Comparing the effects of calibration and climate errors on a statistical crop model and a process-based crop model

**James Watson · Andrew J. Challinor ·
Thomas E. Fricker · Christopher A. T. Ferro**

Abstract Understanding the relationship between climate and crop productivity is a key component of projections of future food production, and hence assessments of food security. Climate models and crop yield datasets have errors, but the effects of these errors on regional scale crop models is not well categorized and understood. In this study we compare the effect of synthetic errors in temperature and precipitation observations on the hindcast skill of a process-based crop model and a statistical crop model. We find that errors in temperature data have a significantly stronger influence on both models than errors in precipitation. We also identify key differences in the responses of these models to different types of input data error. Statistical and process-based model responses differ depending on whether synthetic errors are overestimates or underestimates. We also investigate the impact of crop yield calibration data on model skill for both models, using datasets of yield at three different spatial scales. Whilst important for both models, the statistical model is more strongly influenced by crop yield scale than the process-based crop model. However, our results question the value of high resolution yield data for improving the skill of crop models; we find a focus on accuracy to be more likely to be valuable. For both crop models, and for all three spatial scales of yield calibration data, we found that model skill is greatest where growing area is above 10-15 %. Thus information on area harvested would appear to be a priority for data collection efforts. These results are important for three reasons. First, understanding how different crop models rely on different characteristics of temperature, precipitation and crop yield data allows   us to match the model type to the available data.

J. Watson (✉) · A. J. Challinor
Institute for Climate and Atmospheric Science, School of Earth and Environment,
University of Leeds, Leeds, UK
e-mail: j.watson@uq.edu.au

T. E. Fricker · C. A. T. Ferro
College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Second, we can prioritize where improvements in climate and crop yield data should be directed. Third, as better climate and crop yield data becomes available, we can predict how crop model skill should improve.

1 Introduction

While knowledge of crop physiology comes from experiments at the field scale, climate models have skill at the regional scale. Regional scale crop models have been developed as principled frameworks for upscaling field scale knowledge to the regional scale, in order to help capture and explore the key crop-climate processes. While regional scale crop models can differ significantly in their structure and assumptions, they all rely on the quality of available climate and crop yield data. 'Quality' of this data is not necessarily a matter of higher temporal/spatial resolution. Rather it depends on whether the model-significant statistics of the input data accurately reflect reality.

The projected response of crops to climate variability and change can vary significantly according to the methodology chosen (Challinor et al. 2014). This variation can be ascribed to three causes: structural differences between crop models, differences in crop calibration data, and differences in weather inputs. Structural differences in models result from the choice of parameterisations for representing crop growth and development (White et al. 2011). These choices are often related to the spatial scale for which the model is designed (Challinor and Wheeler 2008). Choices regarding model calibration are also related to the spatial scale of the assessment: regional-scale models typically have less crop growth and development data for calibration. Calibration and application of models at regional-scales invariably involves simplifying spatial heterogeneity, and can therefore result in aggregation error (Hansen and Jones 2000).

Disentangling these three sources of uncertainty is not trivial. Efforts to separate structural model uncertainty from calibration uncertainty have begun, and show promise (Asseng et al. 2013). Frameworks for measuring and interpreting climate model uncertainty have, at least for the case of climate change, a somewhat longer history (Ramirez-Villegas et al. 2013). Recent work includes assessments of the uncertainty associated with bias correction of climate model output (Hawkins et al. 2013b; Koehler et al. 2013). In order to identify the precise sources of climate-induced uncertainty in crop yield, recent studies have also systematically perturbed weather inputs (Berg et al. 2010; Lobell 2013; Watson and Challinor 2013).

There are significant structural differences between process-based and statistical model crop models. The latter can do an excellent job of reproducing historical temperature-induced yield variation at regional scales (Hawkins et al. 2013a), whilst the former are useful for determining the causes of yield variation (Lobell et al. 2013), and may be more robust to non-stationarity in the relationships between weather and crop yield. Comparisons between process-based and statistical models are at an early stage. Estes et al. (2013) found that a statistical model produces larger climate losses than a process-based model for maize and wheat in South Africa, leading them to recommend increased intercomparison of these two types of model.

This study draws on the research described above in order to examine interactions between structural model uncertainty, input weather uncertainty, and input calibration uncertainty. We apply systematic perturbations to observed weather and record the impact on the skill of two regional-scale models: one process-based and one statistical. In order to assess calibration uncertainty, three calibration configurations are used. Each configu-

ration uses yield at one of three spatial scales, plus crop harvested area. Thus we define "calibration" as any use of any crop observational data to improve crop model results.

By analyzing the effects of calibration uncertainty and systematic weather errors using two structurally distinct crop models, this study quantifies how different model types can be influenced by different input data uncertainties. This information is important for determining whether model and input data are fit for purpose in a given study, and informs the allocation of resources for future improvements. In addition, analyses such as this study provide information on how crop model skill may improve as weather and calibration data improve.

2 Materials and methods

Statistical crop models are typically designed for a particular study. Thus, to perform a direct comparison between a statistical crop model and a process-based crop model, this analysis replicated the maize crop hindcast scenario of Hawkins et al. (2013a). A process-based crop model was ported to the same scenario. Both these models are described below.

