



OPEN

## Robustly forecasting maize yields in Tanzania based on climatic predictors

Rahel Laudien<sup>1✉</sup>, Bernhard Schauburger<sup>1</sup>, David Makowski<sup>2</sup> & Christoph Gornott<sup>1,3</sup>

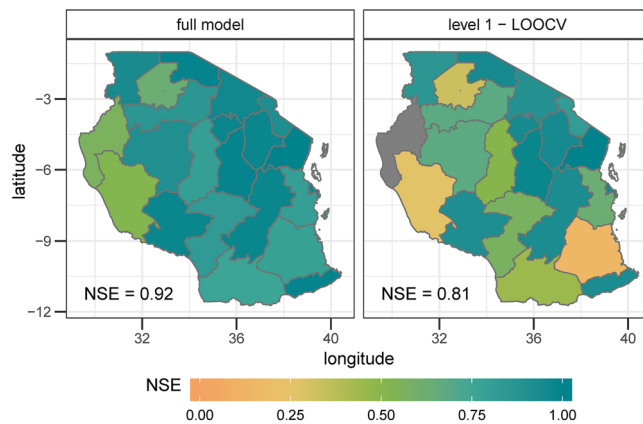
Seasonal yield forecasts are important to support agricultural development programs and can contribute to improved food security in developing countries. Despite their importance, no operational forecasting system on sub-national level is yet in place in Tanzania. We develop a statistical maize yield forecast based on regional yield statistics in Tanzania and climatic predictors, covering the period 2009–2019. We forecast both yield anomalies and absolute yields at the sub-national scale about 6 weeks before the harvest. The forecasted yield anomalies (absolute yields) have a median Nash–Sutcliffe efficiency coefficient of 0.72 (0.79) in the out-of-sample cross validation, which corresponds to a median root mean squared error of 0.13 t/ha for absolute yields. In addition, we perform an out-of-sample variable selection and produce completely independent yield forecasts for the harvest year 2019. Our study is potentially applicable to other countries with short time series of yield data and inaccessible or low quality weather data due to the usage of only global climate data and a strict and transparent assessment of the forecasting skill.

To support food security planning in face of unfavourable weather conditions, accurate yield forecasts at sub-seasonal to seasonal timescales, i.e. from weeks to months ahead, are important. Such yield forecasts can be used for early warning so that actions can be taken before the disaster occurred. On a regional and national scale, a seasonal yield forecast allows to adjust food imports so that food shortages in case of harvest losses or failures can be alleviated. If available on a high spatial resolution, a yield forecast can also inform farm management practices, such as fertilizer use, and decisions on sale prices and grain storage<sup>1,2</sup>.

Despite the importance of yield forecasts, as expressed by the Division of National Food Security of the Ministry of Agriculture in Tanzania (personal communications), no operational yield forecasting system on sub-national level exists up to now. The Famine Early Warning Systems Network (FEWS-net), which provides regular food insecurity reports for East Africa, stopped providing yield forecasts for Tanzania since December 2017. Ogotu et al. (2018)<sup>3</sup> proposed a yield forecasting system for East Africa but this system relies on model simulations that were not validated with observational data at a subnational scale so that the accuracy of the forecast is unknown. Liu and Basso (2020)<sup>2</sup> provide a yield forecast for three case studies in Tanzania based on a process-based model with a lead time of 14–77 days. They calibrate and validate the forecast using survey data at field scale. However, the practical application of this forecast for an operational forecasting system is hampered due to the limited spatial (three regions) and temporal scale (year 2017) and the high potential implementation costs (necessity to collect survey data for every new forecast).

In this study, we provide a statistical yield forecast for the whole country covering the time period from 2009 to 2019, which can be operationalized in a technical and cost-efficient way. We provide a within-season forecast of maize yields at the subnational level about 6 weeks before harvest and relate the yields to weather and sea surface temperature data. We focus on maize (*Zea mays* L.), as this is the main staple crop in Tanzania<sup>4</sup>. We conduct a strict model validation—not only comprising of an out-of-sample validation, but also an out-of-sample variable selection—to assess the forecasting skill.

<sup>1</sup>Potsdam Institute for Climate Impact Research (PIK), P.O. Box 60 12 03, 14412 Potsdam, Germany. <sup>2</sup>National Research Institute for Agriculture, Food and Environment (INRAE), UMR 518 AgroParisTech Université Paris-Saclay, 16 rue Claude Bernard, 75231 Paris Cedex 05, France. <sup>3</sup>Agroecosystem Analysis and Modelling, Faculty of Organic Agricultural Sciences, University of Kassel, Mönchebergstraße 19, 34109 Kassel, Germany. ✉email: laudien@pik-potsdam.de



**Figure 1.** Model performance for yield anomalies from 2009 to 2018 based on selected region-specific weather variables measured in the Nash–Sutcliffe efficiency coefficient (NSE). The NSE can range from 1 (100% agreement between observed and modelled data) to  $-\infty$ ; a value of zero can be interpreted in a way that nothing of the observed variability can be explained by the model. The left panel shows the model performance when the complete time series for each region is included. The right hand panel shows the performance for the leave-one-out cross validation. The colour grey indicates an NSE lower than 0. The median NSE over all regions of Tanzania is shown in the left corner of the panels.

## Results

**The influence of weather on maize yield variability.** In the first step, we assessed how weather influenced maize variability from 2009 to 2018. The median Nash–Sutcliffe efficiency coefficient<sup>5</sup> (NSE) of all models for the estimation results is 0.92 (coined ‘NSE’ hereinafter). The NSE of the level 1 validation (coined ‘NSEv1’) is 0.81. The models show a high performance for almost the whole country (18 out of 21 regions have an NSEv1 of higher than 0.3; Fig. 1). Extreme years, like e.g. the high yields in 2010 in Dar es Salaam and Manyara or low yields in 2016 in Kagera, as well as average yields can be reproduced by the models (Fig. 2).

The variable selection reveals the strong influence of extreme weather events on maize yields in Tanzania. In general, maize yields seem to benefit from higher minimum temperatures, more precipitation and the absence of more than 5 consecutive dry days (Fig. 3).

In the vegetative phase, high rainfall events (precip.p99\_v, i.e. precipitation events with more rain than the 99% percentile) are related with higher yields, whereas the occurrence of consecutive dry days of more than 5 days (cdd5) are mostly negatively correlated with yield. Consecutive dry days play an important role in explaining maize variability. In total, cdd5, cdd10 and cdd15 have been selected 17 times. They are mostly negatively correlated with yields, in particular in the vegetative phase. Temperature events below the 1% percentile of the minimum temperature in the reproductive phase (tas.min.p01\_r) are related with lower yields—indicating that too low minimum temperatures are detrimental for maize yields.

