0.011	0.91664
s(Ind_Seed)	5.825	6.968	9.791	< 2e- 16***
s(UREA_Fert)	2.305	2.951	8.014	4.28e- 05***
s(DAP_Fert)	2.864	3.645	1.912	0.12448
s(Other_Fert)	1.068	1.128	0.006	0.98871
s(Dam_Perc)	1.001	1.001	2.124	0.14571
s(Num_OXEN)	1.001	1.001	10.147	0.00154**
s(FruBear_Trees)	1	1	0.135	0.71365
s(Zone)	62.063	80	5.241	< 2e- 16***
Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
R-sq.(adj) = 0.918 Deviance explained = 93.8%				
fREML = 1469.7 Scale est. = 3.176 n = 598				

Table 4. Parametric coefficients and smooth terms of the fitted additive mixed model.

proportion to 50% and starts to decrease for more than 50% proportion), household age, farmers association production (quintals), indigenous seed used (kg), and only UREA fertilizer used (kg) have a significant nonlinear relationship/effect (with large edf indicating wiggly smooth curves and small p-values < 0.05) (Fig. 6a) with/on log crop production.

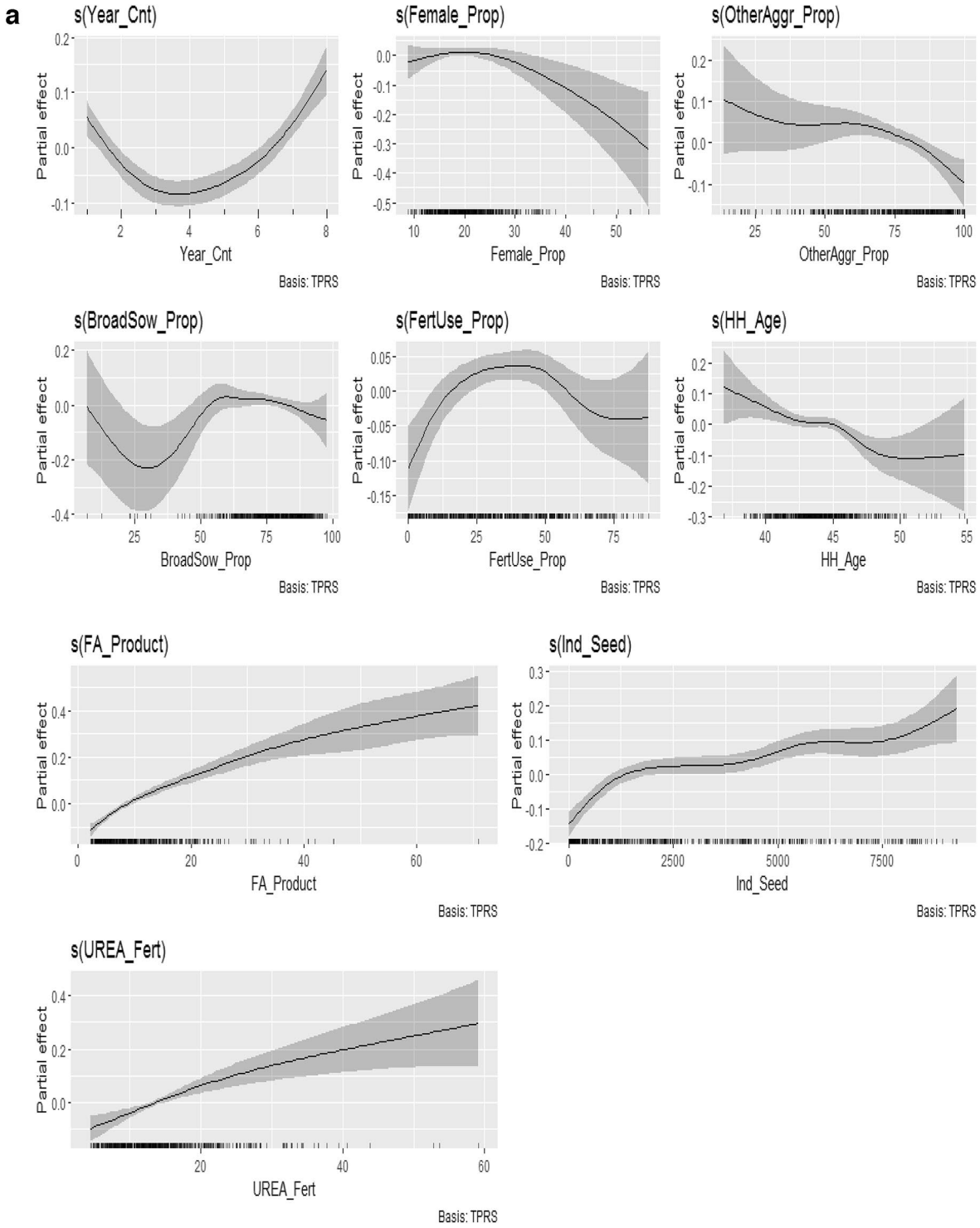

Fig. 6. **a** Partial effects of significant non-linear smooth terms. **b** Partial effects of significant linear smooth terms. **c** Partial effects of non-significant non-linear smooth terms. **d** Partial effects of non-significant linear smooth terms.

Among these significant nonlinearly affecting covariates, year, proportion of farmers who practice other (not mixed and pure) agriculture, proportion of farmers who used any type of fertilizer, household age, farmers association production (quintals), indigenous seed used (kg), and only UREA fertilizer used (kg) were identified to have a significant linear effect on log crop production under a previous study using a linear mixed model analysis[32].

Covariates proportion of educated farmers (decrease of log crop production in zones with a high proportion), proportion of farmers who practice pure agriculture, proportion of farmers who hold only crop, proportion of