2.1 Statistical crop model

The statistical crop model developed by Hawkins et al. (2013a) for maize in France was used as a case study for this analysis. This empirical model relates temperature, precipitation, and maize yield and is trained using historical data. This statistical model is defined as follows:

$$Y(t) = g(t) + \beta_1 X(t) + \beta_2 (P(t) - \bar{P}) + \beta_3 X(t) P(t) + e(t) \tag{1}$$

where for a given year t: Y is maize yield, X is the number of days above a temperature threshold (32 °C) for the June-August growing season, P is mean precipitation for June-August, g is the expected yield given average precipitation and no hot days, and e is a stochastic error term. The β parameters and g are trained using a penalized likelihood function, with g being a cubic regression spline. For further details of this model, see Hawkins et al. (2013a). Since some studies assume a linear technology trend (e.g., Lobell and Asner 2003; de Wit and van Diepen 2007), we analyzed the model just described ($STAT_{nonlinear}$), as well as the case where a linear trend is used for $g(t)$ ($STAT_{linear}$).

The historical daily temperature and precipitation data was obtained from the E-OBS gridded observational dataset (Haylock et al. 2008). Whereas Hawkins et al. used E-OBS version 5.0 on a 0.5° x 0.5° grid, we used version 7.0 on a 0.25° x 0.25° latitude/longitude grid. This higher resolution grid provided 1,035 locations in the study region.

The seasonal yield data was taken from two sources, which together provided observations from three spatial scales (see Fig. 1). Country and regional scale data was obtained for France from EUROSTAT,[1] and departmental scale observations were from AGRESTE - Statistique Agricole Annuelle, obtained via ARVALIS - Institut du végétal. A total of 22 regions and 96 departments were used in this study. Yield observations covering all three of these spatial scales were only available from 1980–2007, so this study only investigated that time period. To allow a direct comparison of input datasets with the GLAM crop model (described below), the yield data was linearly detrended with the level set to that of the start year.

---

[1]http://epp.eurostat.ec.europa.eu/
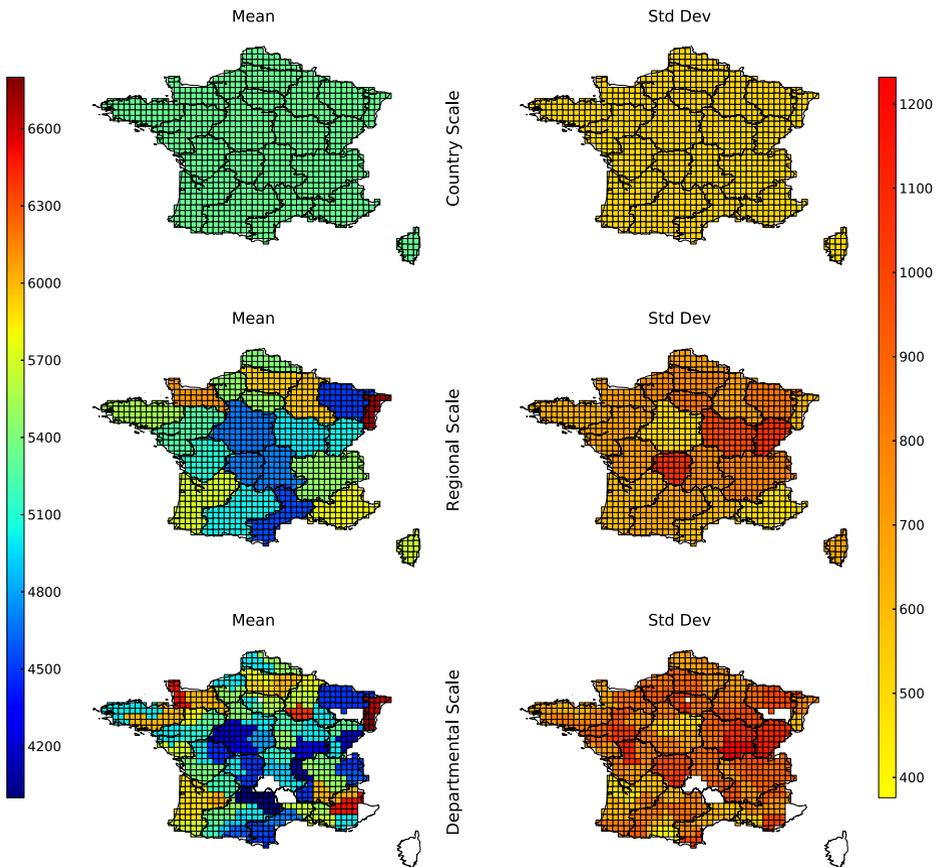
Fig. 1 Mean and standard deviation of yield observations for each E-OBS grid cell. There are clear differences in the observations used at each location, depending on whether the source of the data was country scale (*top row*), regional scale (*middle row*) or departmental scale (*bottom row*). While the country scale mean is near the mid-point of the means reported across all datasets (5320.0 kg ha^{-1}; *left column*), country scale standard deviation is low *(511.88 kg ha^{-1}; right column)*. Locations without yield observations for the entire study period (1980–2007) were ignored

The statistical crop model was run at each location on the E-OBS grid, so each of the kg ha^{-1} yield observation datasets were regridded to the E-OBS data's resolution. The weighted mean of harvest area was calculated for each E-OBS grid cell according to the dataset provided by Monfreda et al. (2008).