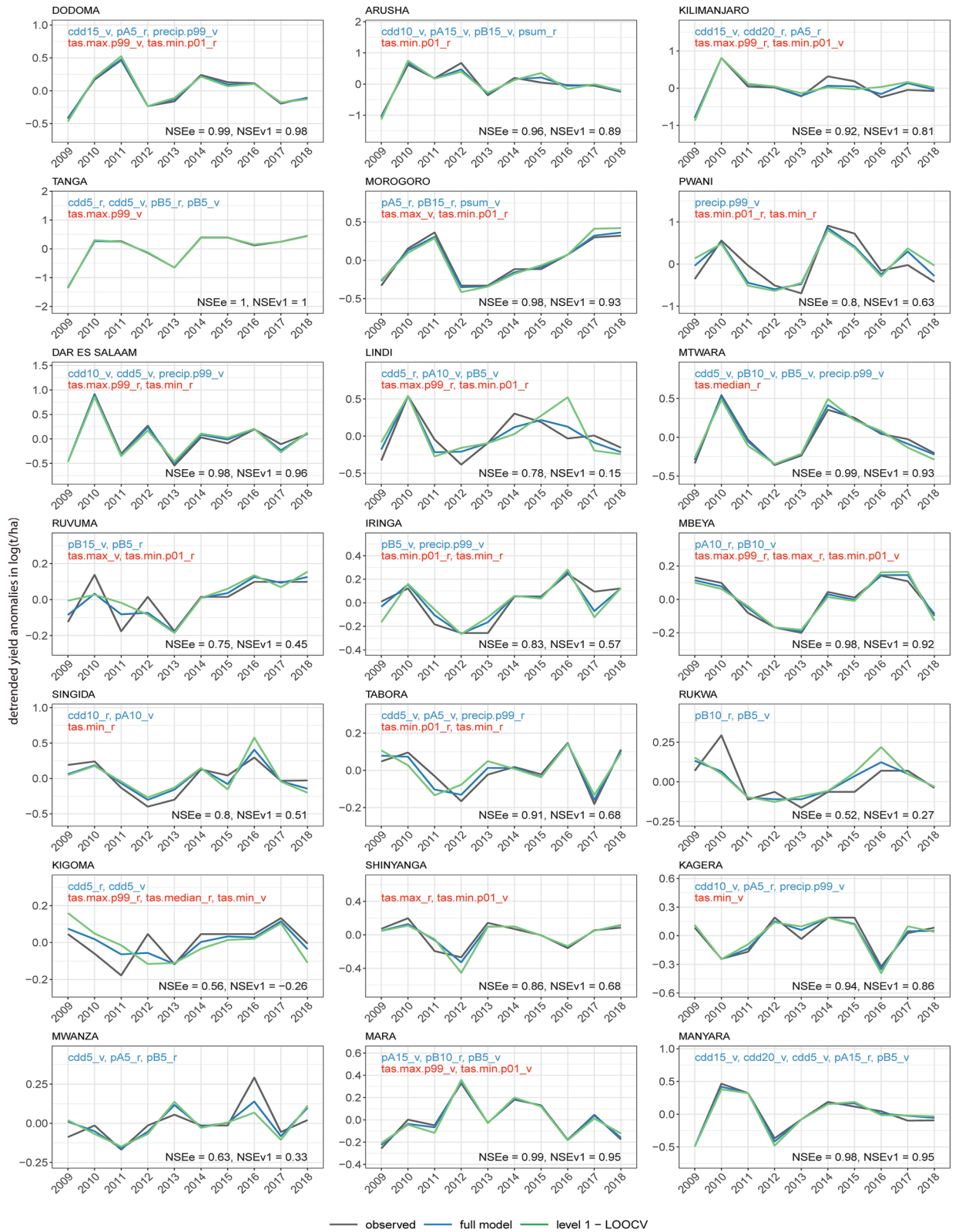
**The influence of sea surface temperatures on maize yield variability.** The usage of sea surface temperature (SST) indicators in a separate, alternative model formulation led to a median model performance of an NSE of 0.29 (level 1-LOOCV), which shows that SST variables can explain a substantial part of yield variability, but that the influence of weather is stronger.

The SST of the West Pacific (WP) with a lead time of 120 days has the strongest influence on maize variability from 2009 to 2018. Maize variability shows a positive correlation with the number of times the SST falls below the 1% percentile of the WP SST (Fig. 4). The SST of the Indian Ocean Dipole (IOD) with a lead time of 30 days influences mostly the bimodal rainfall regions in Tanzania. The number of times the SST of the IOD falls below the 1% percentile is mostly positively correlated with yield, whereas the median IOD shows a mostly negative correlation with yield.

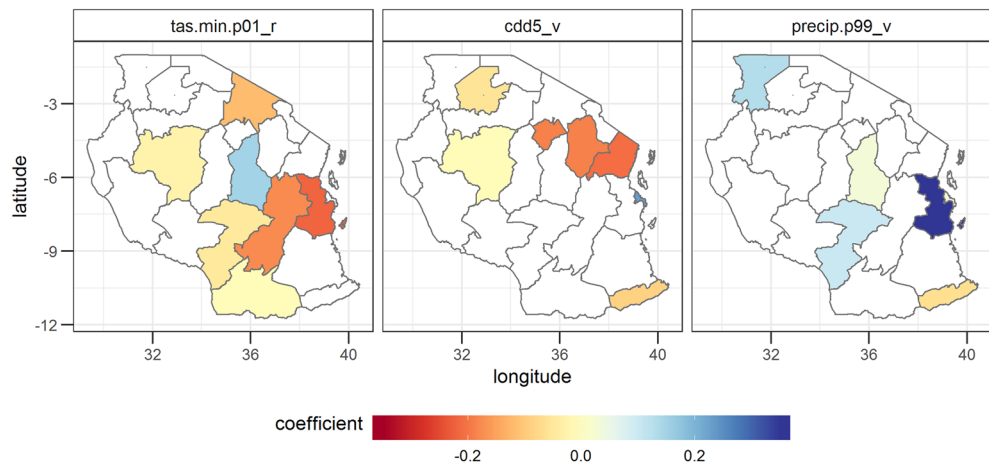
**Forecasting yield anomalies.** We performed a within-season forecast by including the weather and SST variables related to the vegetative phase of the growing season. This allows us to provide a yield forecast around 6 weeks before the calculated harvest date. To assess the hindcast-based operational performance of the forecast, we include the level 1 (i.e. the out-of-sample validation) and level 2 validation (i.e. the out-of-sample variable selection validation).

The forecast shows a high accuracy for the full model (NSE = 0.91) and the level 1 validation (NSE = 0.72). The results of the level 2 validation suggest that the variable selection has a high year-to-year variability so that the performance of this validation is lower. The level 2 validation has a high skill (NSE > 0.3) for 5 regions in Tanzania (Fig. 5). The middle panel of Fig. 8 shows the yearly performance of the forecast of anomalies.

We tested the performance of the forecast compared to a constant model that only takes the mean yield excluding the year that is forecasted as a predictor. The lower RMSE of the forecast compared to a constant model (SI Table 4) underlines the robustness of the forecasting results. A strong correlation between SST and weather



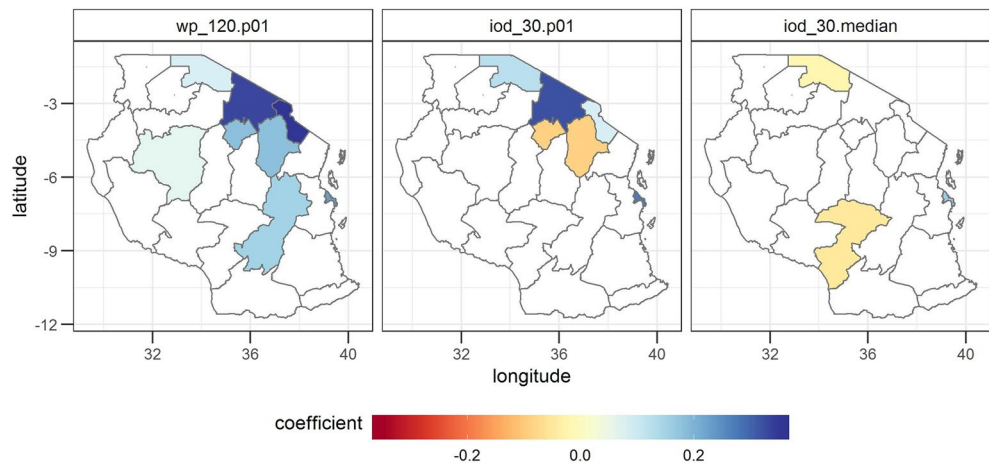
**Figure 2.** Observed and modelled maize yield anomalies in Tanzanian regions based on region-specific selected weather inputs. Grey lines show observed yield anomalies, blue lines show modelled anomalies using the full time series for model development, green lines indicate out-of-sample (level 1-LOOCV) estimates. The y-axis shows detrended yield anomalies in logarithmic form. The names of the inputs are shown in blue for precipitation and red for temperature-related variables. Explanations of abbreviations for input names are provided in supplemental information (SI) Tables 1, 2 and 3.