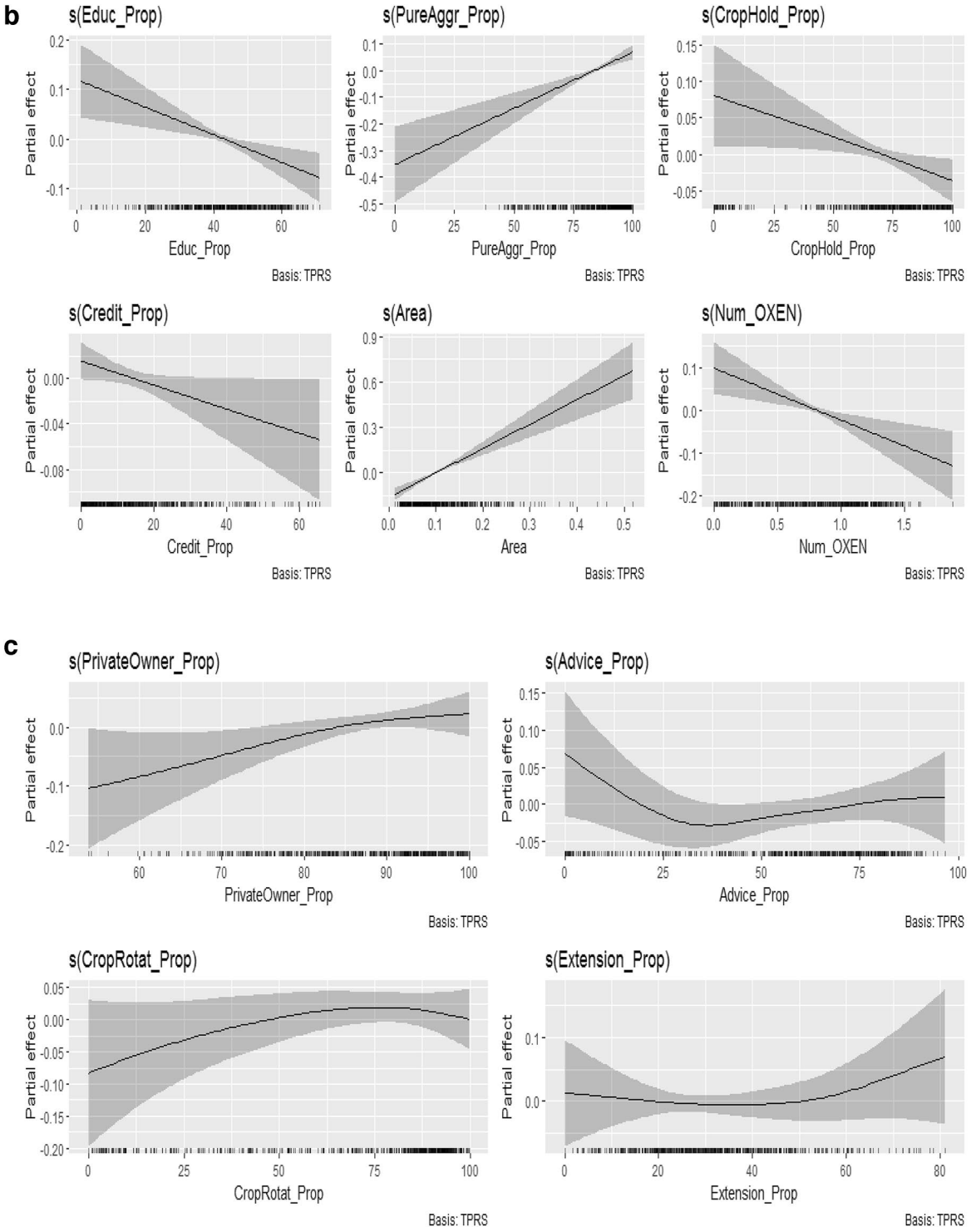

Figure 6. (continued)

farmers who get credit service, area used (hectares) (increase of log crop production in zones with more area used), and number of oxen in the household have a significant linear relationship/effect (one edf with small p-values < 0.05 indicating these covariates have non-zero slopes) (Fig. 6b) with/on log crop production. Number of oxen in the household is negatively affecting crop production in our case, and similarly higher number of livestock in the household was identified to reduce the average crop yield by Wakeyo and Elias[31].

Among these significant linearly affecting covariates, only the proportion of farmers who practice pure agriculture and area used (hectares) were also identified to have a significant linear effect on log crop production, but covariates such as the proportion of educated farmers, the proportion of farmers who hold only crop, the

