Originally published as:

Schauberger, B., Gornott, C., Wechsung, F. (2017): Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting. - Global Change Biology, 23, 11, 4750-4764

DOI: 10.1111/gcb.13738

Global evaluation of a semi-empirical model for yield anomalies and application to within-season yield forecasting

(Semi-empirical modeling of yield anomalies)

Authors: Bernhard Schauberger*[1,2], Christoph Gornott[1], Frank Wechsung[1]

* Corresponding author: schauber@pik-potsdam.de, +33 1 69 08 77 24

[1] Potsdam Institute for Climate Impact Research (PIK), Telegrafenberg A31, 14473 Potsdam, Germany

[2] Laboratoire des Sciences du Climat et de l'Environnement, Institut Pierre-Simon Laplace (IPSL), 91191 Gif sur Yvette, France

Keywords: yield anomaly, maize, wheat, soybeans, global, weather, semi-empirical model, forecast

Type: Primary Research Article

ABSTRACT

Quantifying the influence of weather on yield variability is decisive for agricultural management under current and future climate anomalies. We extended an existing semi-empirical modeling scheme that allows for such quantification. Yield anomalies, measured as inter-annual differences, were modeled for maize, soybeans and wheat in the US and 32 other main producer countries. We used two yield data sets, one derived from reported yields and the other from a global yield data set deduced from remote sensing. We assessed the capacity of the model to forecast yields within the growing season.

In the US, our model can explain at least two thirds (63-81%) of observed yield anomalies. Its out-of-sample performance (34-55%) suggests a robust yield projection capacity when applied to unknown weather. Out-of-sample performance is lower when using remote-sensing derived yield data. The share of weather-driven yield fluctuation varies spatially, and estimated coefficients agree with expectations. Globally, the explained variance in yield anomalies based on the remote-sensing data set is similar to the US (71-84%). But the out-of-sample performance is lower (15-42%). The performance discrepancy is likely due to shortcomings of the remote-sensing yield data since it diminishes when using reported yield anomalies instead. Our model allows for robust forecasting of yields up to two months before harvest for several main producer countries. An additional experiment suggests moderate yield losses under mean warming, assuming no major changes in temperature extremes.

We conclude that our model can detect weather influences on yield anomalies and project yields with unknown weather. It requires only monthly input data and has a low computational demand. Its within-season yield forecasting capacity provides a basis for practical applications like local adaptation planning. Our study underlines high-quality yield monitoring and statistics as critical prerequisites to guide adaptation under climate change.

INTRODUCTION

Strongly varying crop yields can endanger farmers' livelihoods and can lead to national production shortages. Yields are determined by weather and agronomic management influences as well as by stress factors like pests or diseases. For calculating crop yields under current or a changing climate it is important to quantify these influences. Therefore we devise a semi-empirical modeling scheme which allows for quantifying weather influences with high explained variance. We use two different yield data sets with different qualities, one based on reported yield data and the other on remote sensing combined with yield statistics. We show the ability of the model to predict yield anomalies up to two months before harvest.

Two approaches are widely used to simulate crop yields (Di Paola *et al.*, 2016, Jones *et al.*, 2016, Lobell & Burke, 2010). Process-based models simulate physiological processes like carbon assimilation to calculate yields. Statistical models correlate yields with yield-determining factors to elicit contributions of individual factors. Both approaches, and hybrids between them, can aid in understanding and forecasting weather-related yield variability (Liu *et al.*, 2016). Their application to conditions (e.g. climate) out of the training scope is a contested area, however (Lobell & Burke, 2010, Rötter *et al.*, 2011).

Here we extend an existing statistical framework for modeling inter-annual yield variability. The approach is "semi"-empirical as known physiological influences are reflected in the exogenous variables, following the naming of Rahmstorf (2007). The concept was introduced in Wechsung *et al.* (2008) and later successfully applied to German maize and winter wheat yields (Gornott & Wechsung, 2016). We extend the model by adding temperature-stress related variables, using more crops, applying it to 34 countries and providing two application cases: forecasting yield anomalies up to two months before harvest and gauging of yield losses under moderately increased temperatures.

We analyze four staple crops: maize, wheat (spring and winter separately) and soybeans, which cover approx. 34% of the global harvested area (Portmann *et al.*, 2010). We use reported crop yield data in seven countries and a global gridded yield data set that downscaled reported yield statistics utilizing satellite data (here used for 33 countries). Subnational yield data are needed for quantifying spatial differences of yield influences. Though these data are increasingly available, there are still data-scarce regions especially in developing countries. The global and publicly available data set supplied by Iizumi *et al.* (2013b) might serve as alternative. The dataset uses annual remote sensing information to downscale national and subnational yield statistics. The algorithms utilized therein to separate reflectance data spatially and temporally into crops or vegetation necessarily introduce uncertainty, which increases with the share of other vegetation types in grid cells. Despite these caveats we test the potential of this global gridded data set for quantifying yield anomalies, as it may be helpful when subnational yield data are not accessible.

We apply a two-step procedure: the model performance is first analyzed in depth in the US and then, second, extended to all main producing nations. We start with US yields, since the high-quality yield data base curated by the US Department of Agriculture (USDA, 2015) allows for rigorous model evaluation. The model is applied in parallel to the USDA and the Iizumi *et al.* (2013b) data. The US are one of the largest crop producers (FAO, 2016) and have highly diverse climate and soils. We employ one model specification based on selection results by Gornott and Wechsung (2016), but test its sensitivity regarding variations in yield-influencing factors and transformation of variables. Additionally, we include penalty terms for heat and frost.

Instead of absolute yields we consider yield anomalies to remove trends, systematic biases and time-invariant farm- or county-specific influencing factors. Normalizing anomalies of yield and exogenous variables by the logarithm allows a comparison of influences across scales and variables. Only weather variables are included in the model, explicitly neglecting agronomic influences like acreage, shifting land use or fertilizer application on inter-annual yield fluctuation (Mueller *et al.*, 2012, Ray *et al.*, 2015). But these data do not increase model performance in Germany (Conradt *et al.*, 2016) and are difficult to obtain as time series on a spatially explicit level with large spatial coverage; they would therefore enlarge uncertainty. We only use monthly weather values which are deemed to provide more reliable information than daily weather data from models due to aggregation effects (Kilsby *et al.*, 2007, Lobell, 2013, Maurer *et al.*, 2010). This also avoids the use of downscaling methods when using climate model outputs (Glotter *et al.*, 2014, Iizumi *et al.*, 2012).

MATERIALS AND METHODS

Input data

Yield data

We employed two sets of yield data for maize, soybeans, spring and winter wheat (all in t/ha). For the US we used either USDA (USDA, 2015) yields at county level, from 1980 to 2010, or gridded yield data from Iizumi *et al.* (2013b) from 1982 to 2006 (henceforth "GGYD" for "Gridded Global Yield Data"). Both were re-gridded to 0.5° spatial resolution (about 50 km at the equator) to match with the resolution of the weather and land-use data. USDA county yields were assigned to each 0.5° grid cell that completely fall within a county or intersect with its boundaries; yields for grid cells intersecting with several counties were averaged. GGYD yields are provided at 1.125° resolution and were interpolated to 0.5° with second order conservative remapping (preserving fluxes and spatial gradients). Additional county-level yields for Germany, Russia, Tanzania, Australia, Brazil and Burkina Faso (from the respective statistical offices) allowed for further model and yield data quality assessments. National yield time series from FAO (FAO, 2016) were used for comparison of aggregated yield time series. We considered those countries as main producers (Figure 1, SI Table S3) which, sorted by total production, together accounted for more than 90% of world production for a specific crop between 2000 and 2011 (FAO, 2016).

Weather data

We used AgMERRA climate data (Ruane *et al.*, 2015) at 0.5° spatial and monthly temporal resolution, providing minimum, maximum and average temperature, precipitation and shortwave radiation from 1980 to 2010. AgMERRA has been designed for use in agricultural research focusing on reproducing both average and extreme values.

Growing season and land-use data

We utilized static MIRCA2000 crop- and irrigation-specific land-use fractions around 2000 on 0.5° spatial resolution (Portmann *et al.*, 2010). Growing seasons were also taken from MIRCA2000, using the sub-crop with the largest harvested area. Winter and spring wheat were distinguished by their growing season length: eight or more months were classified as winter wheat, four months or less as spring wheat. Remaining ambiguities were resolved by considering the sub-crop with the maximum (minimum) growing season length as winter (spring) wheat. Soybeans have a prolonged flowering period (Ritchie *et al.*, 1993) at the transition between vegetative and reproductive season. Although it could be physiologically reasonable, we restrained from reflecting this period in a separate set of exogenous variables to avoid collinearities and rank deficiencies (many variables for few data).

Regression scheme

Definition

We applied an ordinary least squares (OLS) regression scheme based on the Cobb-Douglas production function with different model specifications. The function relates inter-annual changes of crop yields to a product of inter-annual changes of weather variables (equation 1; SI equation SE3). The natural logarithm linearizes all terms into a sum.

$$\log y_t{}' = \log \beta_0 + \sum_{j=1}^{J} \beta_j \log x'_{jt} + \log u'_t \text{ , with } j = 1, \dots, J \text{ and } t = 1, \dots, M \qquad \text{(eq. 1)}$$

Variables are yield (*y*), weather (x_j) and error term (*u*). Estimated coefficients are $\beta_{0..J}$ and denote intercept (β_0) and weather influences. All variables are provided per grid cell. Years are indexed with *t*. Anomalies are denoted with a prime (′). We calculated yield anomalies as first differences ($y'_t = y_t - y_{t-1}$) between adjacent years, making an explicit time variable obsolete. We used two regression methods: STSM (Separate Time Series Model) and PDM (Panel Data Model). While STSM estimates an independent model for each grid cell, the PDM parametrizes relationships across grid cells, allowing for spatial variation in mean yields with grid cell-specific fixed effects. These choices are justified by earlier results (Conradt *et al.*, 2016, Gornott & Wechsung, 2016) and the similarity of results under different techniques (SI Section 3). Whether spatial correlation poses a problem for the PDM method is tested (see below). In the US we considered nine climatic regions (SI Figures S1-2). Other, larger main producers were split into administrative boundaries for PDM estimation; for all others only one national PDM was estimated (SI Table S3).