**Figure 3.** Estimated regression coefficients for the three most often selected variables. The three most often selected variables are temperature events below the 1% minimum temperature percentile in the reproductive phase (*tas.min.p01\_r*), consecutive dry days of more than 5 days in the vegetative phase (*cdd5\_v*) and precipitation events above the 99th precipitation percentile in the vegetative phase (*precip.p99\_v*). Note that this analysis excludes the model coefficients with a performance lower than an NSE of 0.3 in the level 1 validation, because we assumed these models as not robust enough for further analysis. Negative (positive) values indicate that yields tend to decrease (increase) as a function of the values of the variables. The coefficients show standardized values, i.e. they show the change in yield per standard deviation of the input variable. Regions, in which the variable was not selected, are shown in white. The statistical significance of the estimated regression coefficients is shown in SI Fig. 1.

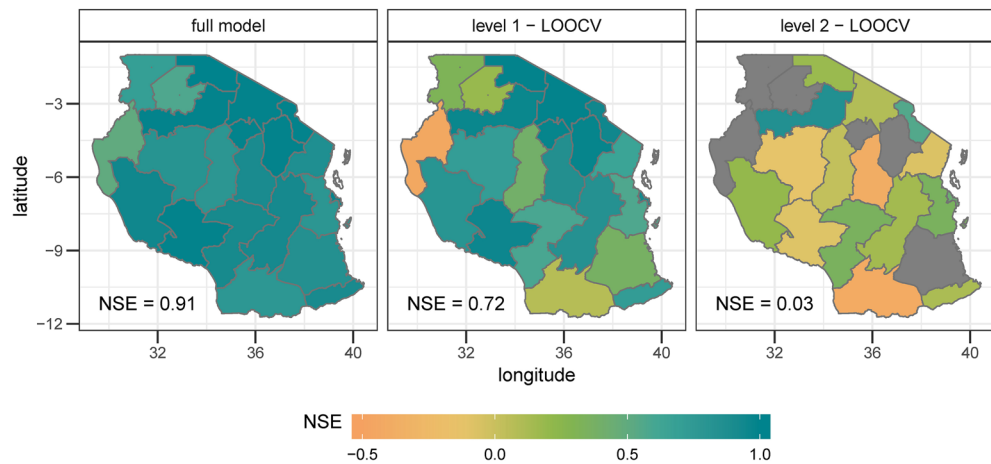
variables exists only in few, non-systematic cases (SI Fig. 4) and is eliminated due to the applied filter during the variable selection. Thus a collinearity between SST and weather variables is not existent.

The variable selection of the forecast reveals similar variables compared to the model based on variables of the vegetative and the reproductive phase. The three most often selected variables are shown in Fig. 6. Precipita-



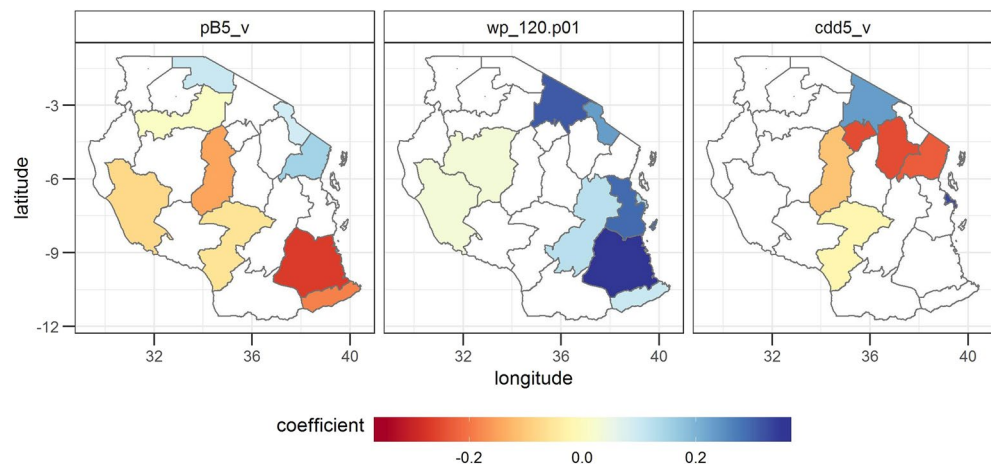
**Figure 4.** Coefficients of the most often selected sea surface temperature (SST) variables. The three most often selected variables are the number of times the SST falls below the 1% percentile of the West Pacific considering a lead time of 120 days (*wp\_120.p01*), the number of times the SST falls below the 1% percentile of the Indian Ocean Dipole considering a lead time of 30 days (*iod\_30.p01*) and the median SST of the IOD considering a lead time of 30 days (*iod\_30.median*). Further explanations can be found under Fig. 3. The statistical significance of the estimated regression coefficients is shown in SI Fig. 2.

tion events below 5 mm (pB5) are positively correlated with maize yields in the North, but negatively correlated with yields in the rest of the country. Whereas the negative correlation could indicate a negative influence of low precipitation events on yields due to insufficient total rainfall amounts in the vegetative phase, the positive



**Figure 5.** Performance of the forecast of anomalies based on selected region-specific weather and SST variables. Areas in grey show an NSE of lower than -0.5. Further explanation can be found under Fig. 2.

correlation could indicate a positive influence of moderate and well distributed rainfall. The number of times the SST in WP falls below the 1% percentile with a lead time of 120 days ( $wp_{120,p01}$ ) is positively correlated with yields, which could be explained by above average rainfall amounts in case of negative SST anomalies in



**Figure 6.** Coefficients of the three most often selected variables in the forecast based on weather and SST variables. Further explanations can be found under Fig. 3. The statistical significance of the estimated regression coefficients is shown in SI Fig. 3.

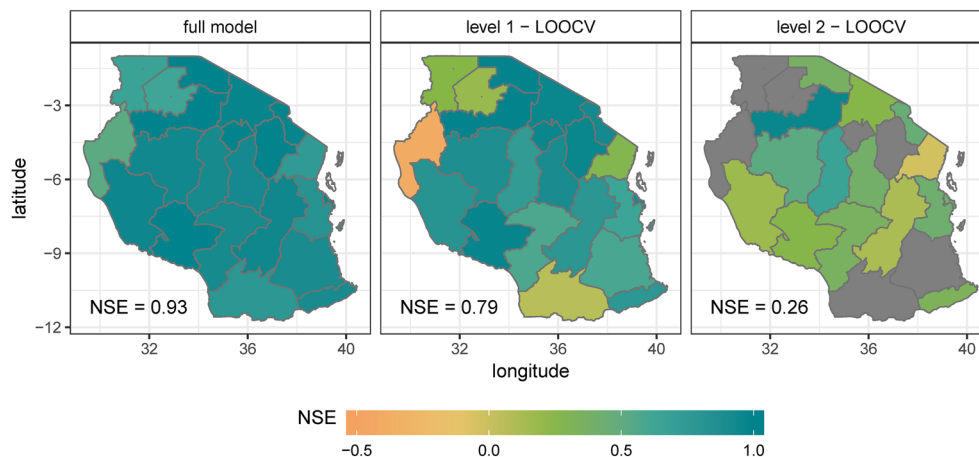
the West Pacific. Dry conditions represented by the number of consecutive dry days of more than 5 days ( $cdd5$ ) are mostly negatively correlated with yields.