2.2 Process-based crop model

We used the General Large Area Model for annual crops (GLAM; Challinor et al. 2004) as a case study of a process-based regional-scale crop model. This model has been used to simulate a range of crop types in both present-day and future climates (Koehler et al. 2013; Challinor et al. 2005). The maize version of the model was adapted from an African model developed by Greatrex (2012). GLAM relies on values for precipitation, minimum and maximum temperature, solar radiation, CO_2 level, planting window and soil hydrological properties to simulate crop development and yield. Calibration of the model is performed

by adjusting the yield gap parameter (YGP) such that the difference between simulated and observed mean yields is minimized. Maize accounts for approximately 56 % of the irrigated area of France (data retrieved from AQUASTAT provided by the FAO),[2] but the ratio of irrigated to non-irrigated maize area varies significantly across the country.[3] Since we wanted to compare the effects of systematic error in both temperature and precipitation, two configurations of GLAM were analyzed – one with irrigation ($GLAM_{irr}$) and one that was rainfed ($GLAM_{rfd}$).

GLAM was run on the same E-OBS grid as the statistical model. The daily temperature and precipitation values for the study period 1980–2007 were taken from the E-OBS dataset as described in Section 2.1. The yield and harvest area input data were identical to that used for the statistical model. The CO_2 concentration was set to 357.07 ppm, the value observed at Mauna Loa for the mid-baseline of 1993 (Tans and Keeling data retrieved 2013). The solar radiation data for the time period was taken from the ECMWF's ERA-Interim reanalysis,[4] and was regridded to the E-OBS data resolution using the area weighted average. The planting window was set according to the dataset produced by Sacks et al. (2010). The soil hydrological values of saturated volume, lower limit volume, and drained upper limit were taken from the WISE Soil Database for Crop Simulation Models version 1.1 (Romero et al. 2012).

The YGP value was calibrated using increments of 0.01. The remaining GLAM parameters were set to their default values with the exception of transpiration efficiency (TE) and the maximum value of normalized TE (TEN_MAX). These parameters can have a significant impact on simulated yield. The value of TE was set to 5.45 pa and TEN_MAX was set to 6.0. These values were taken from Tallec et al. (2013), and are more realistic for the temperate region of this study than the default values. Unlike the statistical model, GLAM simulates processes such as planting and emergence, so the start of simulation was set to April (Birch et al. 2003). Initial tests of this experimental setup showed that GLAM was consistent with the maize development periods for this region as described by Sacks et al. (2010).

2.3 Simulating errors in climate data

In order to compare the effect of systematic climate data errors on these two types of crop models, we deconstruct the temperature datasets described in Sections 2.1 and 2.2 into terms that represent different time scales, and then perturb these terms to alter the mean and variance of the data at different temporal scales. As mentioned in Section 2.2, the statistical model uses temperature and weather data for the June-August season, while GLAM uses April-November. To ensure that both models were presented with identical input perturbations, the transformations described below were applied for the full year January-December.

Each time series was deconstructed as follows. Let $z(y, m, d)$ be the observed value of a given temperature value on day d in month m of year y. The complete 1980–2007 time series was then deconstructed as:

$$z(y, m, d) = \mu + \alpha_y + \beta_m + \gamma_{ym} + \delta_{ymd} \qquad (2)$$

where

$$\begin{aligned}
\mu &= z(\cdot, \cdot, \cdot), \\
\alpha_y &= z(y, \cdot, \cdot) - \mu, \\
\beta_m &= z(\cdot, m, \cdot) - \mu, \\
\gamma_{ym} &= z(y, m, \cdot) - (\mu + \alpha_y + \beta_m), \\
\delta_{ymd} &= z(y, m, d) - (\mu + \alpha_y + \beta_m + \gamma_{ym}).
\end{aligned}$$

In this notation the symbol $\cdot$ indicates a mean over the missing index, so for example:

$$z(y, m, \cdot) = \frac{1}{30} \sum_{d=1}^{30} z(y, m, d).$$

In the above deconstruction, μ is the overall mean, α_y is the average deviation for year y, β_m is the deviation for month m averaged over all years (so $\beta_1 \dots \beta_{12}$ represents the mean seasonal cycle), γ_{ym} is the year-dependent deviation from the mean seasonal cycle, and δ_{ymd} is the daily deviation from the monthly mean. Figure 2 illustrates the components of this deconstruction for a sample of minimum daily temperatures.

By adjusting each of the terms in the deconstruction, we can perturb the mean and variance components of the input time series to get a new time series z^*:

$$z^*(y, m, d; \theta) = \mu_\theta \mu + \alpha_\theta \alpha_y + \beta_\theta \beta_m + \gamma_\theta \gamma_{ym} + \delta_\theta \delta_{ymd}, \tag{3}$$
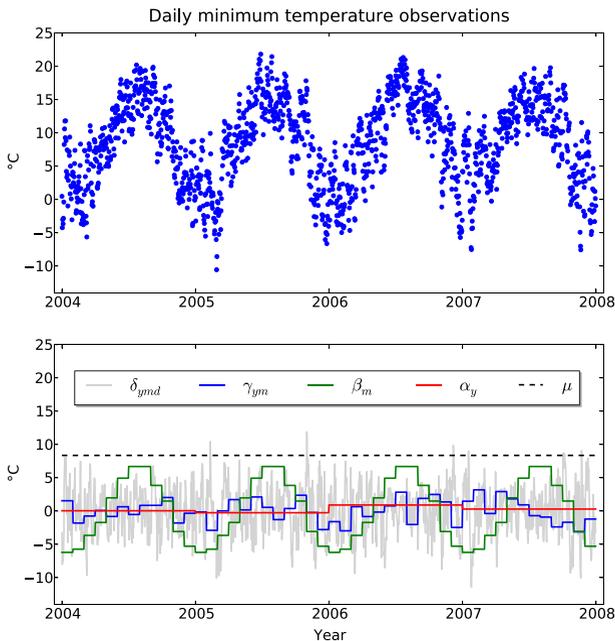
Fig. 2 Observed daily minimum temperatures at E-OBS location 44.625 degrees latitude, 0.625 degrees longitude, for the period 2004 – 2007 (*top panel*). The bottom panel shows the deconstruction components calculated for this time series. The location and time period were chosen for illustrative purposes

for some parameters $\theta = (\mu_\theta, \alpha_\theta, \beta_\theta, \gamma_\theta, \delta_\theta)$, i.e., each of the parameters in θ are multiplicative adjustments to their respective terms in the deconstruction. The original time series is recovered by setting $\theta = (1, 1, 1, 1, 1)$. Figure 3 shows examples of these adjustments made to the time series of Fig. 2.