Exogenous variables

Exogenous variables either describe potential growth or stress factors that reduce growth, included for their known physiological relevance. They are tested for statistical significance, but the model formulation stays constant. We therefore consider the model as “semi”-

empirical following the argumentation of Rahmstorf (2007). A combined temperature-radiation variable relates yields to potential growth. Temperature-normalized solar radiation (SRT, equation 2) is used to account for co-linearity in both variables. Killing (KDD) and freezing degree days (FDD) were added to better account for the non-linear influence of extreme temperatures on crop yields (Barlow *et al.*, 2015, Schlenker & Roberts, 2009). They are defined as the temperature sum above or below a crop-specific threshold, respectively (equations 3,4). The KDD threshold T^{KDD} was 32°C for all crops, while the FDD threshold T^{FDD} was -15°C for the two wheat types and 0°C for maize and soybeans (Hatfield *et al.*, 2011, Luo, 2011, Porter & Gawith, 1999, Sanchez *et al.*, 2014).

$$SRT = \frac{R_S}{T_{avg}+20} \qquad \text{(eq. 2)}$$

$$KDD = \sum_{d=1}^{N} \max(T_d - T^{KDD}; 0) \qquad \text{(eq. 3)}$$

$$FDD = \sum_{d=1}^{N} \min(T_d - T^{FDD}; 0) \qquad \text{(eq. 4)}$$

Further stress variables comprised potential evapotranspiration (PET) and precipitation. Both variables map the yield-reducing effect of inadequate demand and supply of water by PET and precipitation, respectively. PET was calculated from VPD according to Haude (1955) as in Gornott and Wechsung (2016) except that the month-specific correction factor f_H was considered constant for the sake of a simpler model. For winter wheat only the reproductive part of SRT was considered, while for the other crops only the vegetative part was used. The full regression specification is provided in SI section 2. Further agronomic justifications are provided in Gornott and Wechsung (2016). Economic variables like fertilizer price and harvested area were not considered since these only added little explanatory power in Germany (Conradt *et al.*, 2016) and are generally not available on larger areas across the world.

PET and precipitation were split between the vegetative and reproductive part of the growing season. The identification of both parts was based on phenological heat units. The first month of the reproductive period was defined as the first month where the temperature sum, accumulated over the growing season until this month, exceeds 50% of the total temperature sum, accumulated over the whole growing season (supplementary equations SE4,5).

Aggregation

After estimation yield anomaly time series (observed, predicted and one-out-of-sample predicted yield anomalies) were aggregated from grid cells to climate regions or countries (supplementary equations SE1,2). Aggregation was performed unweighted, i.e. treating each grid cell as equal, or weighted by land-use patterns according to MIRCA2000. Performance measures (see below) were then calculated for aggregated time series.

Model evaluation

Performance

Six performance indicators were calculated: coefficient of determination (R^2), root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE), one-out-of-sample R^2 (henceforth: R^2_{O1}), out-of-temperature R^2 (R^2_{OOT}) and out-of-precipitation R^2 (R^2_{OOP}). The first three are standard model evaluation indices and measure the explained variance, the mean deviation and a combined measure of model bias and variability, respectively. They indicate the capacity of the model to explain yield anomalies, which is important for interpreting coefficients. R^2_{O1} was calculated by subsequently and separately stripping each year from the estimation data, estimating the model with the reduced data and eventually predicting yield anomalies for the stripped year with this reduced model. R^2_{O1} thus indicates the model's capacity to project yields from weather data that have not been used for model training. R^2_{OOT} and R^2_{OOP} were similarly calculated by omitting the six first-differences towards and from the three warmest (driest) years, defined by highest growing season mean temperature (lowest precipitation over PET). Thus the model was trained on six yield anomalies less and was then used to predict these missing anomalies. The correlation between these predicted and observed anomalies in only the warmest (driest) years, calculated across aggregation regions, indicates the capacity to project yield anomalies under warmer (drier) climate. Performance measures were calculated on nationally aggregated time series, but are also available for each grid cell.

Statistical tests

The adequacy of the linear model for capturing yield anomalies was examined with six statistical tests. The regression equation specification error test (RESET) evaluated whether quadratic variables would improve the model. The Lagrange multiplier test according to Breusch–Pagan (LM) was used to examine spatial independence of the data. The Breusch–Godfrey test was applied to assess autocorrelation and the Breusch–Pagan test to probe heteroscedasticity (Croissant & Millo, 2008, Wooldridge, 2013). Normal distribution of residuals was tested using the Shapiro–Wilk test. Whether multi-collinearity of exogenous variables poses a problem was assessed with the condition index following Belsley *et al.* (1980). All analyses were performed with R (R Core Team, 2016).

Model application

Two practical applications of the model were performed.

Yield forecasting

The model was applied to forecast yield anomalies during the growing season up to two months before harvest. We clipped the last one or two months, respectively, from the MIRCA2000-defined growing season and calculated all weather variables based on this reduced season. Afterwards the model was trained on the reduced weather data set, relating yield anomalies to weather anomalies observed up to one or two months before harvest. The one-out-of-sample performance of this reduced model is then a measure for its forecasting capacity.

Yield effects from temperature warming

Effects of moderate warming were calculated as a model application case. Temperature in every *second* growing season of the AgMERRA climate was raised by 0.9 or 1.4 °C, corresponding to the difference between the 0.6 °C of warming already present in 1986-2005 (Schleussner *et al.*, 2016) and current climate change targets of 1.5 or 2 °C. Differences in warming over land and ocean (IPCC, 2013) were neglected. Precipitation and radiation were not modified since we assume stochastic changes with mean zero for this temperature range (IPCC, 2013). Differences in CO_2 concentrations would be relevant for absolute yields, but were not considered due to rather minor changes (plus ~30 or 60 ppm for 0.9 or 1.4 °C warming, respectively, compared to 1980-2010 average concentrations; IPCC (2013)). The CO_2 increase of ~60 ppm during the historical period is not relevant for this application when assuming a similar increase in the warmed period – first differences cancel the trend in both time series. Yield anomalies were predicted with coefficients estimated from unmodified climate and exogenous variables from the artificial climate data. Grid-cell yield time series were nationally aggregated without weighting. The first-difference approach allows interpreting yield changes between adjacent years as effects of temperature increases. Yield changes (unmodified to modified and modified to unmodified years, with inverted signs) were averaged and the logarithm removed. A temperature change of 0 °C was used for deriving normalization constants with which all other yield changes were multiplied. Uncertainty of predictions u was calculated by adding RMSE of the one-out-of-sample model ($RMSE_{O1}$) and variance of the temperature-modified yield time series (eq. 5):

$$u = \sqrt{(RMSE_{O1})^2 + Var(mod.time\ series)} \qquad \text{(eq. 5)}$$

RESULTS

Results for the contiguous US

The model had a substantial capacity for explaining and predicting yield anomalies. Yield anomaly time courses for USDA-based models are shown in Figure 2. Results for each of the eight crop-yield data set combinations are displayed in Table 1. All grid cells where the specific crop is grown are included. Either unweighted or weighted aggregation was used, decided on the higher R^2_{O1} for each crop individually. Time series for US regions are provided in SI Figure S11. A performance comparison of different model specifications is provided in SI Figure S6. All statistical tests indicated that the OLS model estimation is adequate (SI section 4).

The model achieved at least two thirds of explained variance (R^2) and a robust (i.e. at least 25%) one-out-of-sample performance (R^2_{O1}) for all four crops with USDA data. Extremely low yields, like those occurring during the US heat and drought wave in 1988 for maize and wheat, were captured by the model, though not in full magnitude. For the two wheat types, yield loss quantities over the whole time series were comparable between model and observations, and for winter wheat also between one-out-of-sample model and observations. The set of three years of most negative yield anomalies (bottom decile) was equal for observed and modeled time series in 7 out of 12 cases. The observed top decile was captured in 8 out of 12 cases. For the one-out-of-sample predicted yields the correspondence for the bottom decile was less accurate with only 3 out of 12 cases. The direction of change and the sign of modeled anomalies matched with the input data for all crops, with only few exceptions.

The model performed differently for different crops, judged by R^2_{O1}. The regression method, variable set or difference method influenced model performance (SI Figure S6). Unweighted aggregation was better for maize, soybeans (except GGYD soybeans where R^2_{O1} was low) and spring wheat, but disfavored for winter wheat. Model performance differed between the two yield data sets. Although R^2 values were similar or higher for GGYD yields, R^2_{O1} values with GGYD data (Table 1**Error! Reference source not found.**, SI Figure S6) were lower in three of four cases. Differences between R^2 and R^2_{O1} were thus higher for

GGYD yields. STSM models showed, on average over all crops and specifications, slightly higher R^2 and R^2_{O1} values than PDM models (SI Figure S6). R^2 and R^2_{O1} were correlated for USDA yields (*r* = 0.97, *p* = 0, *n = 24*), but not GGYD yields (*r* = 0.29, *p* = 0.17, *n = 24*). NSE and R^2 showed larger differences for GGYD than USDA yields. Thus the model's explanatory power was not an indicator for the model's projective power with GGYD yields. The out-of-temperature and out-of-precipitation performance (where six anomalies were omitted for training) was lower than the one-out-of-sample performance. All out-of-temperature values with USDA yields are, nevertheless, above 0.25, thus higher than expectable by chance (corresponding to *r* = 0.5). One-out-of-sample performance in the three warmest years is hardly different from modeled values. Out-of-precipitation values are above 0.25 only for wheat.

The explained variance varied spatially (Figure 3). There was a substantial fraction of grid cells where the model was able to capture yield variability to a large (green shades) or an intermediate extent (yellow shades). But there were also several regions where the model failed to capture variability (red shades). For all crops these were located in areas where yield variability was lower compared to other regions. In regions with substantial yield variation (coefficient of variation CV, defined as standard deviation over mean, is larger than 15%) the model achieved a higher R^2 more often (SI Figure S10; SI Table S2). There was a moderate fraction of grid cells (11-27%) that exhibited low yield variability and was not well explained by the model.

Model coefficients indicated crop-specific patterns of weather influence. The influence of coefficients depended on the crop, but was independent from the estimation method (Figure 4). All STSM coefficient means except two were significantly different from 0 (t-test at 95% confidence level). For all crops a high PET in the reproductive period was clearly negative. Precipitation was positive for summer crops during the vegetative period and for soybeans and winter wheat also during the reproductive period. For spring wheat and maize too much precipitation during the reproductive period was negative. Normalized solar radiation was negative for maize and soybeans (vegetative period), but strongly positive for spring and winter wheat. Any day above 32°C was damaging for all crops (not significant for winter wheat), whereby maize was most affected. Days below -15°C or 0°C, respectively, were damaging for all crops, but did not occur during the spring wheat growing season. There was a marked difference of coefficient values between the two yield data sets (USDA, GGYD).

This was the case for STSMs (SI Figure S7) and PDMs (SI Figure S8).