**Forecasting absolute yields.** As required for operational forecasting systems, we also provide a forecast of absolute yields. To obtain absolute yields, we added the previously subtracted trend and the mean yield to the forecasted yield anomalies.

The full model evaluation and the level 1 validation indicate a high skill (median NSE of 0.93 and 0.79) of the within-season forecast of absolute yields for the whole country (Fig. 7). The level 2 validation has a median NSE of 0.26 and shows a high performance for 10 of 21 regions, which have an NSE higher than 0.3. The median RMSE is 0.07 t/ha for the estimation, 0.13 t/ha for the level 1 validation and 0.3 t/ha for the level 2 validation.

The reason for the improved forecasting performance is, in most regions, due to the inclusion of the trend, in particular for the level 2 validation (Fig. 8).

**Application of the forecast for year 2019.** We used the model trained on yield data from 2009 to 2018 to provide a completely independent forecast for the harvest year 2019. The forecast of absolute yields for 2019 shows an overall high performance of the forecast in regions with a unimodal rainfall regime (NSE of 0.89 in the



**Figure 7.** Assessment of the forecasted absolute yields. Areas in grey show an NSE of lower than  $-0.5$ .

level 2—LOOCV) and a weak performance in bimodal rainfall regions (NSE of  $-7.73$  in the level 2—LOOCV; Fig. 9).

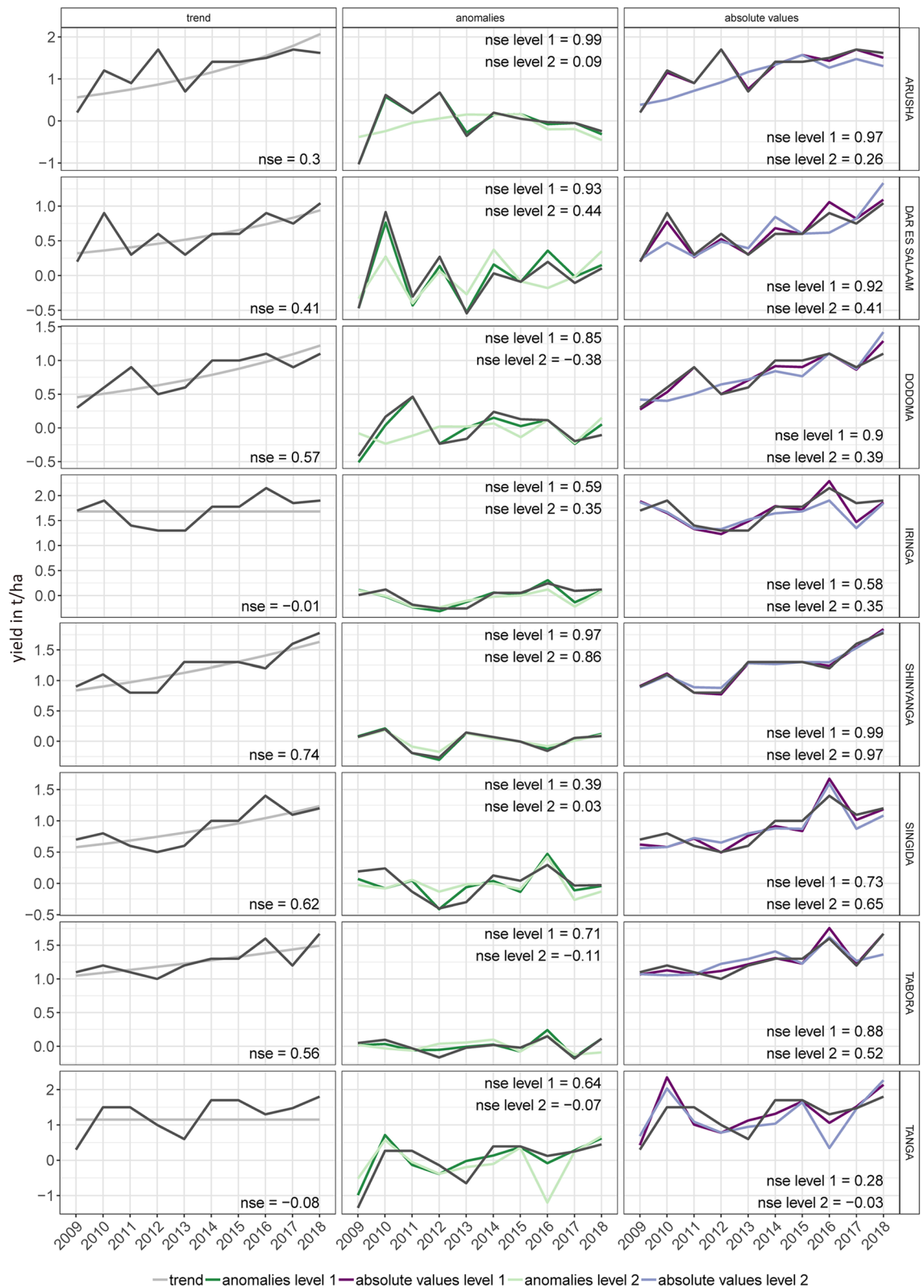
## Discussion

Our study provides a within-season maize yield forecast for entire Tanzania and is, to our best knowledge, the first of its kind. In contrast to existing forecasting studies in East Africa<sup>2,3</sup>, we provide an out-of-sample validation and an out-of-sample variable selection. Furthermore, we test the robustness by providing a completely independent forecast for the harvest year 2019. This strict and transparent assessment of uncertainties of the forecast is particularly important in practice, when e.g. policy makers use a yield forecast to inform food security planning.

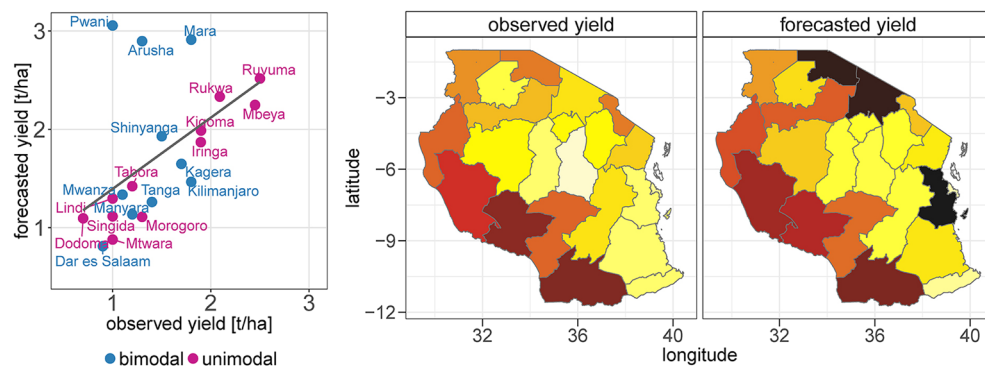
Over all regions, the forecast of anomalies produces a median NSE of 0.91 in the model estimation and a median NSE of 0.72 in the level 1 validation, which indicates an accurate and robust forecast for large parts of the country. The forecast reproduces both extreme years as well as average yields with high accuracy at a lead time of on average 6 weeks.