The precipitation data P was perturbed using a different scheme, so that (1) transformed datasets could not contain negative values and (2) the pattern of days with no precipitation was retained, i.e., just the intensities of days with precipitation are perturbed. In this scheme the logarithm of the monthly means was used, and daily fluctuations are dealt with separately. That is, we let $z(y, m) = log[P(y, m, \cdot)]$ and deconstruct as

$$z(y, m) = \mu + \alpha_y + \beta_m + \gamma_{ym}, \tag{4}$$
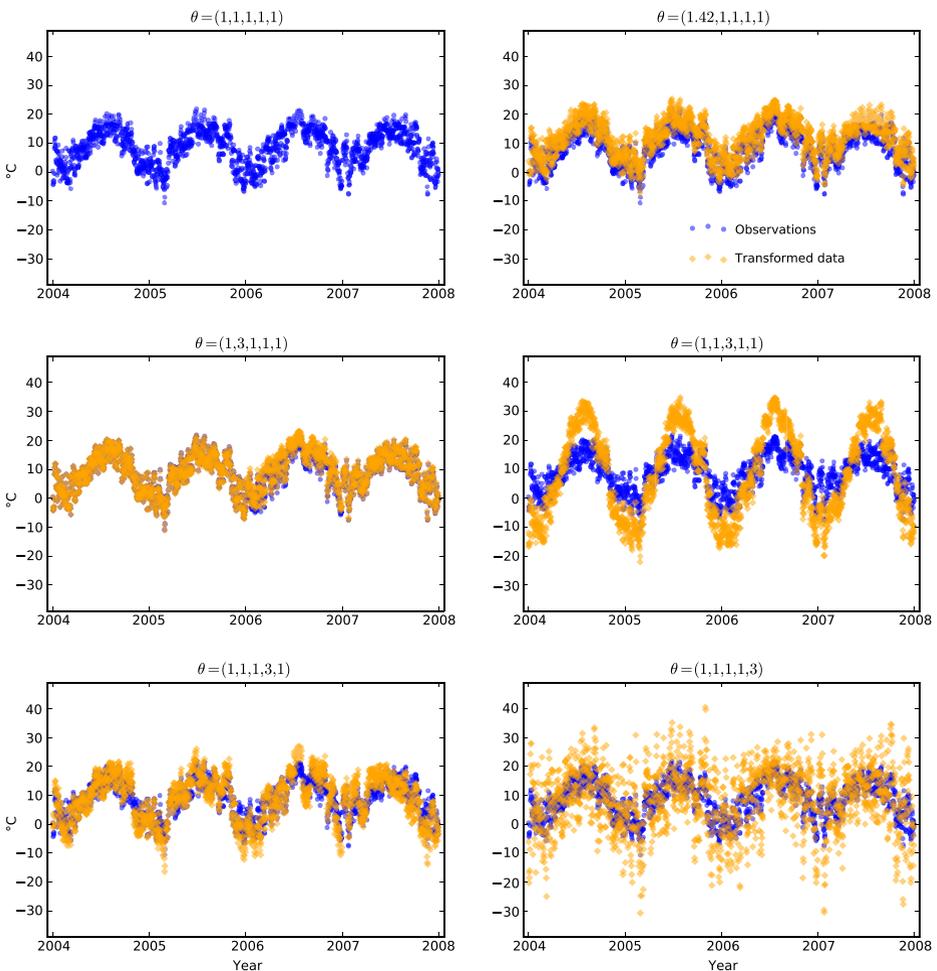
Fig. 3 The effect of different values of θ on the observations described in Fig. 2. The θ values illustrated are the maximum values used in this study. The original data is shown as blue circles, while the time series resulting from the application of each transformation is shown as orange diamonds

where

$$\mu = z(\cdot, \cdot),$$
$$\alpha_y = z(y, \cdot) - \mu,$$
$$\beta_m = z(\cdot, m) - \mu,$$
$$\gamma_{ym} = z(y, m) - (\mu + \alpha_y + \beta_m).$$

We then apply the perturbation scheme as described above with the θ parameters, but omit the δ terms to get $z^*(y, m)$. A transformed daily precipitation time series P^* is recovered by setting

$$P^*(y, m, d) = \zeta_{ymd} exp[z^*(y, m)], \tag{5}$$

where

$$\zeta_{ymd} = \frac{P(y, m, d)}{P(y, m, \cdot)}.$$

1mm of rainfall is added to the first day of months which have no precipitation, as otherwise $P^*(y, m, d)$ would be undefined (i.e., $P(y, m, \cdot) = 0$). This value is insignificant with respect to the simulation of maize in both crop models.