Coefficients varied between climate regions (Figure 5). A high PET during the vegetative season was positive for maize yield in the northern climate zones, but negative in the south. Vegetative PET was positive everywhere for soybeans. For spring wheat a high PET was negative everywhere except the northwest. For winter wheat a high PET during the reproductive season was positive only in the northeast, but negative elsewhere. The effect of precipitation did not show pronounced regional diversity: it was positive in most regions for all crops, with few exceptions. Elevated SRT during the vegetative period had a positive effect on maize yields in mid and western states, but not elsewhere. Enhanced SRT was negative for soybeans in all regions. For spring wheat, by contrast, higher SRT was positive everywhere except the northwest. For winter wheat more SRT had positive effects during the reproductive period in almost the whole US, with a positive gradient to the southeast. Days above 32°C were harmful everywhere for maize, spring and winter wheat (-2 to -4% yield loss for each day).

A mapping sensitivity test, where climate, land-use and growing seasons were interpolated from grid cells to counties rather than yields from counties to grid cells, showed similar or slightly higher R^2 (0.82, 0.74, 0.65 and 0.68 for maize, soybeans, spring and winter wheat, respectively) and R^2_{O1} values (0.61, 0.55, 0.34 and 0.30). We kept the mapping of yields to grid cells, though, to maintain a common framework for both yield data sets.

Results for global main producers

The model explains more than two thirds of yield variance in main producer countries. The robust out-of-sample performance in the US supported an extension of the evaluation to other main producers (SI Table S3; Figure 1). Only GGYD yields could be used as generally available source here. Nationally aggregated GGYD yield anomalies mostly corresponded well with FAO yield anomalies (SI Figure S12), motivating the usage of this data set. The performance (R^2 and R^2_{O1}) for all crops is displayed in Figure 6. The explained variance among main producers, weighted by total production, was 84%, 72%, 71% and 71% for maize, soybeans, spring and winter wheat, respectively. The weighted average one-out-of-sample performance was 42%, 22%, 33% and 15%. The cumulative production share (within the main producers) of nations which achieved an R^2_{O1} of at least 25% is 64%, 18%, 68% and 30% for maize, soybeans, spring and winter wheat, respectively. Analyses with PDM estimation led to similar, though slighty lower performances (SI Figure S14). Calculating aggregated model performance as average performance over all grid cells in a country, rather than by correlating previously aggregated yield time series, resulted in lower model performances: mean R^2 [R^2_{O1}] STSM values over countries were 0.47 [0.18], 0.44 [0.15], 0.48 [0.19] and 0.36 [0.10] for maize, soybeans, spring and winter wheat. This aggregation effect, as discussed in Gornott and Wechsung (2016) for Germany, was thus confirmed globally.

Yield time series for selected main producers can be found in the supplement (SI Figure S13). Mean performance was best for maize (highest R^2 and R^2_{O1}). While R^2 was similarly high for soybeans, the R^2_{O1} was rather low (22%). For winter and spring wheat the model achieved equal mean R^2, while mean R^2_{O1} was substantially higher for spring wheat. There was no obvious influence of harvested area, length of yield time series, share of rainfed agriculture, mean yield level or standard deviation on model performance. Countries where GGYD yields were constructed from subnational data (Table S1 in Iizumi *et al.* (2013b)) tended to have a larger R^2_{O1}, but not significantly. There are some notable discrepancies between R^2 and R^2_{O1}, especially for winter wheat: for example in India or Egypt an R^2 of 0.93 and 0.73, respectively, was accompanied by an R^2_{O1} of 0.04 and 0.03. In both cases, this discrepancy is due to extreme yield values captured by the model, but not the one-out-of-sample model (data not shown). If these extremes are removed, R^2_{O1} increases to 0.16 and 0.22, respectively. Differences between R^2 and R^2_{O1} are generally due to an out-of-sample time series which is less variable and captures fewer extreme values than the modeled time series.

Yield data quality influences the detection of weather influences. There was a marked difference in model performance when using either reported sub-national yield data or gridded yield data derived from remote sensing. R^2_{O1} values for USDA data were 55%, 45%, 34% and 35% for maize, soybeans, spring and winter wheat, respectively, while for GGYD data these were 59%, 18%, 32% and 26%, thus lower except for maize (Table 1**Error! Reference source not found.**). This difference was also visible for Germany, Russia, Burkina Faso, Tanzania and Brazil (SI Table S4).

The average explained variance over all main producing countries and crops was 41.8% with GGYD yields. This was slightly higher than the 32-39% which have been found by Ray *et al.* (2015) with reported data. For maize the average R^2 was 44% with our model, compared to 39% in Ray *et al.*, and for soybeans it was 42%, compared to approx. 35%. For wheat (average over spring and winter) it was 42% with our model, compared to 35%.

Yield anomalies are forecasted with high accuracy within the growing season in several countries. The model was used for a simple forecasting of yields up to two months before harvest. The results for countries with reported yields are shown in Figure 7, for all main producers using GGYD yields in SI Figure S15. In all but five (out of 14) cases the one-out-of-sample performance is equal or even higher than the standard model when omitting the last month of the reproductive season for training and prediction. In seven cases this holds also when omitting the last two months. In ten cases yield anomalies can be predicted better than by chance ($R^2_{O1} > 0.25$) two months before harvest, and in six cases this prediction accuracy is more than 50%. When using GGYD yield data, 25 of 63 cases can be predicted with at least 25% accuracy two months before harvest (representing 4-86% of global production depending on the crop), and in six cases with 50% accuracy (representing 0-51% of global production).

Mean warming suggests negative yield effects. When increasing temperatures by 0.9 or 1.4 °C above the 1980-2010 average, yields are predicted to lose 3-18% (excluding Australian wheat and Brazilian soybeans) in comparison to reported yield data (Table 2). Results for Russia had high uncertainties due to large $RMSE_{O1}$ values and standard deviations. Projections based on GGYD yields were not performed due to low R^2_{OOT} scores (Table 1**Error! Reference source not found.**).

DISCUSSION

We have applied a semi-empirical regression model to estimate weather influences on yields of maize, soybeans, spring and winter wheat. The model achieves good performance in explaining and predicting inter-annual yield variation in the US. For all main producer countries a high average explanatory power but varying out-of-sample prediction capacity is attained. The model shows medium to high accuracy for yield anomaly forecasts during the growing season up to two months before harvest. An application of the model with artificially increased temperatures suggests negative effects of moderate warming on crop yields.

Modeling yield anomalies in the US

The fraction of explained yield variation was at least two thirds and the one-out-of-sample yield prediction accuracy achieved 34-55%. The model also achieved a quantitative reproduction of negative yield anomalies in most cases, which is of particular importance when studying non-linear economic responses. When validating the model in the warmest or driest years its out-of-sample capacity is better than 25% in six of eight cases (Table 1**Error! Reference source not found.**, USDA).

Explanation (R^2) and projection (R^2_{O1}) capacity were strongly different (up to 0.65) in some cases, and more so for GGYD yields (SI Figure S6), underlining that both model fit and out-of-sample performance should be considered when evaluating the quality of a model (Holzkämper *et al.*, 2015, Landau *et al.*, 2000, Refsgaard *et al.*, 2013). Differences between NSE and R^2 values could be due to an over-proportional influence of outlier values or scale effects on the NSE.

The different out-of-sample performance of the model with USDA and GGYD yield data, in particular for soybeans and winter wheat, suggests several uncertainties of the gridded yield data. First, the combination of reported yields with remote sensing data and growing season modeling might not be apt for winter crops as these are more easily mixed with other vegetation. Second, the time series of the GGYD data is shorter by six years, leaving less data for out-of-sample estimations. Yet a regression with USDA yields in the shorter GGYD time frame produced similar results as with the full range (data not shown), thus the shorter time series alone is unlikely to explain different performances. Third, the equal or higher average R^2 with GGYD yield data (SI Figure S6) could possibly result from an implicit consideration of weather influences in the GGYD data set or the fitting of the model to more extreme values which arose in the GGYD construction but are not necessarily caused by weather. A misestimation of the true weather influence with our model would ensue. FAO yields, which are used in GGYD construction to calibrate remote sensing data, are often combined from reported and estimated data, adding a further layer of uncertainty. Fourth, yield variability from small plot sizes, in particular in developing countries, could be flattened at the coarse aggregate scale and thus blur weather influences. Fifth, GGYD yields showed lower CVs than USDA yields (except spring wheat, SI Table S2). This may explain the larger differences between R^2 and R^2_{O1} for GGYD yields, as low CVs together with shorter time series can lead to high correlations, but instable models i.e. a low R^2_{O1}. Similar differences in model performance between observed and remote sensing-derived yields in other nations (SI Table S4) further support our conclusions.

The geographical variation of model performance could have several causes. Different management techniques eliminate different shares of weather influence on crop yield. In particular irrigation, which is more prominent in the Western US (Schlenker & Roberts, 2009), marginalizes the effect of precipitation and also temperature (Lobell & Bonfils, 2008, Schauberger *et al.*, 2017). This is underlined by a lower model performance in this region (Figure 3). Thus, a low explanatory power might reflect a limited influence of weather on yields, as our model only detects weather impacts. Other reasons could include unconsidered, indirect weather influences (e.g. pests or diseases), errors in observations or aggregation effects. This may also explain the substantial share of grid cells with high yield variability but low explanatory power (SI Table S2**Error! Reference source not found.**). Low yield variability is difficult for any model to capture. Combined analysis of yield variation and model explanatory power reveals that areas with low yield variability are more likely to have a lower R^2 (SI Table S2, SI Figure S10). Areas with a high USDA yield CV, by contrast, have equal shares of high and low explained variance. Uncertainties introduced by interpolating yield or weather statistics could destroy their associations (Hansen & Jones, 2000). A comparison of our results using GGYD data to the global study by Ray *et al.* (2015), using reported data, revealed a similar or larger share of grid cells with substantial yield variability but unsatisfactory explained variance ($R^2 < 0.45$) in Ray *et al.* Our results suggest, again, that yield variability in many agricultural areas is influenced by more factors than only weather. These could include changing land-use patterns (Olmstead & Rhode, 2011), economic influences like fertilizer usage or stressors like ozone or pests.