Weather influences explain a substantial share of observed maize yield variability in Tanzania as shown by the national median NSE of 0.92 in the estimation and 0.81 in the level 1 validation of the yield model based on weather inputs. Precipitation as well as higher minimum temperatures positively influence maize yields. The variable *consecutive dry days* is the most often selected variable across the regions, which underlines the detrimental effect of dry weather conditions on maize yields. This is in line with Cairns et al. (2013)<sup>6</sup> and Rowhani et al. (2011)<sup>7</sup> who found drought stress to be one of the main weather-related reasons for low maize yields in Sub-Saharan Africa. Apart from weather inputs, the SST of the West Pacific (WP) provides most explained yield variability in the forecast. This can be explained by a strong correlation between the SST in the WP and East African rain<sup>8</sup>. Low SST in the WP, which are related with higher rainfall amounts during the long rains in East Africa<sup>8–10</sup>, are positively correlated with maize yields. The variable selection shows that maize yield variability in Tanzania is strongly influenced by extreme events. Also the study of Rowhani et al. (2011)<sup>7</sup> focusing on climate variability and crop production in Tanzania seems to be in line with our finding. They found a strong relationship between the coefficient of variation (cv) of weather variables and cereal yields. In contrast to the cv, the percentile variables can explicitly represent lower or upper extremes and are therefore easier to interpret. The importance of extreme events for explaining maize yield variability in Tanzania implies that there is an upper quality bound for yield forecasts in case of extreme events that are out of the considered range of the training data. The variables selected in the estimation and the forecast show strong similarities for most regions, which suggests that the variable selection is robust.

In contrast to the high estimation and level 1 validation performance, the level 2 validation, which additionally integrates an out-of-sample variable selection, only shows good skills for a limited number of regions in Tanzania. This rigorous level 2 validation is most relevant in practice, as at the moment of forecasting (1–2 month before the harvest), the variable selection can only be done based on yield information from previous years. Despite its importance, this validation is rarely applied by forecasting studies and more than half of the studies even lack a level 1 validation<sup>11</sup>. The study of Ogotu et al. (2018)<sup>3</sup> who provide a forecast for East Africa does not even assess the model results based on observational data. To guarantee that the weaker level 2 performance is not due to model overfitting for the short time series at our disposal, we used a precautionary approach of only allowing a maximum of 5 variables to be selected for a time series of 10 years. This leaves sufficient degrees of freedom (5 DF in the full model and 4 DF for the validation) so that we assume model overfitting to the training as not the main reason for lower performance. Instead, the weaker level 2 performance could be related to the following reasons. First, the short length of the time series of only 10 years may not be sufficient to provide a stable out-of-sample variable selection. In our study period, several ENSO events (e.g. El Niño in 2015/2016 or 2009/2010 and La Nina in 2010/2011 or 2011/2012<sup>12</sup>) and Indian Ocean Dipole phases (e.g. positive phases in 2019 and 2015 and negative phases in 2016 and 2010<sup>13</sup>) occurred, which bring high year-to-year variation in weather



**Figure 8.** Regional performance of the forecasts derived from the model combining SST and weather inputs. Model assessment was done separately for the trend (left panel), the variability (middle panel) and absolute yields, which is a combination of trend and variability (right panel) for 8 of 21 regions in Tanzania. Plots for all 21 regions can be found in SI Fig. 5. The dark colour (dark green and dark purple) shows the forecast when the level 1 validation is applied. The corresponding NSE value is shown as ‘nse.l1’. The light colour (light green and light purple) and the ‘nse.l2’ show the results of the level 2 validation. Because the trend was fitted based on logarithmic values, the transformation back to linear values results in a slightly curved shape.



**Figure 9.** Assessments of the forecasted absolute yields for 2019 in regions with unimodal and bimodal rainfall regimes.

influences and consequently in the variable selection. The time series may not be long enough to train the model for this highly variable weather. The forecast should be repeated and tested on a longer time series when more data becomes available. Also the prediction of ENSO Events<sup>14,15</sup> can be used to inform the yield forecast. Second, uncertainties in the calculated start and end of the growing season, as well as the calculated separation between the vegetative and the reproductive phase can lead to a lower performance of the forecast or inaccuracies in the lead time. The calculated dates may not reflect the real phenological phases and the static dates do not account for the inner-seasonal variability due to differing weather conditions. Third, the forecast relies on reported yield data, which is prone to reporting errors and data quality issues, which may change the forecasting performance.

Because of its relevance in practice, we also provide a forecast of absolute yields, which in addition to the yield variability also includes information about the trend. The results of the absolute yield forecasts are slightly better (median NSE of 0.79 in the level 1 validation) due to the explanatory power of the trend. The disentangling of trend and variability provides transparency about the reasons for the forecasting skills so that in case of a weak weather signal, the trend can be used as the best available information about the expected yields. The forecast is only valid when assuming that the trend of the previous years is going to continue for the following years. For grounding such an assumption, a causal analysis of trends taking into account national policies<sup>16</sup> and international market conditions, as it has been examined for East African countries<sup>17</sup> and the Usungu plain in Tanzania<sup>18</sup>, would be mandatory.

The completely independent forecast of yields for the harvest year 2019 shows a high performance for the unimodal rainfall areas and a poor performance in bimodal rainfall regions. The poor performance in the bimodal regions could be related to the outbreak of the fall armyworms, which particularly hit the bimodal rainfall areas in northern, north-eastern and the coastal areas. The below average rainfall during the long rains in these areas fostered the outbreak of the fall armyworms so that an infestation level of 50% was reached in some regions<sup>19</sup>. This result underlines the importance of incorporating local expert knowledge in the evaluation of an operational forecast. Yields in some regions may be strongly influenced by factors that the model does not account for, such as pests and diseases or political, economic or social changes. Therefore, the forecast—as a tool that provides quantitative information about the expected weather-related yields—should be embedded in a forecasting system that integrates several sources of information for the evaluation of the food security situation.

In this study, we constructed an empirical model to decipher climatic influences on maize yield variation in Tanzania. We applied this model to forecast maize yield anomalies and absolute yields at a lead time of about 6 weeks before the harvest on region level in Tanzania. The model provides accurate and robust yield forecasts for large parts of the country. The strict and transparent assessment of the forecasting skill and the low requirements of input data make the forecast potentially suitable to inform operational yield forecasting systems in other countries with inaccessible or low quality weather data and short time series yield data. The proposed model can contribute to better informed local agronomic management strategies and support the implementation of regional agricultural development programs so that food shortages caused by unfavourable weather conditions can be mitigated. This has the potential to accelerate investments in these programs and thus, reduce food insecurity.

## Methods

**Yield data.** We used official maize yield data from 2009 to 2019 obtained from the National Food Security Division of the Ministry of Agriculture of Tanzania. Maize is the most important staple crop in Tanzania<sup>4</sup> and had a total harvest area of 3,428,630 ha in 2019, which represents ca. 25% of the total crop land in Tanzania<sup>20</sup> (statistics from 2017). Maize yields in Tanzania show a high variability and were on average at 1.6 t/ha from 2016 to 2018 (SI Fig. 6). The official statistics report yearly yield data covering all 21 regions of Mainland Tanzania (excluding the island Zanzibar). The administrative boundaries of some regions changed in 2012. To obtain a consistent time series, we used the administrative boundaries from before 2012 by applying a weighted average based on area.