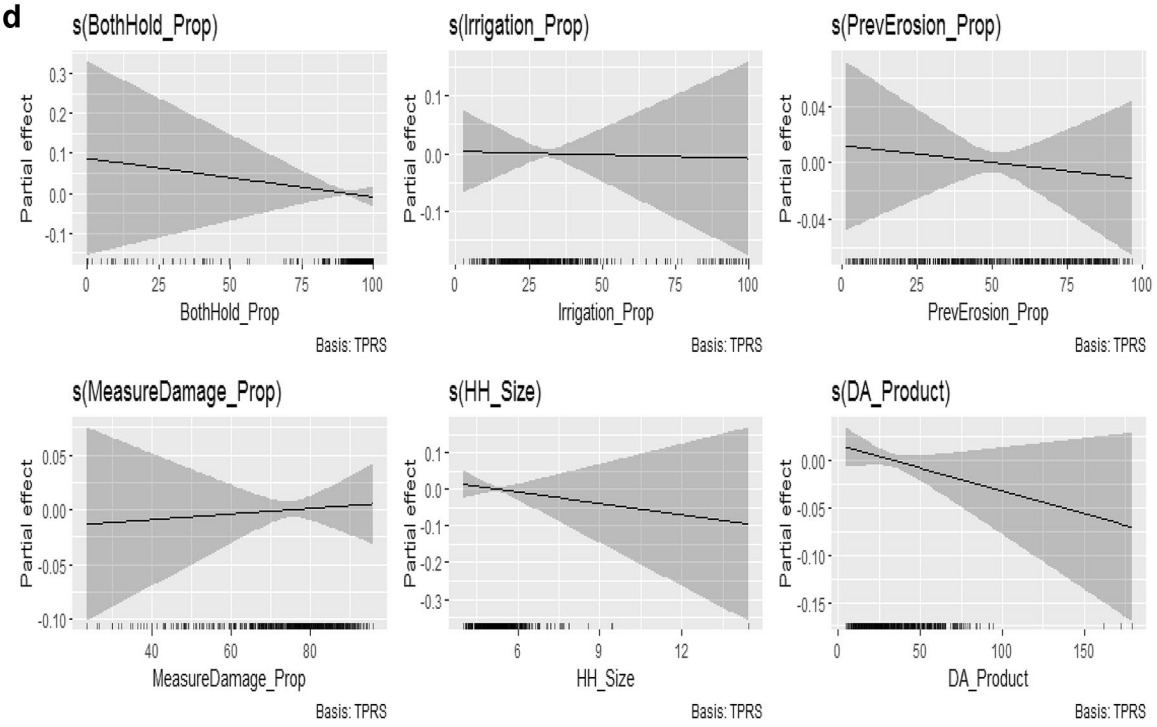

Figure 6. (continued)

proportion of farmers who get credit service, and the number of oxen were not found significant in explaining log crop production under a previous study using a linear mixed model analysis[32].

The covariates proportions of farmers who own the farm privately, who received advisory service, who practice crop rotation, who involved in extension program, who used improved seeds (kg), who reported crop damage, and only DAP fertilizer used (kg) (Fig. 6c) have large edf (wiggly smooth curves) with large p-values (> 0.05), indicating their effect on log crop production is unlikely to be significant. These covariates were also found insignificant in explaining log crop production under a previous study using a linear mixed model analysis[32].

Covariates like proportions of farmers who hold both crop and livestock, who used irrigation, who prevented their farmland from erosion, who took prevention measure for crop damage, household size, development agent production (quintals), and number of fruit bearing trees have one and closer to one edf (indicating a linear effect or relationship) with large p-values (> 0.05 indicating the slopes are zero) so that we fail to reject the null hypothesis of horizontal line indicating log crop production is not varying with the different values of these covariates (Fig. 6d). The insignificant effect of these covariate was also indicated from the linear mixed model analysis[32].

Covariates improved seed used (kg), other fertilizer (except UREA and DAP) used (kg), and percent of damaged crop have one and closer to one edf with large p-values indicating their insignificant effect under the additive mixed model but they have a significant linear effect under linear mixed model analysis[32].

The smooth term s(Zone) at the last row of the nonparametric part (Table 4) has a significant wiggly smooth curve (p-values <0.05), indicating that zones have their own wiggly curve besides their different random intercepts, but these individual wiggly curves have the same trend or slope.

Model diagnostic for the final fitted model is checked using residuals plots and Q-Q plot of the random intercepts shown under Fig. 7a and b given bellow, respectively, and from the upper and lower left panels of Fig. 7a the assumption of normality are justified, from lower right panel we can see a strong relationship between observed and fitted values both lie closely on the diagonal line indicating the accuracy of the fitted model and from upper right panel we see a scatter plot of residuals and predictor values which form a horizontal band centered around zero residual justifying the assumption of constant variance; from Fig. 7b we can see the plot of the ninety random intercept effects for each zone and these are very close to the theoretical gaussian distribution justifying the assumption that random effects are normally distributed. From these plots, we can generalize that the basic assumptions are satisfied and the final fitted model is the best fit for predicting log of crop production.

Conclusion

From the smooth plots of selected covariates and the box plot of log crop production for each zone and the summary results from ANOVA and AIC tables, we have seen that allowing covariates to have a nonlinear effect on log crop production and incorporating zone as a random term in the model specification help to improve the performance of the model, and based on this, we concluded that the additive mixed effects model is the best fit for the data on hand.