The estimated coefficients and their geographical distributions agree with expectations. Maize reacted negatively to a high PET in the reproductive season and to very hot days (KDD) in particular in warmer regions – which agrees with previous findings (Lobell *et al.*, 2013, Schlenker & Roberts, 2009). This is contrary to expectations that C_4 crops would not experience much damage from mild heat (Sage & Kubien, 2007), but is likely due to water stress prior to direct heat damages (Schauberger *et al.*, 2017). This effect also explains the higher model performance for maize and soybeans in the South, where water stress is more dominant. PET in the vegetative season and solar radiation affected maize positively only in cooler regions, confirming previous studies (Long *et al.*, 2006, Rötter & Van de Geijn, 1999). Precipitation effects seem limited, though vegetative precipitation was usually positive. This conforms with a larger water demand of maize during the vegetative season (Hlavinka *et al.*, 2009). The relatively low precipitation coefficient values, despite its prominent importance (Barnabas *et al.*, 2008, Troy *et al.*, 2015), are due to comparably high and strongly varying input values (Gornott & Wechsung, 2016, Lobell *et al.*, 2013).

Differences in C_3 (soybeans, wheat) and C_4 (maize) photosynthesis efficiencies (Long *et al.*, 2006, Rötter & Van de Geijn, 1999) are reflected in a lower positive effect of SRT for maize. KDDs were less negative for winter wheat than for maize, since these hardly occur during the growing season – winter wheat is usually harvested before heat waves build up. A higher PET in the reproductive cycle was more detrimental than a higher PET in the vegetative cycle of either winter wheat or maize due to a more developed canopy. This also applies to precipitation effect differences between the reproductive winter wheat and the vegetative maize cycle. The model performance was low for all crops in the Northwest, and only slightly higher in the East North Central region. These regions seem more stable against weather fluctuations.

Six independent statistical tests indicated that our OLS estimation approach is applicable. Quadratic variables would not improve the model fit although this technique is often used to capture non-linear influences (Lobell *et al.*, 2011, Ray *et al.*, 2015). Autocorrelation occurring in many grid cells (SI Figure S9) points to periodically occurring yield variability, which might lead to an underestimation of standard errors with OLS. But this autocorrelation is due to autocorrelation in the raw yield data (55%, 32%, 31% and 37% of grid cells for maize, soybeans, spring and winter wheat, respectively, at 95% confidence level with a Ljung-Box test) and the first difference approach which produces correlated yield differences. Therefore we assume it as unproblematic for our analysis. The nationally aggregated time series was weakly autocorrelated for soybeans and winter wheat and not autocorrelated for maize and spring wheat.

When calculating yield variability on spatially aggregated level, a land-use weighting is usually applied to capture spatially divergent contributions to agricultural production. But model performance was better with unweighted yields except for winter wheat, whose growing area is less concentrated (SI Figure S3). Land-use patterns can be considered as an indirect function of climate since crops more favored by a certain climate also tend to have more area share. Thus there is an implicit inclusion of land-use patterns in the estimated coefficients, which makes the weighting negligible when inspecting aggregated yield variability. The differences are not substantial in all cases, which further suggests that land-use weighting can be omitted. This is beneficial for model generalization since weighting is another level of uncertainty (Cohn *et al.*, 2016, Porwollik *et al.*, 2016).

The model only used monthly aggregated weather data as input. This is an advantage over models requiring daily weather input since monthly aggregates are the preferred output from climate models (Taylor *et al.*, 2012) and are also less sensitive to outliers. The yield-anomaly approach of our model additionally eliminates any time-dependent systematic bias. It is therefore particularly apt for usage with data from climate models, which often require a bias correction before impact assessments (Hempel *et al.*, 2013).

Application to main producers

The generally good correlation between GGYD and FAO yield anomalies (SI Figure S12) allows us to interpret aggregated production from GGYD yields and MIRCA2000 areas as representative for main producing countries. The average R^2_{O1} was at least one third for maize and spring wheat. For soybeans and winter wheat average R^2_{O1} was low, which is likely due to shortcomings of GGYD data with these crops (see above and below). This is supported by the increased performance of the model when using reported yield data (SI Table S4).

More than half of the global maize and spring wheat production anomalies could be well explained by our model (R^2_{O1} at least 25%). This enables the usage of our model in global economic assessments. We assume this share to rise with more reported yield data.
Countries with a high predictive capacity of the model (R^2_{O1} above or around 50%) all have water-dominated yield variability, i.e. the majority of cultivated area being rainfed and a rather high alternation between deficient and sufficient precipitation. This suggests that the model particularly captures water-limiting signals, though this may be questioned by the low R^2_{OOP} with GGYD yields (Table 1). Wheat grown in Morocco and Turkey was classified as winter wheat due to its relatively long growing season (7-11 months) over the local winter, but is different from "classical" winter wheat grown in cooler nations where the crop experiences a vegetative pause over the winter. This could bias results towards lower R^2 values. The performance of our semi-empirical model, when run with reported yield data, was equal or superior to several previously applied statistical approaches (Iizumi *et al.,* 2013a, Lobell & Field, 2007, Ray *et al.*, 2015, Urban *et al.*, 2012).

We analyzed GGYD yields as an alternative to reported yields in areas where such data are currently not available. But the model-based nature of the data set could introduce a bias to our results. The robust performance of the semi-empirical model in the US, Germany, Russia, Burkina Faso, Tanzania and Brazil allows its usage for identifying cases where GGYD yields presumably suffer from a construction bias. We speculate that an existing weather influence on crops could be blurred by GGYD construction steps and is therefore less detectable with our (or any weather-driven) model. R^2 and R^2_{O1} values are then further apart, for example due to GGYD-processing induced yield extremes that are uncoupled from weather influences. The less convincing results for soybeans and winter wheat match with the evaluation by Iizumi *et al.* (2013b) suggesting that GGYD data likely requires improvement for both crops. A remaining concern is whether estimating a statistical model from a data set (GGYD) and then using the same model to evaluate these data may confound conclusions. But two additional analyses confirm our assumption that estimation problems occur more likely when GGYD yields are involved. First, the out-of-sample performance of models trained on reported yields is clearly superior to models trained on GGYD yields (SI Table S4). Second, a cross-comparison of model-predicted yields with reported FAO data, but where the model has been estimated with GGYD data (SI Figure S14), shows that there are discrepancies for all crops. Differences between predicted yields and FAO are usually smaller when using reported yields for training the model (dashed blue lines in Figure 2). Nevertheless we esteem the unique ability of GGYD yields to cover all regions of the globe where subnational yield data are otherwise difficult to obtain. Usage of latest satellite data with more sophisticated land-use separation methods may reduce counter-factual error sources and thus increase the reliability of satellite-derived yield statistics (Iizumi & Ramankutty, 2016).

Yield forecasting and warming experiment

The model concept allows for a simple extension towards forecasting of yields few months before harvest. This study presents a first example application in this direction. The forecasting is robust (R^2_{O1} > 50%) up to two months before harvest in several major producing countries, but requires improvement in others, in particular for soybeans and winter wheat. The performance is thus comparable to previous approaches (Bolton & Friedl, 2013, Johnson, 2014, Sakamoto *et al.*, 2014), but has been done here without any particular adaptation to country-specific conditions or model formulation. In several cases the reduced growing season leads to higher R^2_{O1} values than the full season. This could stem from three reasons. First, crop climatic requirements can be different in grain filling and maturity phase (Barnabas *et al.*, 2008), which are not distinguished in our reproductive season and could lead to meaningless coefficients in the default model. Second, the growing season dates in MIRCA2000 could be wrong, leading to an improvement when omitting a too long part. Third, the vegetative and reproductive season split could be misplaced. These reasons will have to be investigated in further studies. Again, the importance of high-quality input yield data for model training is highlighted: only then reliable within-season forecasts are possible, as evidenced by the lower performance with GGYD yields.
The forecasting scheme could be modified in two directions. Both require near-term monthly weather forecasts published, for example, by the NOAA (NOAA Climate Forecast, 2017). First, the full growing season can be used for training. In the season where yields should be predicted before harvest the missing part of the weather information is supplied by a near-term forecast. Second, both approaches can be combined: a reduced growing season, e.g. withholding the last two months of the season, is used for training. Yield predictions are then calculated for three or more months before harvest by supplying the missing weather information up to two months before harvest with near-term weather forecasts.

Predicting yields with counter-factual temperature increases is another model application case. The approach neglects CO_2 trends, variation of cofactors like precipitation and comes with high uncertainties (out-of-temperature performances in Table 1 and the *u* measure according to equation 5 provide a first, maybe too high estimate), which might mask effects. This could change if real climate scenarios were used including drifts in temperature extremes and precipitation. But impacts seem plausible in direction and magnitude compared to previous studies (Challinor *et al.*, 2014, Giannakopoulos *et al.*, 2009, Schleussner *et al.*, 2016). The low R^2_{OOT} performance for GGYD yields underlines the importance of high-quality yield data when projecting future yields. The average decline in wheat yields, when averaged over spring and winter wheat at 0.9°C warming (Table 2), is 6% – in agreement with the results by Liu *et al.* (2016). Thus the semi-empirical model described here can be considered a fourth method next to the three methods considered therein.

The model scheme presented in this study is an open concept that can be extended to incorporate further weather or economic factors. The prediction of yields within the growing season is highly sought after for timely adaptation measures in management, storage or marketing. Our model will be further developed in this direction. The differential performance between observed and remote-sensing based yield data calls for better and publicly available yield data from statistical offices in all countries. These can aid in planning adaptation or evaluating, for example, agricultural micro-insurance schemes.

ACKNOWLEDGEMENTS

BS acknowledges funding from the German National Academic Foundation. The research was supported by projects Trans-SEC, funded by BMBF and co-financed by BMZ, and EXTRA, funded by BMBF. We thank T. Iizumi for data provision and four anonymous reviewers that helped to improve the manuscript. Author contributions: FW initiated, BS and FW designed the study with support by CG. BS performed the study and wrote the manuscript with contributions by FW and CG.

REFERENCES

Barlow KM, Christy BP, O'leary GJ, Riffkin PA, Nuttall JG (2015) Simulating the impact of extreme heat and frost events on wheat crop production: A review. Field Crops Research, **171**, 109-119.

Barnabas B, Jager K, Feher A (2008) The effect of drought and heat stress on reproductive processes in cereals. Plant Cell Environ, **31**, 11-38.

Belsley DA, Kuh E, Welsch RE (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York, Wiley.

Bolton DK, Friedl MA (2013) Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology, **173**, 74-84.

Challinor AJ, Watson J, Lobell DB, Howden SM, Smith DR, Chhetri N (2014) A meta-analysis of crop yield under climate change and adaptation. Nature Climate Change, **4**, 287-291.

Cohn AS, Vanwey LK, Spera SA, Mustard JF (2016) Cropping frequency and area response to climate variability can exceed yield response. Nature Climate Change, **6**, 601-604.