**Weather data.** We created weather variables based on two climate data sources. For precipitation, we used observed daily rainfall totals from CHIRPS (Climate Hazards group Infrared Precipitation with Stations) at a



resolution of  $0.25^\circ \times 0.25^\circ$ <sup>21</sup>. CHIRPS provides reliable precipitation information for East Africa and outperforms previous products, like ARC2 and TAMSAT<sup>22,23</sup>. For daily mean, maximum and minimum temperature, we used ERA5 data<sup>24</sup> provided at a spatial resolution of  $0.25^\circ \times 0.25^\circ$ . ERA5 is the most recent reanalyses product and outperforms ERA-Interim for East Africa<sup>25</sup>. The weather variables were created for the time period of available yield statistics (2009–2019), including weather data for the year 2008 in case of growing seasons starting already at the end of 2008.

**SST data.** In addition to weather data, we included monthly sea surface temperature (SST) anomalies<sup>26</sup> of (1) the El Niño 3.4 zone ( $170^\circ\text{W}$ – $120^\circ\text{W}$ ,  $5^\circ\text{S}$ – $5^\circ\text{N}$ ), (2) the West Pacific Box (WP) ( $130^\circ\text{E}$ – $160^\circ\text{E}$ ,  $10^\circ\text{S}$ – $10^\circ\text{N}$ ) and (3) the Indian Ocean Dipole (IOD), which is the non-normalized difference between the West Indian Ocean ( $50^\circ\text{E}$ – $70^\circ\text{E}$ ,  $10^\circ\text{S}$ – $10^\circ\text{N}$ ) and the Eastern Indian Ocean ( $90^\circ\text{E}$ – $110^\circ\text{E}$ ,  $10^\circ\text{S}$ – $0^\circ\text{N}$ ). We included these SST indicators due to their documented influence on East African rainfall. SST anomalies in the El Niño 3.4 zone and the IOD have shown to be positively correlated with rainfall over East Africa during the short rains, whereas SST anomalies in the WP show a negative correlation with East African rainfall during the long rains<sup>8–10</sup>. The influence of these indicators on East African precipitation has a lag of some weeks to months offering the potential to provide yield forecasts at longer lead times. We tested different lead times and included those that showed the highest correlation with yield (SI Fig. 7), i.e. a lead time of 120 days for El Niño 3.4 and WP; and a lead time of 30 days for IOD.

**Growing season.** The North, North-East and the coastal areas of Tanzania are characterized by a bimodal rainfall pattern. The so-called ‘short rains’ (or Vuli) occur from October to December and the ‘long rains’ (Masika) last from March to May. The rest of the country has a unimodal rainfall pattern called Musumi with rainfall occurring from December to April<sup>9</sup>. Because of the low availability of irrigation facilities in Tanzania<sup>20</sup>, we did not distinguish irrigated from non-irrigated agriculture and considered the growing seasons to be aligned with the onset of the rain.

We defined the start of the growing season following the approach of Dodd and Jolliffe (2001)<sup>27</sup>, which was tested for tropical and subtropical conditions. They define the onset of the growing season when the following three criteria are fulfilled:

1. at least 25 mm rainfall within 6 days
2. starting day and at least 2 other days in this 6-day period are wet ( $>0.1$  mm)
3. no dry period of 10 or more consecutive days within next 40 days

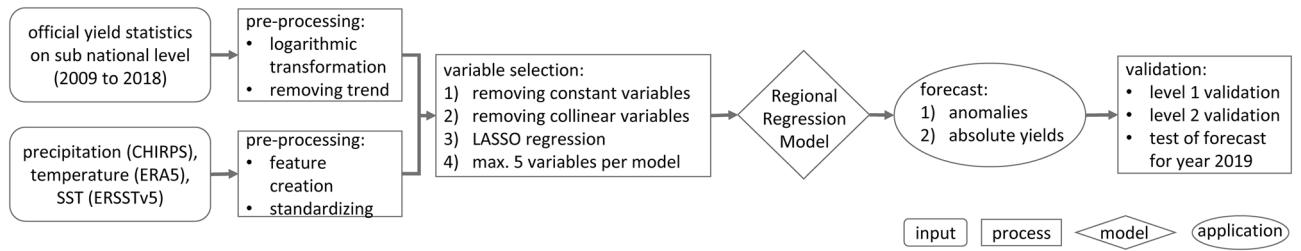
Because of the bimodal rainfall pattern in North and North-East Tanzania, two onsets of the growing season are found for some grid points. In this case, we considered the onset of the long rains (Masika), which is the main growing season. We used a static crop calendar at region level by first calculating the onset and the end of the growing season per grid point for each harvest year and then calculating the median over all grid points within a region and then a median over all years. We also tested other approaches (SI Section 8) to test the robustness of the results.

We defined the end of the growing season as 110 days after the start, which corresponds to the average time from sowing to maturity of 4 typical maize cultivars in Tanzania (i.e. Stuka, Staha, TMV1 and Pioneer HB3253). This is a reasonable choice since individual growing season lengths vary in a small range from 105 to 114 days<sup>28</sup>.

**Model inputs.** Based on temperature and precipitation data, we created variables that account for the climate drivers that influence maize development and yields (SI Tables 1, 2, 3 show a list of model inputs).

In addition to the median daily mean, maximum and minimum temperature over the growing season (*tas.median*, *tas.max*, *tas.min*), we included variables related to extreme temperatures. Above the optimum temperature range, photosynthesis is reduced, whereas respiration rates rise, such that net photosynthesis rates decline<sup>29</sup>. Particularly during flowering, maize is sensitive to heat stress, as it can lead to the desiccation of pollen or a reduction in grain numbers. High temperatures accelerate the development rate and result in a shortened growing season and thus a reduction in light perception. Particularly if the time for grain filling is reduced, grain size and consequently yields decline<sup>30</sup>. To account for the damaging impacts of too high temperatures, we included the number of days where the daily maximum temperature exceeds the region-specific long term 99% percentile of the maximum temperature in the growing season (*tas.max99*; SI Section 9 Eq. 1). Also temperatures below the optimum temperature range, i.e. any temperature below  $25^\circ\text{C}$  according to Rötter and Van De Geijn (1999)<sup>31</sup>, are detrimental for maize growth. Therefore, we include the number of times the daily minimum temperature falls below the region-specific long-term 1% percentile of the minimum temperature (*tas.min01*).

The optimal rainfall required in the growing season for maize is around 500–800 mm<sup>32</sup>. We included the precipitation sum (*Psum*) in the growing season to represent the overall water availability for maize. For optimal plant development, the timing and duration of water supply are equally critical. During flowering, maize requires sufficient water supply. Before pollination water stress leads to kernel abortion, even if at the time of pollination sufficient water is available. Excessive rain, in contrast, can lead to soil water saturation and oxygen deficiency, which limits root respiration and the growth of roots. Rainfall that exceeds the water holding capacity of the soil can lead to the leaching of nutrients and nutrient deficiencies of the plant<sup>33</sup>. To represent different precipitation ranges, we included the number of days with precipitation above a threshold of 5, 10 and 15 mm (*pA5*, *pA10*, *pA15*, respectively) and below a threshold of 5, 10 and 15 mm (*pB5*, *pB10*, *pB15*, respectively). Because the distribution of rainfall within the growing season is of particular importance for plant growth, we included



**Figure 10.** Modelling flow chart.

consecutive dry days of more than 5, 10, 15 and 20 days (*cdd5*, *cdd10*, *cdd15*, *cdd20*, respectively). Extremely high precipitation events are covered by the number of times the daily precipitation sum exceeds the region-specific long-term 99% percentile of the daily precipitation sum.