Values for θ were chosen to encompass a wide range of systematic errors that may occur in climate model datasets, to inform future studies whether there are potential issues when assessing if a dataset is fit for purpose. We chose values for μ_θ to be +/-45 % of observed μ. In this study's region for the reference period 1970–1999, the maximum difference in mean maximum summer temperature found between E-OBS observations and the CMIP3 ensemble (Meehl et al. 2007) was 9 °C, which is 37 % of the observed mean (Hawkins et al. 2013b). Note that an analysis of the more recent CMIP5 climate model ensemble (Taylor et al. 2012) indicates that its range of ensemble spread is not reduced compared to CMIP3 (Knutti and Sedláček 2013). It is difficult to obtain systematic errors in the standard deviation of climate models at a range of temporal scales. Climate models have the potential to have both higher and lower variance than observations depending on the selected scale. Thus, to ensure we evaluated the effect of a wide range of systematic errors, we compared crop model performance with z^* datasets ranging from no variance through to 3x the respective values of E-OBS for α_θ, β_θ, γ_θ and δ_θ). In order to provide a set of reference crop model sensitivities that directly compare the relative importance of temperature and precipitation errors, we transform precipitation with the same θ values as for temperature. Note that the aim is not to directly equate numeric changes made to the temperature and precipitation timeseries, but rather to compare the effects of changes in relative errors of their statistical components.

Each component of the deconstruction was tested independently, e.g., the effect of altering μ was tested for values of μ_θ where $\theta = (\mu_\theta, 1, 1, 1, 1)$). The effect of these errors were assessed in terms of $\Delta RMSE$ and $\Delta CCOEF$ (defined below). $\Delta RMSE$ was chosen to measure the accuracy of predictions, but since this metric heavily penalizes models that correctly capture the weather / yield relationship but incorrectly predict mean yield,[5] $\Delta RMSE$ was assessed in tandem with $\Delta CCOEF$, which measures the correlation between observed and simulated yields. Each crop model was calibrated using weather and crop data for the

---

[5]For example, a model with a flat response close to the mean can have a lower $\Delta RMSE$ than a model that correctly predicts yield response but with an incorrect mean.

period 1980–2002, and $\Delta RMSE$ and $\Delta CCOEF$ were defined as follows for the period 2003–2007. Let $RMSE_{baseline}$ denote the RMSE of simulated yield and observed yield, and $RMSE_{transformed}$ denote the RMSE of simulated yield and observed yield after the transformation has been applied. Similarly, let $CCOEF_{baseline}$ denote the correlation coefficient (CCOEF) of simulated yield and observed yield, and $CCOEF_{transformed}$ denote the CCOEF of simulated yield and observed yield after the transformation has been applied. Then:

$$\Delta RMSE = \frac{RMSE_{transformed} - RMSE_{baseline}}{RMSE_{baseline}} * 100$$

$$\Delta CCOEF = CCOEF_{transformed} - CCOEF_{baseline}$$

$\Delta RMSE$ and $\Delta CCOEF$ were calculated for each E-OBS grid cell using $STAT_{nonlinear}$, $STAT_{linear}$, $GLAM_{rfd}$ and $GLAM_{irr}$, and each assessment was repeated using country, regional and departmental scale yield calibration data.

2.4 Overview of study design

In summary, the $RMSE_{baseline}$ and $CCOEF_{baseline}$ of crop model simulations was measured for each of the 1,035 E-OBS grid locations, using:

–   4 crop model types ($STAT_{nonlinear}$, $STAT_{linear}$, $GLAM_{irr}$, $GLAM_{rfd}$), and
–   3 yield data sources (country, regional, and departmental scale observations).

Then, the above runs were repeated, measuring $\Delta RMSE$ and $\Delta CCOEF$ for all combinations of the following climatic variations:

–   2 perturbed weather variables (temperature and precipitation),
–   5 altered weather variable components (μ_θ, α_θ, β_θ, γ_θ, δ_θ), and
–   6 values for each weather variable component (see Figs. 6 and 7).

3 Results and discussion

Differences in the response to calibration and weather errors were found between the statistical crop models and the process-based crop models. Also, some key data features were critically important to both model types. These results are discussed below.