Conradt T, Gornott C, Wechsung F (2016) Extending and improving regionalized winter wheat and silage maize yield regression models for Germany: Enhancing the predictive skill by panel definition through cluster analysis. Agricultural and Forest Meteorology, **216**, 68-81.

Croissant Y, Millo G (2008) Panel data econometrics in R: the plm package. J. Stat. Softw., **27**, 1-43.

Di Paola A, Valentini R, Santini M (2016) An overview of available crop growth and yield models for studies and assessments in agriculture. Journal of the Science of Food and Agriculture, **96**, 709-714.

Fao (2016) FAOStat, http://faostat3.fao.org/home/E. pp Page.

Giannakopoulos C, Le Sager P, Bindi M, Moriondo M, Kostopoulou E, Goodess CM (2009) Climatic changes and associated impacts in the Mediterranean resulting from a 2 °C global warming. Global and Planetary Change, **68**, 209-224.

Glotter M, Elliott J, Mcinerney D, Best N, Foster I, Moyer EJ (2014) Evaluating the utility of dynamical downscaling in agricultural impacts projections. Proc Natl Acad Sci U S A, **111**, 8776-8781.

Gornott C, Wechsung F (2016) Statistical regression models for assessing climate impacts on crop yields: A validation study for winter wheat and silage maize in Germany. Agricultural and Forest Meteorology, **217**, 89-100.

Hansen JW, Jones JW (2000) Scaling-up crop models for climate variability applications. Agricultural Systems, **65**, 43-72.

Hatfield JL, Boote KJ, Kimball BA *et al.* (2011) Climate Impacts on Agriculture: Implications for Crop Production. Agronomy Journal, **103**, 351-370.

Haude W (1955) Zur Bestimmung der Verdunstung auf möglichst einfache Weise. Mitteilungen des Deutschen Wetterdienstes, **11**.

Hempel S, Frieler K, Warszawski L, Schewe J, Piontek F (2013) A trend-preserving bias correction - the ISI-MIP approach. Earth System Dynamics, **4**, 219-236.

Hlavinka P, Trnka M, Semerádová D, Dubrovský M, Žalud Z, Možný M (2009) Effect of drought on yield variability of key crops in Czech Republic. Agricultural and Forest Meteorology, **149**, 431-442.

Holzkämper A, Calanca P, Honti M, Fuhrer J (2015) Projecting climate change impacts on grain maize based on three different crop model approaches. Agricultural and Forest Meteorology, **214-215**, 219-230.

Iizumi T, Ramankutty N (2016) Changes in yield variability of major crops for 1981–2010 explained by climate change. Environmental Research Letters, **11**, 034003.

Iizumi T, Sakuma H, Yokozawa M *et al.* (2013a) Prediction of seasonal climate-induced variations in global food production. Nature Climate Change, **3**, 904-908.

Iizumi T, Uno F, Nishimori M (2012) Climate Downscaling as a Source of Uncertainty in Projecting Local Climate Change Impacts. Journal of the Meteorological Society of Japan. Ser. II, **90B**, 83-90.

Iizumi T, Yokozawa M, Sakurai G *et al.* (2013b) Historical changes in global yields: major cereal and legume crops from 1982 to 2006. Global Ecology and Biogeography, **23**, 346-357.

Ipcc (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. (eds Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM) pp Page, Cambridge, United Kingdom and New York, NY, USA, .

Johnson DM (2014) An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment, **141**, 116-128.

Jones JW, Antle JM, Basso B *et al.* (2016) Brief history of agricultural systems modeling. Agricultural Systems.

Kilsby CG, Jones PD, Burton A *et al.* (2007) A daily weather generator for use in climate change studies. Environmental Modelling & Software, **22**, 1705-1719.

Landau S, Mitchell RaC, Barnett V, Colls JJ, Craigon J, Payne RW (2000) A parsimonious, multiple-regression model of wheat yield response to environment. Agricultural and Forest Meteorology, **101**, 151-166.

Liu B, Asseng S, Ewert F *et al.* (2016) Similar estimates of temperature impacts on global wheat yield by three independent methods. Nature Climate Change.

Lobell DB (2013) Errors in climate datasets and their effects on statistical crop models. Agricultural and Forest Meteorology, **170**, 58-66.

Lobell DB, Bänziger M, Magorokosho C, Vivek B (2011) Nonlinear heat effects on African maize as evidenced by historical yield trials. Nature Climate Change, **1**, 42-45.

Lobell DB, Bonfils C (2008) The Effect of Irrigation on Regional Temperatures: A Spatial and Temporal Analysis of Trends in California, 1934–2002. Journal of Climate, **21**, 2063-2071.

Lobell DB, Burke MB (2010) On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, **150**, 1443-1452.

Lobell DB, Field CB (2007) Global scale climate–crop yield relationships and the impacts of recent warming. Environmental Research Letters, **2**, 014002.

Lobell DB, Hammer GL, Mclean G, Messina C, Roberts MJ, Schlenker W (2013) The critical role of extreme heat for maize production in the United States. Nature Climate Change, **3**, 497-501.

Long SP, Zhu X-G, Naidu SL, Ort DR (2006) Can improvement in photosynthesis increase crop yields? Plant, Cell and Environment, **29**, 315-330.

Luo Q (2011) Temperature thresholds and crop production: a review. Climatic Change, **109**, 583-598.

Maurer EP, Hidalgo HG, Das T, Dettinger MD, Cayan DR (2010) The utility of daily large-scale climate data in the assessment of climate change impacts on daily streamflow in California. Hydrology and Earth System Sciences, **14**, 1125-1138.

Mueller ND, Gerber JS, Johnston M, Ray DK, Ramankutty N, Foley JA (2012) Closing yield gaps through nutrient and water management. Nature, **490**, 254-257.

Noaa Climate Forecast (2017) http://www.cpc.ncep.noaa.gov/products/forecasts/. pp Page.
Olmstead AL, Rhode PW (2011) Adapting North American wheat production to climatic challenges, 1839 – 2009. Proc Natl Acad Sci U S A, **108**, 480-485.
Porter JR, Gawith M (1999) Temperatures and the growth and development of wheat a review. European Journal of Agronomy, **10**, 23-36.
Portmann FT, Siebert S, Döll P (2010) MIRCA2000-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. Global Biogeochemical Cycles, **24**, GB1011.
Porwollik V, Müller C, Elliott J *et al.* (2016) Spatial and temporal uncertainty of crop yield aggregations. European Journal of Agronomy.
R Core Team (2016) R: A Language and Environment for Statistical Computing. (ed Computing RFFS) pp Page, R Foundation for Statistical Computing.
Rahmstorf S (2007) A Semi-Empirical Approach to Projecting Future Sea-Level Rise. Science, **315**, 368-370.
Ray DK, Gerber JS, Macdonald GK, West PC (2015) Climate variation explains a third of global crop yield variability. Nat Commun, **6**, 5989.
Refsgaard JC, Madsen H, Andréassian V *et al.* (2013) A framework for testing the ability of models to project climate change and its impacts. Climatic Change, **122**, 271-282.
Ritchie SW, Hanway JJ, Benson GO, Herman JC, Lupkes SJ (1993) How a Soybean Plant Develops. pp Page, Ames, Iowa State University of Science and Technology.
Rötter RP, Carter TR, Olesen JE, Porter JR (2011) Crop–climate models need an overhaul. Nature Climate Change, **1**, 175-177.
Rötter RP, Van De Geijn S (1999) Climate change effects on plant growth, crop yield and livestock. Climatic Change, **43**, 651-681.
Ruane AC, Goldberg R, Chryssanthacopoulos J (2015) Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation. Agricultural and Forest Meteorology, **200**, 233-248.
Sage RF, Kubien DS (2007) The temperature response of C(3) and C(4) photosynthesis. Plant Cell Environ, **30**, 1086-1106.
Sakamoto T, Gitelson AA, Arkebauer TJ (2014) Near real-time prediction of U.S. corn yields based on time-series MODIS data. Remote Sensing of Environment, **147**, 219-231.
Sanchez B, Rasmussen A, Porter JR (2014) Temperatures and the growth and development of maize and rice: a review. Glob Chang Biol, **20**, 408-417.
Schauberger B, Archontoulis S, Arneth A *et al.* (2017) Consistent negative response of US crops to high temperatures in observations and crop models. Nat Commun, **8**, 1-9.
Schlenker W, Roberts MJ (2009) Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proc Natl Acad Sci U S A, **106**, 15594-15598.
Schleussner C-F, Lissner TK, Fischer EM *et al.* (2016) Differential climate impacts for policy-relevant limits to global warming: the case of 1.5 °C and 2 °C. Earth System Dynamics, **7**, 327-351.
Taylor KE, Stouffer RJ, Meehl GA (2012) An Overview of CMIP5 and the Experiment Design. Bulletin of the American Meteorological Society, **93**, 485-498.
Troy TJ, Kipgen C, Pal I (2015) The impact of climate extremes and irrigation on US crop yields. Environmental Research Letters, **10**, 054013.
Urban D, Roberts MJ, Schlenker W, Lobell DB (2012) Projected temperature changes indicate significant increase in interannual variability of U.S. maize yields. Climatic Change, **112**, 525-533.
Usda (2015) USDA Quickstats, http://quickstats.nass.usda.gov/. pp Page.
Wechsung F, Lüttger AB, Hattermann F (2008) Projektionen zur klimabedingten Änderung der Erträge von einjährigen Sommer-und Winterkulturen des Ackerlandes am Beispiel

von Silomais und Winterweizen. In: *PIK Report.* pp Page.
Wooldridge JM (2013) *Introductory Econometrics. A Modern Approach*, South Western Cengage Learning.