The variables were separately calculated for the vegetative and reproductive phase of the growing season. The separation between both phases was based on the sum of growing degree days (GDD; SI Section 9 Eq. 2). The days in the growing season until 50% of the full-season GDD sum was reached were allocated to the vegetative phase and the remainder to the reproductive phase, following Schauburger et al. (2017)<sup>34</sup>.

The SST temperatures were aggregated by taking the median over the growing season. We also included the number of times the 99% percentile of the whole time series is exceeded and the number of times the indicator is below the 1% percentile as an additional variable. The SST indices are only created for the vegetative phase (including the lead time) based on the same approach as for the weather variables.

**Model development and assessment.** For each of the 21 regions, we applied the work flow as shown in Fig. 10, i.e. we provide the forecast separately for each region comprising of region-specific variables and parameters to account for the diverse climatic conditions within the country.

For our analysis, we used R<sup>35</sup> (version 3.5.0) with the packages *tidyr*<sup>36</sup> and *plyr*<sup>37</sup> for data pre-processing, the packages *sp*<sup>38</sup> and *rgdal*<sup>39</sup> for spatial data processing, the package *glmnet*<sup>40</sup> to perform LASSO regression and the package *ggplot2*<sup>41</sup> to generate the figures and maps.

**Pre-processing.** We standardized all weather and SST input data to allow for a better comparability of the beta coefficients<sup>42</sup>. We transformed the yield input data to logarithmic values and removed the trend by first testing different de-trending methods (mean, linear, quadratic) and then applying the one that resulted in the lowest Akaike Information Criterion (AIC).

**Variable selection.** We applied the following variable selection process to elucidate important variables for explaining yield variability in the different regions:

- We removed variables that do not show year-to-year variations (i.e. zero variance).
- To avoid multicollinearity, only those variables were selected that are not strongly collinear (i.e. Pearson's  $r > 0.7$ ) with another explanatory variable. If a pair of variables was strongly collinear, then the variable with the higher correlation with yield anomalies was selected.
- Input selection was done using LASSO regression. Through regularization, LASSO performs a co-variate selection, which improves both the prediction accuracy and the interpretability<sup>43</sup>. To select the optimum lambda (the regularization penalty for the LASSO regression), we used the lowest cross-validation (years were omitted subsequently) mean squared error (MSE).
- At last, we removed all variables except for those 5 variables showing the highest correlation with yields. This step was conducted to reduce overfitting of the model based on the rule-of-thumb to include a maximum of half as many independent variables (climate variables) as there are dependent variables (yields).

**Model for estimating and forecasting yields before harvest.** For each region, we applied a separate regional regression model (Eq. 1) following the approach of Gornott and Wechsung (2016)<sup>44</sup> and Schauburger et al. (2017)<sup>34</sup>.

$$\log(y_{it}) = \sum_{k=1}^K \beta_{ki} x_{kit} + \varepsilon_{it} \quad (1)$$

with  $\beta$  as parameters,  $y$  as the demeaned and detrended response variable,  $x$  as the standardized explanatory input variable,  $\varepsilon$  as error term for  $K$  variables ( $k = 1, \dots, K$ ),  $N$  spatial units ( $i = 1, \dots, N$ ) and  $T$  years ( $t = 1, \dots, T$ ).

With the Breusch–Godfrey and the Breusch–Pagan tests we tested autocorrelation and heteroscedasticity of model residuals. We use the Variance Inflation Factor to test for multicollinearity.

For the within-season forecast, we only included variables during the vegetative phase. This provides a lead time of around 55 days, i.e. around 6 weeks before the harvest in a growing season totalling 110 days. The exact lead time differs per region—depending on the start of the reproductive phase in each region (SI Table 5). We provided the forecast for yield anomalies (i.e. variation around a trend) and for absolute values. For the latter, we added the trend and retransformed the values back to the linear form.

**Cross validation.** We applied a two level leave-one-out cross validation (LOOCV).

Level 1—LOOCV: we selected the variables based on all observations with LASSO. Then we subsequently removed observations for 1 year from the data set and used the remaining observations from the other years to fit the model and predict yield changes for the removed year, using the pre-selected set of variables.

Level 2—LOOCV: we subsequently removed observations for 1 year from the data set, then selected variables with LASSO and estimated the model based on the remaining data set. We used this model to predict yield changes for the removed year. This level 2—LOOCV guarantees that no information from the removed year is used for the variable selection or the model estimation. For this validation we allow a maximum of only 4 variables per model (allowing one less than half as many variables as there are observations). This validation simulates the operational context, where no yield information from the year to be forecasted is available for model building.

In addition, we used the model we developed based on the yield data from 2009 to 2018 to provide a completely independent forecast for the harvest year 2019. The yield data for 2019 was not known to us during the model development.

The goodness of fit between the observed data and the predicted data was then evaluated based on the Nash–Sutcliffe efficiency coefficient<sup>5</sup> (NSE). In contrast to the explained variance ( $r^2$ ), NSE does not only evaluate similarities in variability, but also integrates the mean model bias, which makes it a robust measure of the prediction quality. We also calculated the root mean squared error (RMSE) to estimate the average error between observed and predicted yields.

## Data availability

All data supporting the findings of this study are either public data sets, are available within the article and its Supplementary Information files or are available from the corresponding author upon reasonable request.