3.1 The importance of harvest area

Figure 4 shows the RMSE and CCOEF of each of the model types ($STAT_{linear}$, $STAT_{nonlinear}$, $GLAM_{irr}$ and $GLAM_{rfd}$) for each E-OBS grid cell. This data is plotted against the harvest area of the respective grid cell (x-axis) and separated by the scale of the yield calibration data. This figure shows a clear relationship between harvest area and model skill. Regions with low reported harvest areas are relatively more likely to give poor results with regional scale crop models when compared to regions with higher harvest areas. This relationship is robust across both structurally distinct crop model types, which are based on very different maize development and calibration assumptions. The effect is also robust across all three scales of the yield observations, which each have distinct statistical variance.
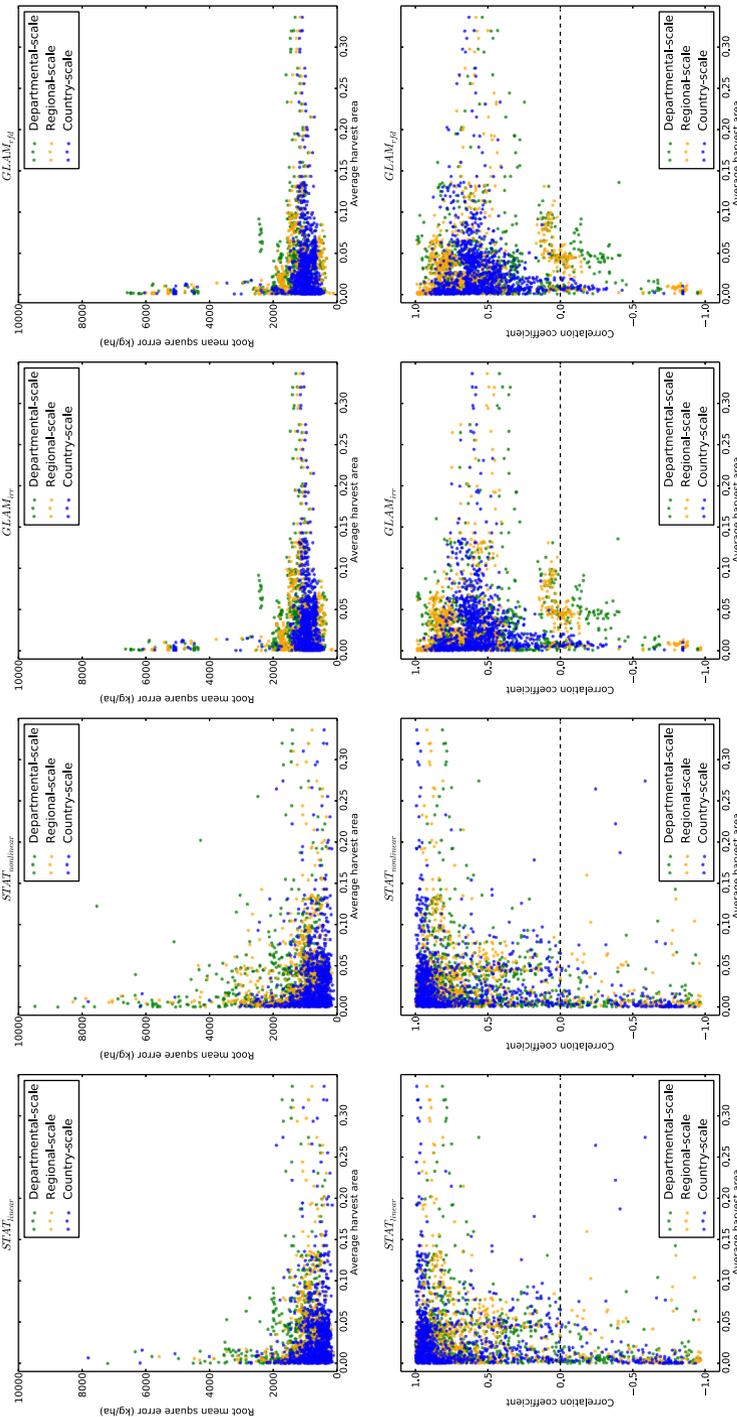
Fig. 4 Aggregated harvest area of each E-OBS grid cell plotted against model skill. The *top row* shows root mean square error, while the *bottom row* plots correlation coefficient. The skill of $STAT_{linear}$, $STAT_{nonlinear}$, $GLAM_{irr}$, $GLAM_{irr}$ and $GLAM_{rfd}$ are shown in columns 1, 2, 3 and 4 respectively. Model runs were calibrated with country scale (*blue*), regional scale (*orange*), and departmental scale (*green*) yield observations. For all models, increasing harvest area results in more consistent performance. Interestingly, at high harvest areas country scale calibration data tended to result in slightly lower RMSE and higher correlation coefficients than regional scale data, and regional scale data often resulted in slightly better performance than departmental data. This difference between the spatial scale of calibration data is most pronounced in $GLAM_{irr}$ and $GLAM_{rfd}$ at low harvest areas

These crop models rely on a signal existing between the model-relevant weather statistics and the model-relevant maize yield statistics. The strength of this signal in turn relies on the statistics of *both* the crop and weather datasets to accurately reflect conditions at the scale the crop model is being run. The lower the harvest area of a given grid cell, the more likely it is that reported kg ha^{-1} crop yields are the result of local sub-grid conditions not reflected in the interpolated grid cell weather. Thus the increasing likelihood of poor model performance in grid cells with low maize harvest area indicates an increasing mismatch between grid cell weather statistics and local crop yield observations.

The spatial scale of yield calibration data also had an effect on crop model skill. At any given value of harvest area, higher resolution yield data generally results in poorer model skill. This effect is particularly evident for GLAM at harvest areas < 10 %, where distinct differences in model skill can be seen depending on the spatial scale of the yield calibration data (see the CCOEF panels for $GLAM_{irr}$ and $GLAM_{rfd}$ in Fig. 4). This finding is contrary to the expectation that aggregation error decreases as spatial resolution increases, since the weather and the yield data are more closely matched to those experienced in reality (Hansen and Jones 2000). If model skill only relied on the spatial compatibility of weather and crop observations, we would expect departmental scale calibration data to outperform country scale data, since departments are still larger than the E-OBS grid cells. A possible explanation is that location-specific variations, such as management practices and weather conditions, cancel out over large areas, thus resulting in a stronger weather/yield relationship in the aggregated data. This finding appears to question the value of high resolution yield data for improving the skill of crop models. A focus on accuracy is likely to be of greater value. However, since these models were not optimized for local conditions, further work is clearly needed before definitive recommendations can be made.