TABLES

***Table 1**: Model performance for eight crop-yield data set combinations in the US. Columns are crop, yield data set, application of land-use weighted aggregation, Nash-Sutcliffe efficiency (NSE), explained variance of the modeled (R^2) and one-out-of-sample time series (R^2_{O1}), out-of-temperature and out-of-precipitation correlation (R^2_{OOT} and R^2_{OOP}) and the share of grid cells for which the model is significant ($p < 0.05$).*

Crop	Yield data	Weighted Aggregation	NSE	R^2	R^2_{O1}	R^2_{OOT}	R^2_{OOP}	Significant Cells
Maize	USDA	No	0.74	0.81	0.55	0.31	0.11	51 %
	GGYD	No	0.70	0.92	0.59	0.08	r<0	47 %
Soybeans	USDA	No	0.69	0.69	0.45	0.38	0.02	60 %
	GGYD	Yes	0.60	0.72	0.18	r<0	r<0	24 %
Spring wheat	USDA	No	0.63	0.63	0.34	0.28	0.42	52 %
	GGYD	No	0.61	0.73	0.32	r<0	0.34	48 %
Winter wheat	USDA	Yes	0.64	0.65	0.35	0.33	0.28	50 %
	GGYD	Yes	0.55	0.91	0.26	0.00	0.00	10 %

Table 2*: Yield effects (as fraction of average historic yields) of artificial temperature increases, using only reported yield data. Fractions were normalized with T+0 offset. Values in brackets are uncertainty measures u (+/-) of the fraction according to equation 5.*

Crop	Country	T +0.9 °C	T +1.4 °C
Maize	USA	0.96 (0.07)	0.95 (0.07)
	Russia	0.88 (0.87)	0.85 (0.86)
	Brazil	0.97 (0.19)	0.95 (0.20)
	Germany	0.96 (0.09)	0.94 (0.09)
	Burkina Faso	0.95 (1.00)	0.94 (1.00)
Soybeans	USA	0.97 (0.16)	0.96 (0.17)
	Brazil	1.00 (0.12)	1.00 (0.12)
Spring wheat	USA	0.95 (0.16)	0.92 (0.17)
	Australia	1.05 (0.71)	1.07 (0.74)
	Russia	0.89 (0.77)	0.84 (0.83)
Winter wheat	USA	0.97 (0.07)	0.95 (0.07)
	Russia	0.88 (0.72)	0.82 (0.78)
	Germany	0.95 (0.06)	0.92 (0.07)
	Brazil	0.89 (0.32)	0.85 (0.36)

FIGURE CAPTIONS

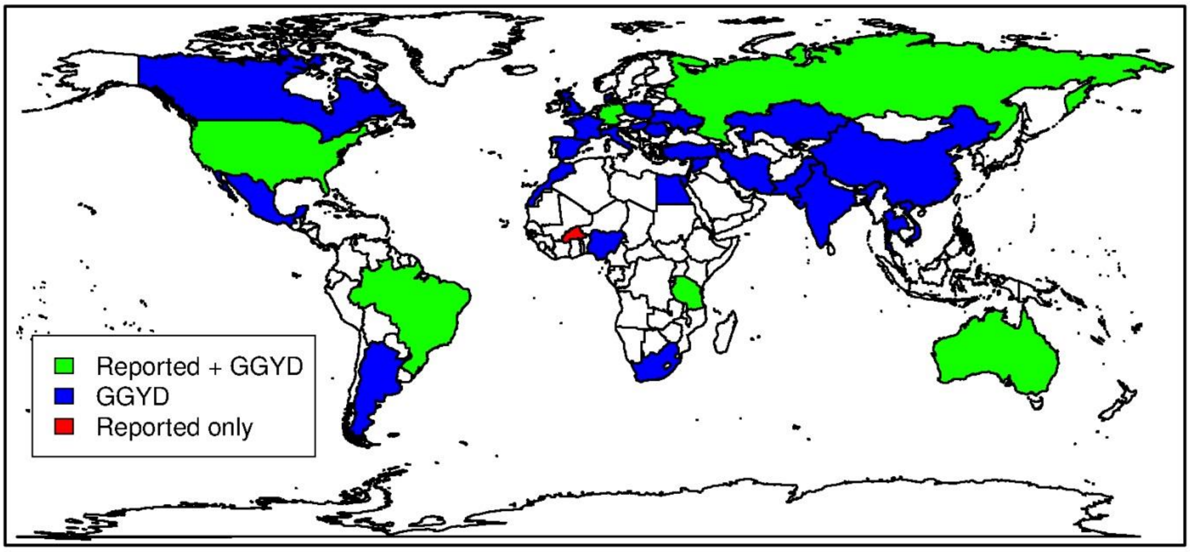


***Figure 1**: World map of countries analyzed in this study. Colors of countries denote whether GGYD and reported yields (green), only GGYD yields (blue) or only reported yields (red) are used in this study. Countries in white are no main producers and not analyzed.*

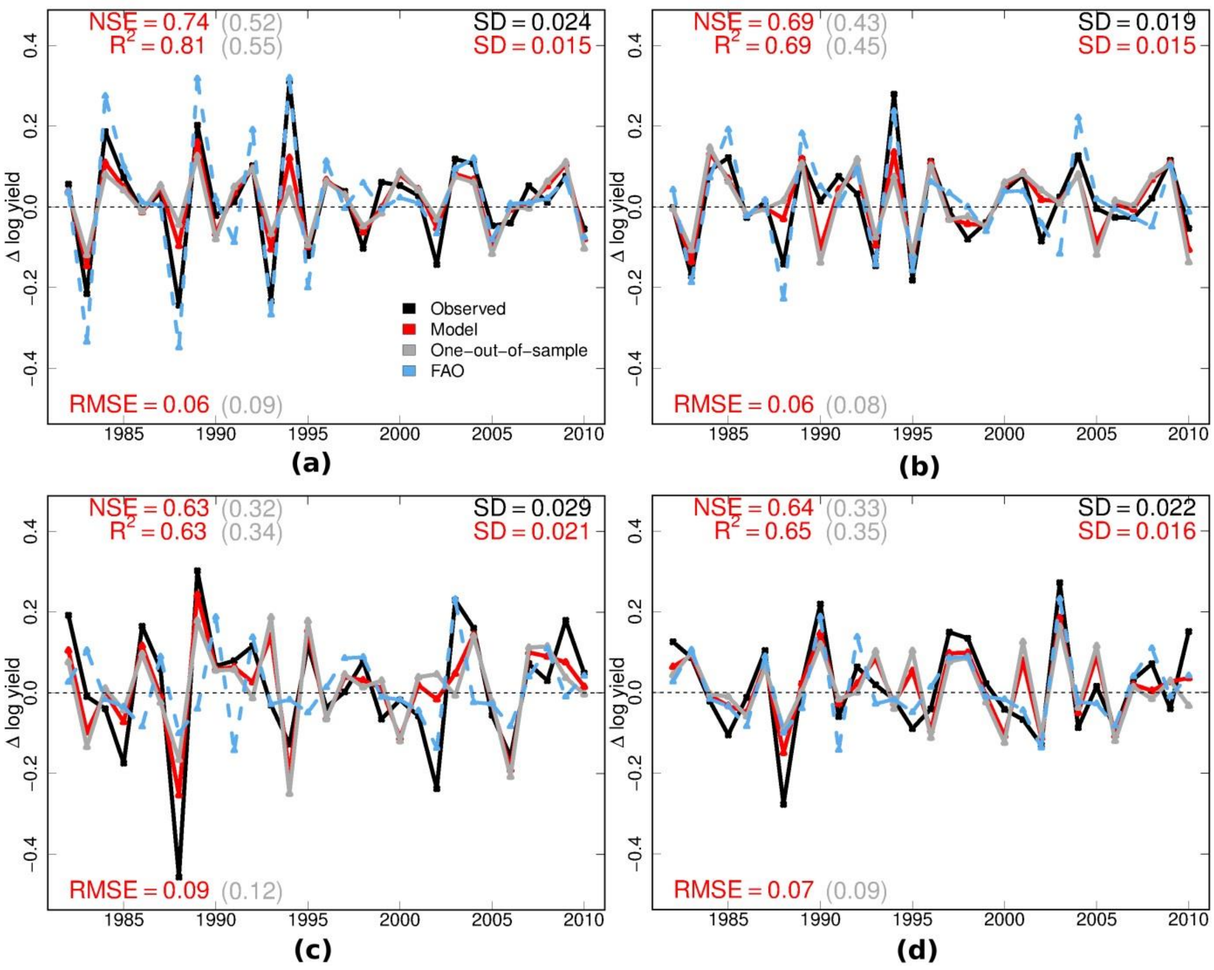


Figure 2*: Observed and modeled time series of national US yield anomalies for maize (a), soybeans (b), spring wheat (c) and winter wheat (d). Black lines are anomalies of reported USDA yields, red lines are anomalies predicted by the model trained on the full data panel, gray lines are anomalies predicted from one-out-of-sample models, and blue dashed lines are FAO yield anomalies. Data points were 56,092, 38,373, 21,291 and 58,877 for maize, soybeans, spring and winter wheat, respectively. Numbers in plots are performance measures and standard deviation (SD); colors of numbers correspond to the respective anomaly series. Modelled and FAO yield anomalies were significantly ($p < 0.05$) correlated for maize (Pearson's $r = 0.87$), soybeans (0.69) and winter wheat (0.68), but not for spring wheat (0.13), since FAO yields combine spring and winter wheat.*

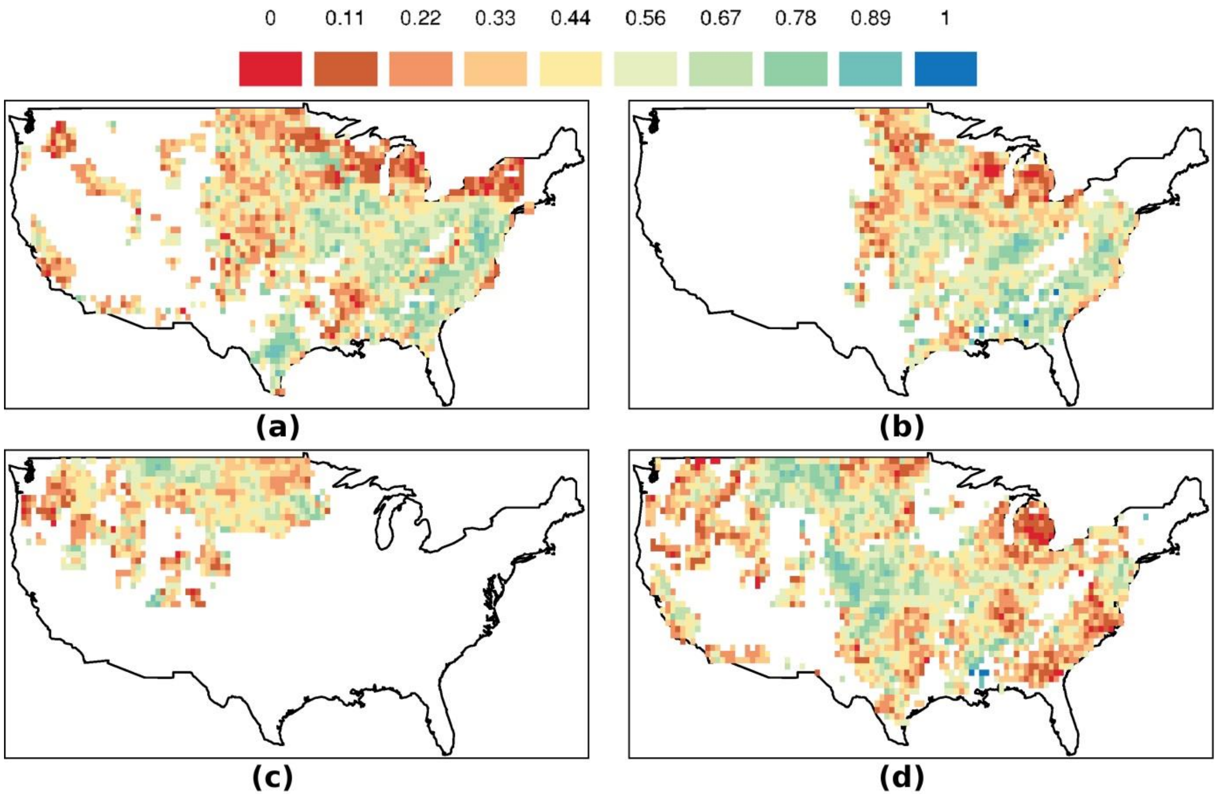


Figure 3*: Explained variance of yield anomalies due to weather anomalies (R^2, color map on top) for maize (a), soybeans (b), spring wheat (c) and winter wheat (d) with USDA yields. White regions have no cropping area.*