Received: 26 June 2020; Accepted: 27 October 2020

Published online: 12 November 2020

## References

- Liu, L. & Basso, B. *Seasonal Crop Yield Forecast: Methods, Applications, and Accuracies. Advances in Agronomy* Vol. 154 (Elsevier Inc., Amsterdam, 2019).
- Liu, L. & Basso, B. Linking field survey with crop modeling to forecast maize yield in smallholder farmers' fields in Tanzania. *Food Secur.* **12**, 537–548 (2020).
- Ogutu, G. E. O., Franssen, W. H. P., Supit, I., Omondi, P. & Hutjes, R. W. A. Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. *Agric. For. Meteorol.* **250–251**, 243–261 (2018).
- National Bureau of Statistics. Basic Demographic and Socio-Economic Profile (2014) [https://tanzania.countrystat.org/fileadmin/user\\_upload/countrystat\\_fenix/congo/docs/2012\\_Tanzania\\_Population\\_and\\_Housing\\_Census-Basic\\_Demographic\\_and\\_Socio-Economic\\_Profile.pdf](https://tanzania.countrystat.org/fileadmin/user_upload/countrystat_fenix/congo/docs/2012_Tanzania_Population_and_Housing_Census-Basic_Demographic_and_Socio-Economic_Profile.pdf) (Accessed 11th July 2020).
- Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I—a discussion of principles. *J. Hydrol.* **10**, 282–290 (1970).
- Cairns, J. E. *et al.* Adapting maize production to climate change in sub-Saharan Africa. *Food Secur.* **5**, 345–360 (2013).
- Rowhani, P., Lobell, D. B., Linderman, M. & Ramankutty, N. Climate variability and crop production in Tanzania. *Agric. For. Meteorol.* **151**, 449–460 (2011).
- Davenport, F., Funk, C. & Galu, G. How will East African maize yields respond to climate change and can agricultural development mitigate this response?. *Clim. Change* **147**, 491–506 (2018).
- Hoell, A. & Funk, C. Indo-Pacific sea surface temperature influences on failed consecutive rainy seasons over eastern Africa. *Clim. Dyn.* **43**, 1645–1660 (2014).
- Funk, C. *et al.* Predicting East African spring droughts using Pacific and Indian Ocean sea surface temperature indices. *Hydrol. Earth Syst. Sci.* **18**, 4965–4978 (2014).
- Schaubert, B., Jägermeyr, J. & Gornott, C. A systematic review of local to regional yield forecasting approaches and frequently used data resources. *Eur. J. Agron.* **120**, 126153 (2020).
- NOAA. Cold and Warm Episodes by Season (2020) [https://origin.cpc.ncep.noaa.gov/products/analysis\\_monitoring/ensostuff/ONI\\_v5.php](https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php) (Accessed 24th April 2020).
- NOAA/ESRL. Dipole Mode Index (DMI) (2020) <https://stateoftheocean.osmc.noaa.gov/sur/ind/dmi.php> (Accessed 24 April 2020).
- Ludescher, J. *et al.* Very early warning of next El Niño. *Proc. Natl. Acad. Sci.* **111**, 2064–2066 (2014).
- Meng, J. *et al.* Complexity-based approach for El Niño magnitude forecasting before the spring predictability barrier. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 177–183 (2020).
- FAO. Review of Food and Agricultural Policies in the United Republic of Tanzania 2005–2011—Country Report (2013) <https://www.fao.org/3/a-at476e.pdf> (Accessed 5 June 2020).
- Aylward, C. *et al.* Maize Yield Trends and Agricultural Policy in East Africa. *Evans School of Public Policy and Governance* (2015) [https://evans.uw.edu/sites/default/files/EPAR\\_UW\\_310\\_National-Level\\_MaizeYieldTrends\\_5.31.16.pdf](https://evans.uw.edu/sites/default/files/EPAR_UW_310_National-Level_MaizeYieldTrends_5.31.16.pdf) (Accessed 4 June 2020).
- Malley, Z. J. U., Taeb, M. & Matsumoto, T. Agricultural productivity and environmental insecurity in the Usangu plain, Tanzania: Policy implications for sustainability of agriculture. *Environ. Dev. Sustain.* **11**, 175–195 (2009).
- FAO. GIEWS Country Brief United Republic of Tanzania (2019) <https://www.fao.org/giews/countrybrief/country/TZA/pdf/TZA.pdf> (Accessed 4 May 2020).
- FAO. FAOSTAT—Food and agriculture data (2019) <https://www.fao.org/faostat/> (Accessed 10 April 2019).
- Funk, C. *et al.* The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* **2**, 1–21 (2015).
- Muthoni, F. K. *et al.* Long-term spatial-temporal trends and variability of rainfall over Eastern and Southern Africa. *Theor. Appl. Climatol.* **137**, 1869–1882 (2019).
- Dinku, T. *et al.* Validation of the CHIRPS satellite rainfall estimates over eastern Africa. *Q. J. R. Meteorol. Soc.* **144**, 292–312 (2018).
- Copernicus Climate Change Service (C3S). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate (2017) <https://cds.climate.copernicus.eu/cdsapp#!home> (Accessed 19 December 2019).
- Gleixner, S., Demissie, T. & Diro, G. T. Did ERA5 Improve Temperature and Precipitation Reanalysis over East Africa? *Atmosphere*. **11**, 1–19 (2020).
- Huang, B. *et al.* Extended reconstructed Sea surface temperature, Version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).

27. Dodd, D. E. S. & Jolliffe, I. T. Early detection of the start of the wet season in semiarid tropical climates of Western Africa. *Int. J. Climatol.* **21**, 1251–1262 (2001).
28. Mourice, S. K., Rweyemamu, C. L., Tumbo, S. D. & Amuri, N. Maize cultivar specific parameters for Decision Support System for Agrotechnology Transfer (DSSAT) application in Tanzania. *Am. J. Plant Sci.* **05**, 821–833 (2014).
29. Barnabás, B., Jäger, K. & Fehér, A. The effect of drought and heat stress on reproductive processes in cereals. *Plant Cell Environ.* **31**, 11–38 (2008).
30. Carter, P. R. & Hestermann, O. B. Handling corn damaged by autumn frost. *National Corn Handbook—Climate and Weather* (1990) <https://www.baycounty-mi.gov/Docs/CitizenCorps/HandlingCornDamageByAutumnFrost.pdf> (Accessed 1 May 2019).
31. Rötter, R. & Van De Geijn, S. C. Climate change effects on plant growth, crop yield and livestock. *Clim. Change* **43**, 651–681 (1999).
32. Critchley, W. & Siegert, K. Water harvesting (1991) <https://www.fao.org/3/U3160E/u3160e04.htm> (Accessed 1 May 2019).
33. Rötter, R. P. *et al.* Linking modelling and experimentation to better capture crop impacts of agroclimatic extremes—a review. *Food Crop. Res.* **221**, 142–156 (2018).
34. Schaubberger, B., Gornott, C. & Wechsung, F. Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting. *Glob. Change Biol.* **23**, 4750–4764 (2017).
35. R Core Team. R: a language and environment for statistical computing (2018) <https://www.r-project.org/>.
36. Wickham, H. & Henry, L. tidy: Easily Tidy Data with 'spread()' and 'gather()' functions (2019) <https://cran.r-project.org/package=tidy>.
37. Wickham, H. The split-apply-combine strategy for data analysis. *J. Stat. Softw.* <https://doi.org/10.18637/jss.v040.i01> (2011).
38. Pebesma, E. J. & Bivand, R. S. Classes and methods for spatial data in R (2005) <https://cran.r-project.org/doc/Rnews/>.
39. Bivand, R., Keitt, T. & Rowlingson, B. rgdal: bindings for the 'Geospatial' Data Abstraction Library. R package version 1.3-2 (2018) <https://cran.r-project.org/package=rgdal>.
40. Friedman, J. *et al.* Regularization paths for generalized linear models via coordinate descent (2008) <https://web.stanford.edu/~hastie/Papers/glmnet.pdf>.
41. Wickham, H. ggplot2: elegant graphics for data analysis (2009) <https://ggplot2-book.org/>.
42. Wooldridge, J. M. *Introductory Econometrics A Modern Approach*. (Cengage Learning, 2014).
43. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
44. Gornott, C. & Wechsung, F. Statistical regression models for assessing climate impacts on crop yields: a validation study for winter wheat and silage maize in Germany. *Agric. For. Meteorol.* **217**, 89–100 (2016).

## Acknowledgements

This work has been funded by the EPICC and the ClimSec project. EPICC is part of the International Climate Initiative (IKI). The Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) supports this initiative on the basis of a decision adopted by the German Bundestag. ClimSec is financed by the German Federal Foreign Office. The work of David Makowski was partly funded by the project CLAND (16-CONV-0003).

We thank Stephanie Gleixner for her contributions to the climate data pre-processing.

## Author contributions

R.L., C.G., B.S. and D.M. designed the research. R.L. performed the research and wrote the manuscript with contributions from C.G. and B.S. C.G. initiated the research. D.M. reviewed the statistical modelling approach and contributed to the interpretation of the results. All authors revised the manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-020-76315-8>.

**Correspondence** and requests for materials should be addressed to R.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/>.

© The Author(s) 2020