3.2 The effect of simulated climate errors

The results of the simulations described in Section 2.3 are summarized in Figs. 6 and 7. Figure 6 shows the mean model skill across grid cells, while Fig. 7 shows the standard deviation. These results are restricted to grid cells with a harvest area >= 20 %, to reduce the impact of crop model inconsistency at low harvest areas described in Section 3.1. These grid cells are shown in Fig. 5. $STAT_{linear}$ and $STAT_{nonlinear}$ were similar in their responses to simulated errors. This was also the case for $GLAM_{rfd}$ and $GLAM_{irr}$. Here we only show the responses of $STAT_{nonlinear}$ and $GLAM_{rfd}$.

Both the statistical model and GLAM were significantly more influenced by transformations applied to temperature than those applied to precipitation. The fact that both model types exhibited this effect indicates that this is a model-independent result. Previous studies have also indicated a stronger maize yield response to temperature than to precipitation (Lobell et al. 2013; Hawkins et al. 2013a). This result can in part be explained by the method used to transform precipitation data. As described in Section 2.3, precipitation data was perturbed differently to temperature, in order to retain the pattern of days with rainfall, and to disallow negative values. By only adjusting the intensity of observed rainfall events, the overall impact of precipitation transformations can be less than those of temperature. Reductions in mean precipitation by up to 42 % did not significantly impact the skill of either model type – this is not a water-limited scenario. The statistical model's performance was affected by precipitation
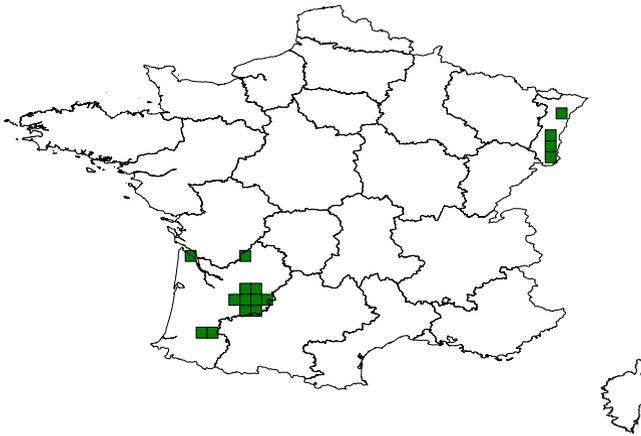
Fig. 5 E-OBS grid cells where the aggregated harvest area is $>= 20$ %

changes in average yearly deviation, and year dependent deviations from the seasonal cycle.

The responses to temperature transformations differed significantly between GLAM and the statistical models. *GLAM* was predominantly affected by transformations in climatic mean (μ) and monthly deviations (β_m) with low and high values of μ_θ and β_θ, respectively. Transformations resulting in over-estimation typically resulted in increases in GLAM $\Delta RMSE$, for all scales of yield observations (country, regional and departmental). An exception is the case for increases in the daily deviations from the monthly mean (δ_{ymd}) when country level crop yield data is used. This yield dataset was by definition common to all the grid cells analyzed, and exhibited the lowest standard deviation of the yield datasets (Fig. 1).

The statistical model was predominantly influenced by reductions in overall mean (μ), monthly deviations (β_m and γ_{ym}) and daily deviations from the monthly mean (δ_{ymd}). Increases in transformations also impacted the performance of the statistical model, but this was highly dependent on the scale of crop yield observations used. Increased overall mean increased statistical model RMSE, but only in the case of country scale yield observations. Higher resolution yield observations resulted in the statistical model exhibiting no negative change in skill for increases in overall mean. Positive changes to the average yearly deviation (α_y) resulted in greater loss of model skill as crop yield resolution increased. Positive changes in monthly deviations and daily deviations in the monthly mean also resulted in some loss of model skill, but again where country level yield observations were used. Overall, the statistical model was significantly more sensitive than GLAM to the spatial scale of the yield calibration data.

The statistical model generally exhibited greater variance than GLAM across the grid cells analyzed (Fig. 7). Unlike the statistical model, GLAM simulates processes in addition to statistical interactions between temperature, precipitation and crop yield, making it less susceptible to these transformations. However, large reductions in the overall mean (μ) and monthly deviations (β_m) of temperature did result in significant variation in GLAM's response.

Care must be taken in drawing conclusions about the relative quality of these two models using these results. Resistance or sensitivity to transformations is not an indication of model quality, but rather a pointer to what data errors should be of concern to impacts modellers. Importantly, the relative effects of (1) individual weather data transformations, and (2) the scale of yield calibration data, differs depending on model type.

3.3 Comparison of statistical and process-based models

Statistical models can test for relatively simple relationships and extrapolate based on observed relationships. Process-based models contain more assumptions (i.e. physiological processes and crop-climate relationships). The statistical model is directly influenced by the loss of statistically significant information in temperature, precipitation and yield data, while the process-based model is affected by altered process interactions resulting from such errors. The main difference between the two models used in this study is that the statistical models are designed to be entirely fit to data, whereas GLAM is more constrained in its behaviour by the processes that comprise the model. For either type of model, errors in relevant input data characteristics will result in the model failing when used out of sample. The degree to which such errors are relevant to each model is not clear a priori. Identifying the data characteristics that each model is sensitive to, and quantifying their effects, are pre-requisites for answering our research questions. Three observations are relevant:

First, the yield calibration data, in the absence of weather perturbations, affect the two models in different ways. GLAM RMSE is more consistent across the three different yield calibration datasets than the statistical models (Fig. 4). In contrast, for CCOEF at low harvest areas, the converse is true. This is not simply the result of more degrees of freedom resulting in lower RMSE, since the non-linear statistical model performs slightly worse than the linear model. Similarly, the greater number of degrees of freedom in GLAM do not automatically result in an improved CCOEF – here GLAM is constrained by the processes it simulates.