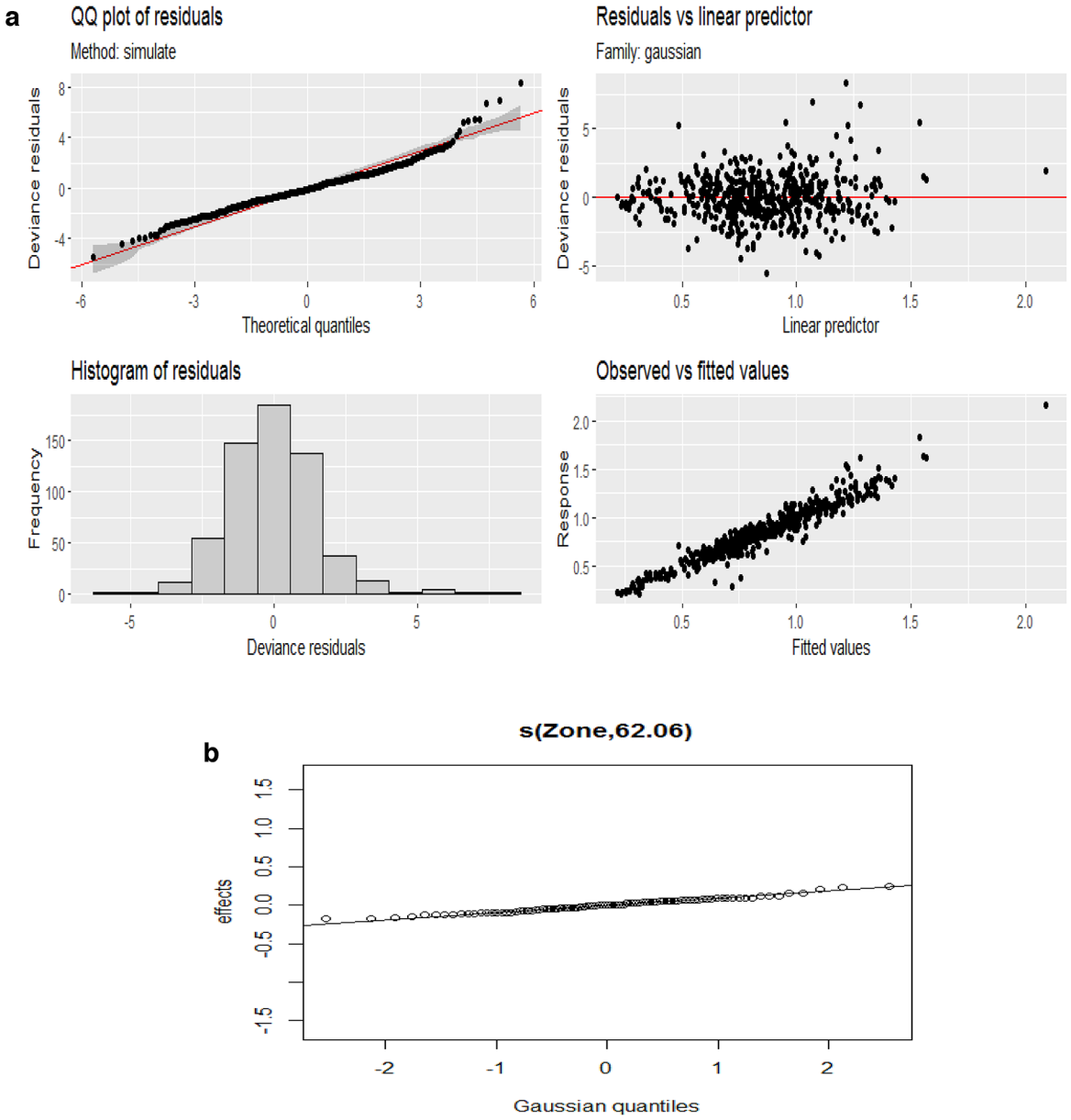

Fig. 7. **a** Model diagnostic plots; Q-Q and histogram of residuals (left panels), residual vs. linear predictor (right-top panel), and response vs. fitted values (right-bottom panel). **b** Quantile-quantile plot of by-zone random intercepts.

The results from additive mixed effects model analysis indicated that Gambella, SNNP and Oromia regions have a significant higher overall mean crop production than the reference Dire Dawa town; year has non-linearly increasing effect on log crop production since 2015/16 Meher season; proportion of female farmers, household age and proportion of farmers who practice other (not mixed and pure) agricultural activities have non-linearly decreasing effect on log crop production while farmers association production condition, UREA fertilizer and indigenous seed used have non-linearly increasing effect on log crop production. Proportion of educated farmers, proportion of farmers getting credit service, proportion of farmers who hold crop (with reference to livestock) and number of oxen in the household have linearly decreasing effect on crop production while area used and proportion of farmers who practice pure agriculture (with reference to mixed agriculture) have linearly increasing effect on log crop production.