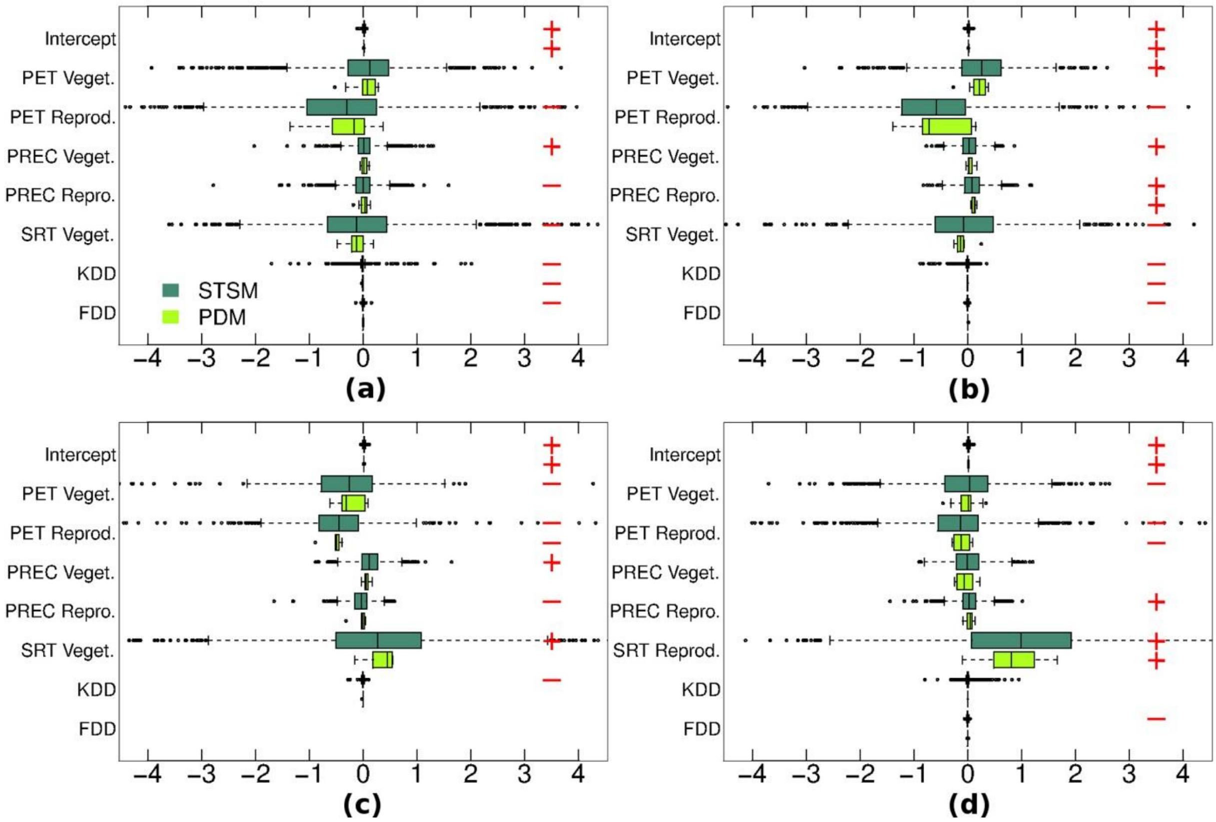


Figure 4*: Coefficient comparison for STSM and PDM model estimation for maize (a), soybeans (b), spring wheat (c) and winter wheat (d) with USDA yields. Blue boxes show coefficients with STSM estimation (estimated for each grid cell), while green boxes show PDM coefficients (estimated for each climate region). The band inside each box is the median, while boxes represent 25% and 75% quantiles. Whiskers are defined as the maximum and minimum as long as both values are within the 1.5 interquartile range from the median. Otherwise the last points in this range are shown with whiskers and outliers are depicted as points. Red +/- symbols indicate a mean significantly larger/lower than 0 (t-test at 95% confidence level).*

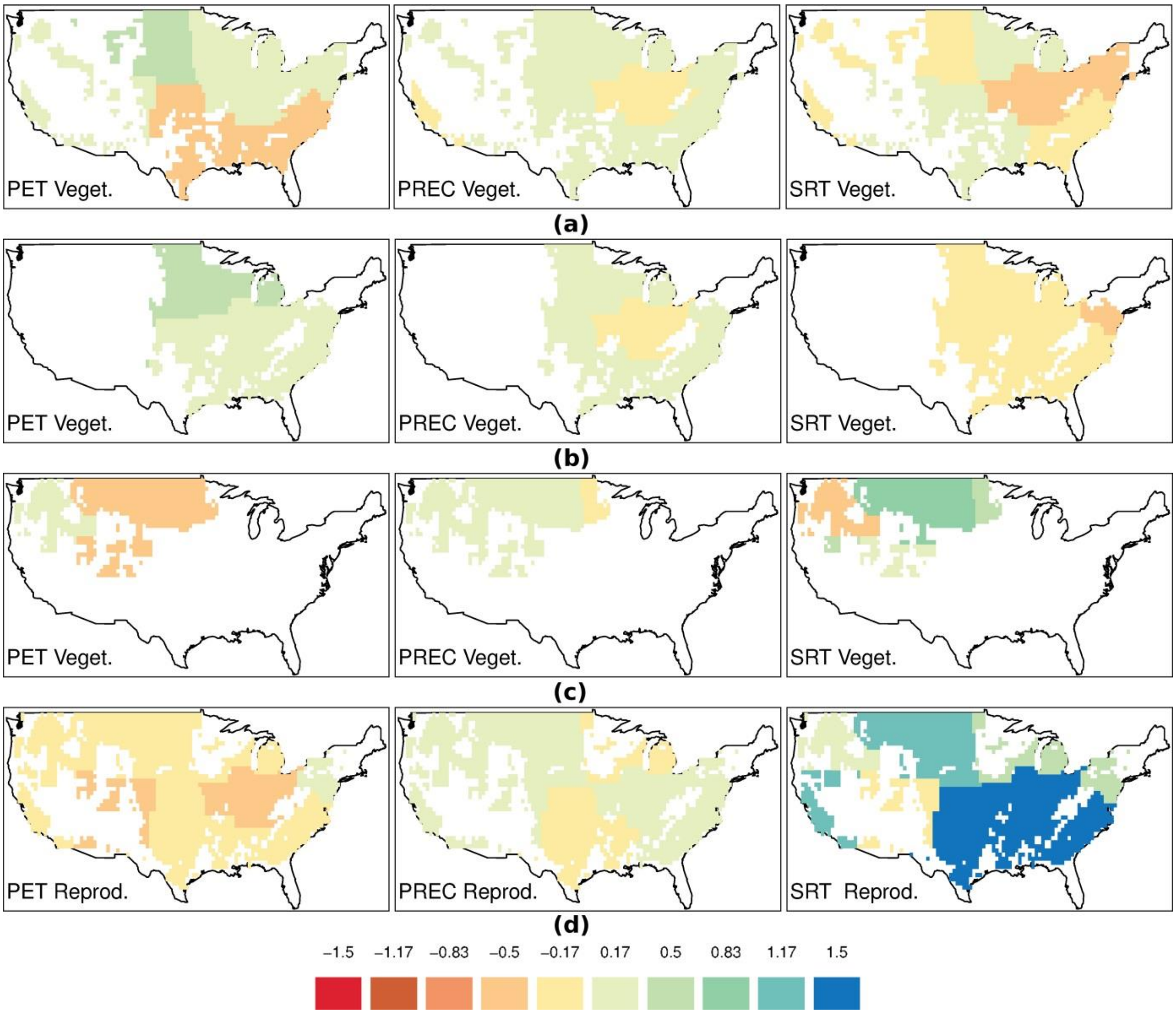
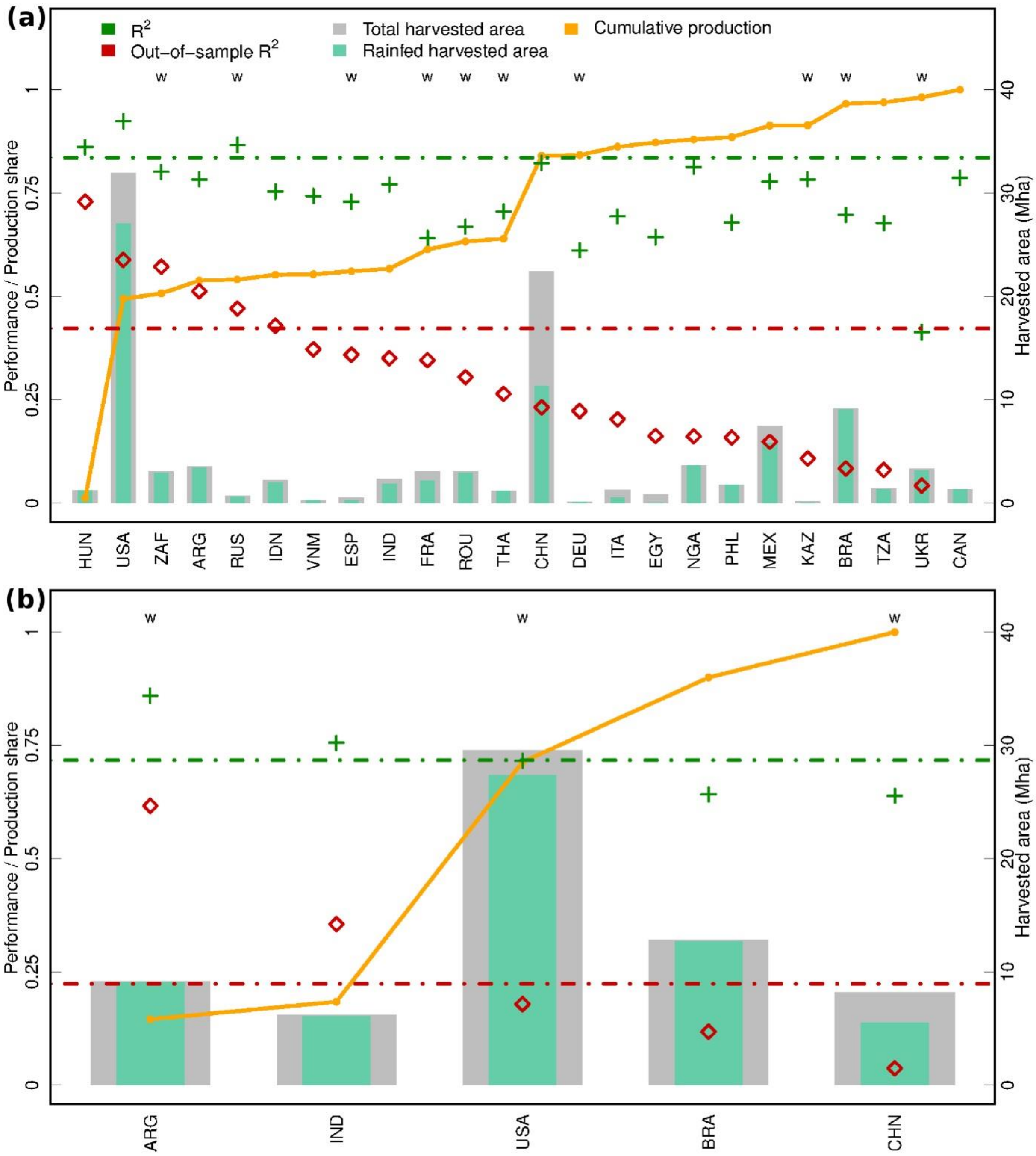