To have an increased crop production, the practical implications of our findings are:

i) Farmers should use their farmland in cluster form; they should practice pure agriculture rather than mixed agriculture; they should use UREA fertilizer and indigenous seeds; and farmers associations should be established.

ii) Policies on credit and advisory services, women's participation, extension programs, and educational systems should be revised in such a way as to help farmers in financing their crop production activities and motivate more female farmers and educated farmers to participate in crop production activities.
iii) Zonal- and regional-level governmental bodies from the sector should provide experience sharing between farmers within and across zones, evaluate the effectiveness of agricultural policies and extension programs, invite and encourage private investors to participate and use mechanized crop production to reduce pre- and post-harvest crop waste and damages, and provide different agricultural inputs in quantity and at affordable prices, especially to support elderly farmers.
iv) Further studies in the sector should consider relaxing the linear functional relationships between covariates and crop production and should incorporate random effects to capture unobserved variability between zones and address the dependencies or correlations between repeated observations in the dataset.

The basic model assumptions are assessed using residuals plots and random effect plot, and from these plots, we concluded that these basic assumptions are not violated and the fitted model can be used for prediction and inference purposes.

The additive mixed model is used to identify the direction and significance of the underlying functional relationship between the covariates and log crop production when the relationships are not strictly linear by considering by-zone variation as a random effect and the dependence of observations from each zone. But it has a limitation of assuming observations from different zones are independent. According to Waldo Tobler's first law of geography, observations from nearby area units (zones) are more similar or have a relatively strong spatial autocorrelation than observations from distant area units (zones)[47]. So, the spatial nature of the spatially structured dataset (where the coordinates of each zone are available) leads to spatial dependence or autocorrelation of observations, which needs spatial analysis in the future that addresses the dependence of observations; otherwise, it leads to spurious results. Also, this study has a limitation of incorporating the climate and agro-ecology factors, or covariates, as we have used the annual agricultural sample survey dataset obtained from the Central Statistics Agency (CSA), which doesn't include these important factors or covariates for all Meher seasons.

Data availability

Data will be made available upon formal request and approval of Central Statistical Agency (CSA) of Ethiopia and the authors (Yidnekachew Mare: yidnekachew.mare08@gmail.com).

Received: 30 April 2024; Accepted: 4 April 2025
Published online: 11 April 2025

References

1. Hadebe, S. T., Modi, A. T. & Mabhaudhi, T. Drought tolerance and water use of cereal crops: A focus on sorghum as a food security crop in sub-Saharan Africa. *J. Agron. Crop. Sci.* **203** (3), 177–191 (2017).
2. Aloo, B. N., Mbega, E. R., Tumuhairwe, J. B. & Makumba, B. A. Advancement and practical applications of rhizobacterial biofertilizers for sustainable crop production in sub-Saharan Africa. *Agric. Food Secur.* **10**, 1–12 (2021).
3. Rubhara, T. T., Mudhara, M., Oduniyi, O. S. & Antwi, M. A. Impacts of cash crop production on household food security for smallholder farmers: A case of Shamva District, Zimbabwe. *Agriculture*, *10*(5), p.188. (2020).
4. Evangelista, P., Young, N. & Burnett, J. How will climate change spatially affect agriculture production in Ethiopia? Case studies of important cereal crops. *Clim. Change.* **119**, 855–873 (2013).
5. Wakjira, M. T. et al. Rainfall seasonality and timing: implications for cereal crop production in Ethiopia. *Agricultural and Forest Meteorology*, *310*, p.108633. (2021).
6. ETHIOPIA, T. F. D. R. O. & AGENCY. AGRICULTURAL SAMPLE, C. S. SURVEY. [cited 2023 April]; (2021). Available from: http://www.statsethiopia.gov.et/wp-content/uploads/2021/06/2020_21-2013-E.C-AgSS-Main-Season-Agricultural-Farm-Management-Report.pdf
7. Abhilash, P. C. et al. Sustainability of crop production from polluted lands. *Energy Ecol. Environ.* **1**, 54–65 (2016).
8. Smith, M. The application of Climatic data for planning and management of sustainable rainfed and irrigated crop production. *Agric. For. Meteorol.* **103** (1–2), 99–108 (2000).
9. McLaughlin, D. & Kinzelbach, W. Food security and sustainable resource management. *Water Resour. Res.* **51** (7), 4966–4985 (2015).
10. Lewandowski, I., Härdtlein, M. & Kaltschmitt, M. Sustainable crop production: definition and methodological approach for assessing and implementing sustainability. *Crop Sci.* **39** (1), 184–193 (1999).
11. Kassie, M., Zikhali, P., Manjur, K. & Edwards, S. August. Adoption of sustainable agriculture practices: evidence from a semi-arid region of Ethiopia. In *Natural Resources Forum* Vol. 33 189–198 (Blackwell Publishing Ltd, 2009).
12. Foley, J. A. et al. Solutions for a cultivated planet. *Nature* **478** (7369), 337–342 (2011).
13. Godfray, H. C. J. et al. Food security: the challenge of feeding 9 billion people. *Science* **327** (5967), 812–818 (2010).
14. Grafton, R. Q., Daugbjerg, C. & Qureshi, M. E. Towards food security by 2050. *Food Secur.* **7**, 179–183 (2015).
15. Prosekov, A. Y. & Ivanova, S. A. Food security: the challenge of the present. *Geoforum* **91**, 73–77 (2018).
16. Lobell, D. B. & Gourdji, S. M. The influence of climate change on global crop productivity. *Plant Physiol.* **160** (4), 1686–1697 (2012).
17. Premanandh, J. Factors affecting food security and contribution of modern technologies in food sustainability. *J. Sci. Food. Agric.* **91** (15), 2707–2714 (2011).
18. Liliane, T. N. & Charles, M. S. Factors affecting yield of crops. *Agronomy-climate change & food security*, p.9. (2020).
19. Heil, K., Heinemann, P. & Schmidhalter, U. Modeling the effects of soil variability, topography, and management on the yield of barley. *Frontiers in Environmental Science*, *6*, p.146. (2018).
20. Jiang, H. et al. Understanding the non-stationary relationships between corn yields and meteorology via a spatiotemporally varying coefficient model. *Agricultural and Forest Meteorology*, *301*, p.108340. (2021).
21. Qader, S. H. et al. Exploring the use of Sentinel-2 datasets and environmental variables to model wheat crop yield in smallholder arid and semi-arid farming systems. *Science of The Total Environment*, *869*, p.161716. (2023).
22. Rencher, A. C. & Schaalje, G. B. *Linear Models in Statistics* (Wiley, 2008).
23. Searle, S. R. & Gruber, M. H. J. *Linear Models* (Wiley, 2016).