Figure 5*: Estimated coefficients for USDA yields. Rows are maize (a), soybeans (b), spring wheat (c) and winter wheat (d). Coefficients were estimated with STSM regression and aggregated from grid cells to climate regions. From left to right the coefficients are PET in vegetative (maize, soybeans, spring wheat) or reproductive (winter wheat) season, precipitation and SRT in the same seasons, respectively. Color map is shown at bottom.*


(a)
R^2
Out-of-sample R^2
Total harvested area
Rainfed harvested area
Cumulative production
Performance / Production share
Harvested area (Mha)
w
HUN
USA
ZAF
ARG
RUS
IDN
VNM
ESP
IND
FRA
ROU
THA
CHN
DEU
ITA
EGY
NGA
PHL
MEX
KAZ
BRA
TZA
UKR
CAN
(b)
ARG
IND
USA
BRA
CHN

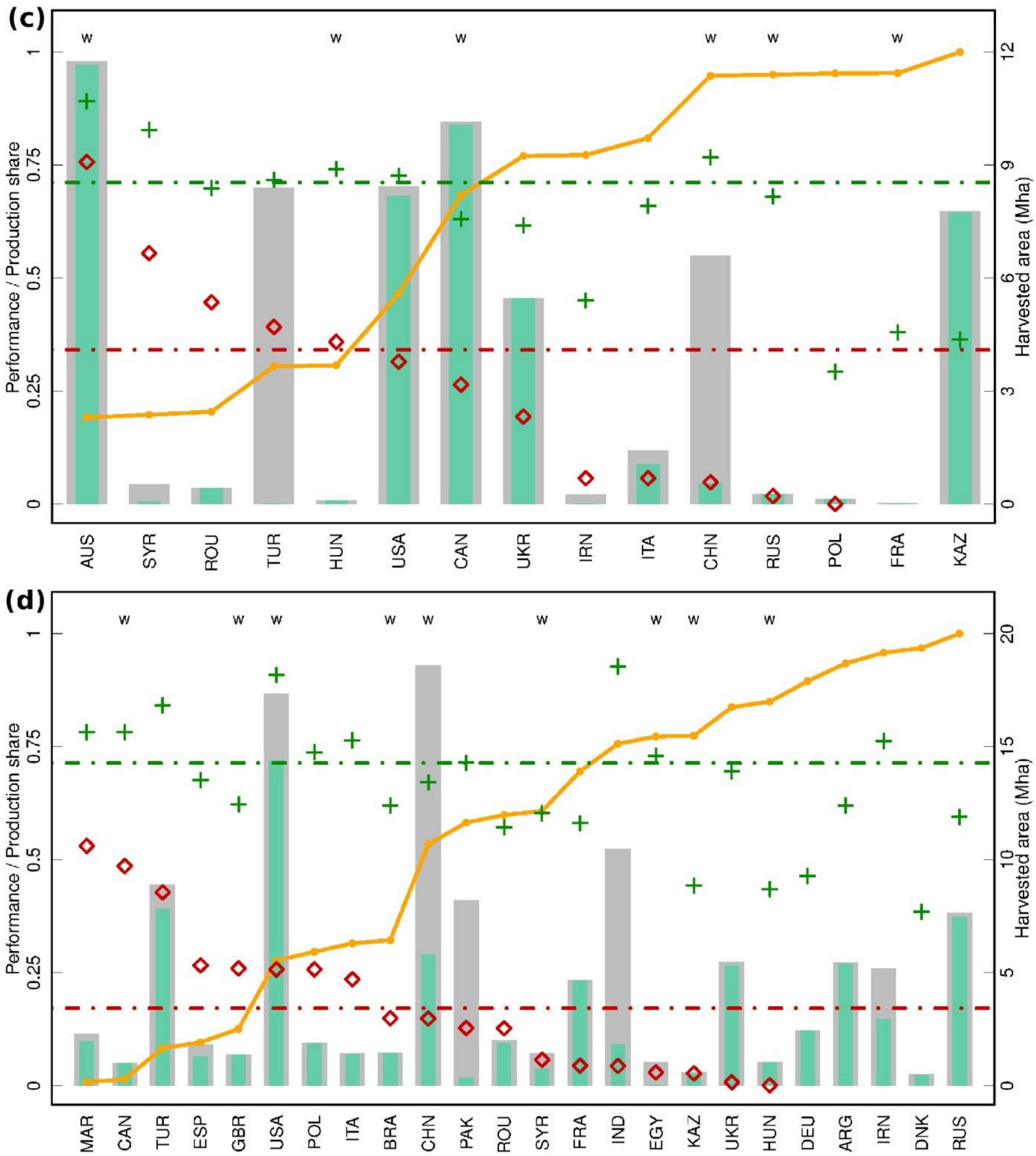


***Figure 6**: Performance of STSM models in main producing countries for maize (panel a), soybeans (b), spring wheat (c) and winter wheat (d). Countries are ordered by descending R^2_{O1}; three-letter codes are provided in SI Table S3. Green crosses mark R^2 and red diamonds R^2_{O1} values (left y axis). The mean R^2 and R^2_{O1} over all main producers, weighted by production, are indicated with dashed green and red lines, respectively. A "w" above countries indicates that the displayed R^2_{O1} value is achieved when including land-use weighting. Gray and blue bars denote total and rainfed harvested area in Mha, respectively (right y axis). The orange line denotes cumulative production share among main producers (left y axis).*

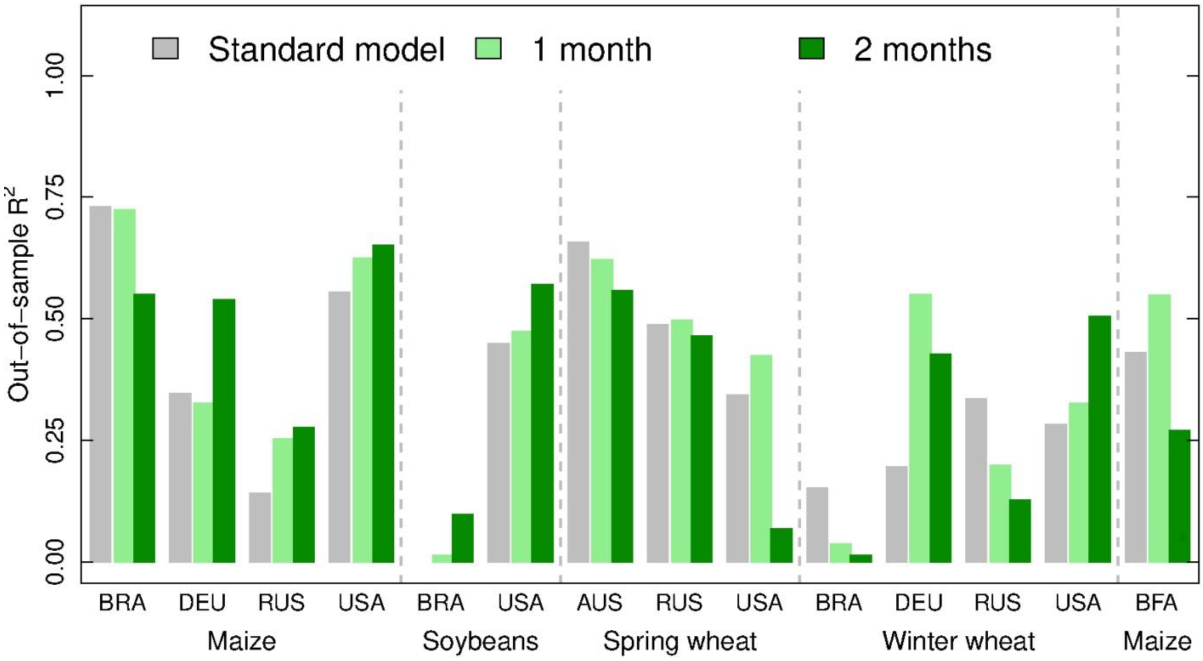


***Figure 7**: Capacity of the model for yield forecasting within the growing season, using only reported yield data. The one-out-of-sample performance R^2_{O1} is shown. Gray bars are the standard model with full growing season used for training and prediction. Green and black bars show performance when withholding one or two months, respectively, for training the model and predicting yield anomalies out of sample. Burkina Faso (BFA) is not a main producer and therefore plotted off set.*