24. Davidian, M. & Giltinan, D. M. Nonlinear models for repeated measurement data: an overview and update. *J. Agricultural Biol. Environ. Stat.* **8**, 387–419 (2003).
25. West, B. T., Welch, K. B. & Galecki, A. T. *Linear Mixed Models: a Practical Guide Using Statistical Software* (Crc, 2022).
26. McCulloch, C. E. & Searle, S. R. *Generalized, Linear, and Mixed Models* (Wiley, 2004).
27. Wood, S. N. *Generalized Additive Models: an Introduction with R* (CRC, 2017).
28. Yang, M. et al. *Seasonal Prediction of Crop Yields in Ethiopia Using an Analog Approach*331p.109347 (Agricultural and Forest Meteorology, 2023).
29. Araya, A., Prasad, P. V. V., Ciampitti, I. A. & Jha, P. K. Using crop simulation model to evaluate influence of water management practices and multiple cropping systems on crop yields: A case study for Ethiopian highlands. Field Crops Research, 260, p.108004. (2021).
30. Rettie, F. M., Gayler, S., Weber, K. D., Tesfaye, T., Streck, T. & K. and Climate change impact on wheat and maize growth in Ethiopia: A multi-model uncertainty analysis. *PloS One*. **17** (1), e0262951 (2022).
31. Wakeyo, M. B. & Elias, H. Mixed farming system for crop yield improvement and adaptation to climate change: evidence from smallholder farmers in Ethiopia. *Ethiop. J. Econ.* **32** (2), 37–73 (2023).
32. Mare, Y., Belay, D. B. & Zewotir, T. Linear modeling of zonal level crop production in Ethiopia. *Heliyon*, **10**(10). (2024).
33. Hastie, T. J. Generalized additive models. In Statistical Models in S (249–307). eds. John M. Chambers and Trevor J. Hastie. Routledge. (2017).
34. Lin, X. & Zhang, D. Inference in generalized additive mixed models by using smoothing splines. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **61** (2), 381–400 (1999).
35. Baayen, R. H. spsampsps Linke, M. An introduction to the generalized additive model. *A practical handbook of corpus linguistics. New York: Springer*, pp.563–591. (2020).
36. CSA. Area and Production for Major Crops (Private Peasant Holdings, Meher Season) 2019-20 (2012 E.C.). 2020 [cited 2022 March]; Available from: https://www.statsethiopia.gov.et/our-survey-reports/
37. Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *J. Royal Stat. Soc. Ser. A: Stat. Soc.* **135** (3), 370–384 (1972).
38. Chen, C. Generalized additive mixed models. *Commun. Statistics-Theory Methods*. **29** (5–6), 1257–1271 (2000).
39. Series, Q. R. W. *Introduction to Generalized Additive Mixed Models (GAMMs)*. [cited 2023 June]; (2023). Available from: https://r.qcbs.ca/workshop08/book-en/introduction-to-generalized-additive-mixed-models-gamms.html
40. Melesse, S. F. & Zewotir, T. Additive mixed models to study the effect of tree age and Climatic factors on stem radial growth of Eucalyptus trees. *J. Forestry Res.* **31** (2), 463–473 (2020).
41. Maindonald, J. Smoothing terms in gam models. *Center for Mathematics and its applications. Austrian National University*. (2010).
42. Bera, D., Chatterjee, N. D. & Bera, S. Comparative performance of linear regression, polynomial regression and generalized additive model for canopy cover estimation in the dry deciduous forest of West Bengal. *Remote Sensing Applications: Society and Environment*, *22*, p.100502. (2021).
43. Austin, P. C., Ghali, W. A. & Tu, J. V. A comparison of several regression models for analysing cost of CABG surgery. *Stat. Med.* **22** (17), 2799–2815 (2003).
44. Seaward, C. El Niño in Ethiopia: Programme Observations on the Impact of the Ethiopia Drought and Recommendations for Action. Oxfam International. (2016).
45. Gleixner, S., Keenlyside, N., Viste, E. & Korecha, D. The El Niño effect on Ethiopian summer rainfall. *Clim. Dyn.* **49**, 1865–1883 (2017).
46. Ewbank, R., Perez, C., Cornish, H., Worku, M. & Woldetsadik, S. *Building Resilience To El Niño-related Drought: Experiences in Early Warning and Early Action from Nicaragua and Ethiopia*43pp.S345–S367 (Disasters, 2019).
47. Tobler, W. R. A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46** (sup1), 234–240 (1970).

Acknowledgements

We would like to acknowledge Central Statistical Agency (CSA) of Ethiopia for providing and allowing the data that we used in this study.

Author contributions

Y.M., D.B.B., and T.Z. conceptualized the research problem. Y.M. data curation, methodology, software and formal statistical analysis. D.B.B. and T.Z. provided resources and validated the analysis. Y.M. writing-original draft preparation. D.B.B. and T.Z. writing-review and editing. T.Z. supervised the overall work. All authors read and approved the final manuscript.

Funding

This research received no external funding.

Declarations

Competing of interests

The authors declare no competing of interests.

Additional information

Correspondence and requests for materials should be addressed to Y.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025