Second, susceptibility to weather errors vary according to the model, the variable and the timescale of perturbation. Overall, GLAM is less susceptible to errors in precipitation than the statistical model (Fig. 6). The calibration process in GLAM results in some precipitation bias being corrected, even though this is not the primary aim of calibration (Challinor et al. 2005). Temperature perturbations tend to produce larger errors. The sign and frequency of perturbations that produce the greatest impact tends to differ between the two models. In the statistical model, reducing the amplitude of temperature variation (α_θ, β_θ, γ_θ and $\delta_\theta < 1$) produces the largest errors. In GLAM, it is predominantly increases in amplitude that have the greatest impact on skill. GLAM is particularly prone to errors in mean temperature (μ_θ) and the magnitude of the seasonal cycle of temperature (β_θ).

Third, the spatial coherence of model response across grid cells differs between GLAM and the statistical models (Fig. 7). Errors in GLAM in response to weather perturbations are generally more spatially systematic (across grid cells) than when those same perturbations are introduced to the statistical models. This result reflects the fact that the spatial calibration of GLAM is restricted to one single parameter, whereas the statistical models have more degrees of freedom across space. This suggests that weather biases may be easier to correct in a process-based model than a statistical model, and agrees with previous work calling for greater attention to measurement error in statistical crop models (Lobell 2013).
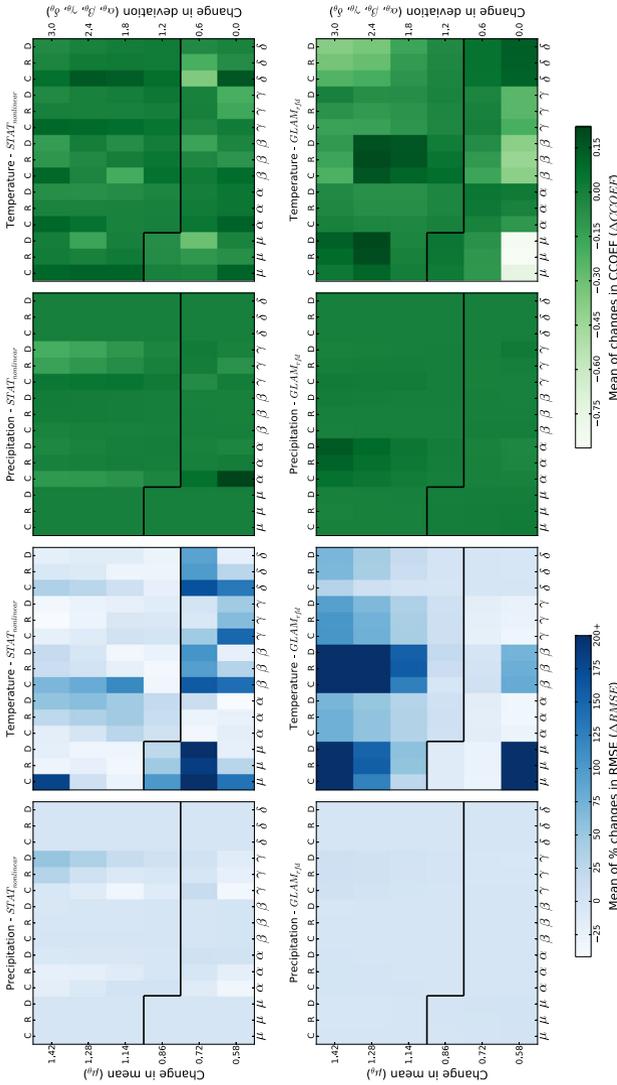
Fig. 6 The mean effect of weather transformations on crop model skill for grid cells with >= 20 % harvest area. The results of transformations in the overall mean (μ_θ), average yearly deviation (α_θ), average monthly deviation (β_θ), year-dependent deviation from the mean seasonal cycle (γ_θ) and daily deviation from the monthly mean (δ_θ) are separated along the x-axes. The '$_\theta$' subscripts are omitted from the figure labels for clarity. The scale of the yield observations used – country (C), regional (R) or departmental (D) – are indicated at the top of each panel. Values of μ_θ are shown on the left y-axis, and values of α_θ, β_θ, γ_θ and δ_θ are shown on the right y-axis. Positive and negative transformations are separated by *black lines*. Mean $\Delta RMSE$ values are shown in *blue* in the left four panels, and mean $\Delta CCOEF$ values are shown in *green* in the right four panels. Both $STAT_{nonlinear}$ (*top row*) and $GLAM_{rfd}$ (*bottom row*) had a significantly stronger response to transformations in temperature (*second and fourth columns*) than in precipitation (*first and third columns*). For $GLAM_{rfd}$ temperature transformations, μ_θ and β_θ had the largest effect, at both low and high values, while positive transformations generally resulted in increases in $\Delta RMSE$. $STAT_{nonlinear}$ temperature change responses were significantly more sensitive to the scale of yield observations compared to $GLAM_{rfd}$, and were generally more sensitive to negative transformations. For $STAT_{nonlinear}$, changes in μ_θ, β_θ and δ_θ generally had more of an effect with country scale yield data, and α_θ had more of an effect with high resolution yield data. $STAT_{nonlinear}$ was also somewhat sensitive to precipitation changes in α_θ and γ_θ. $\Delta RMSE > 200$ was considered a model failure, so plotted values are capped at this value
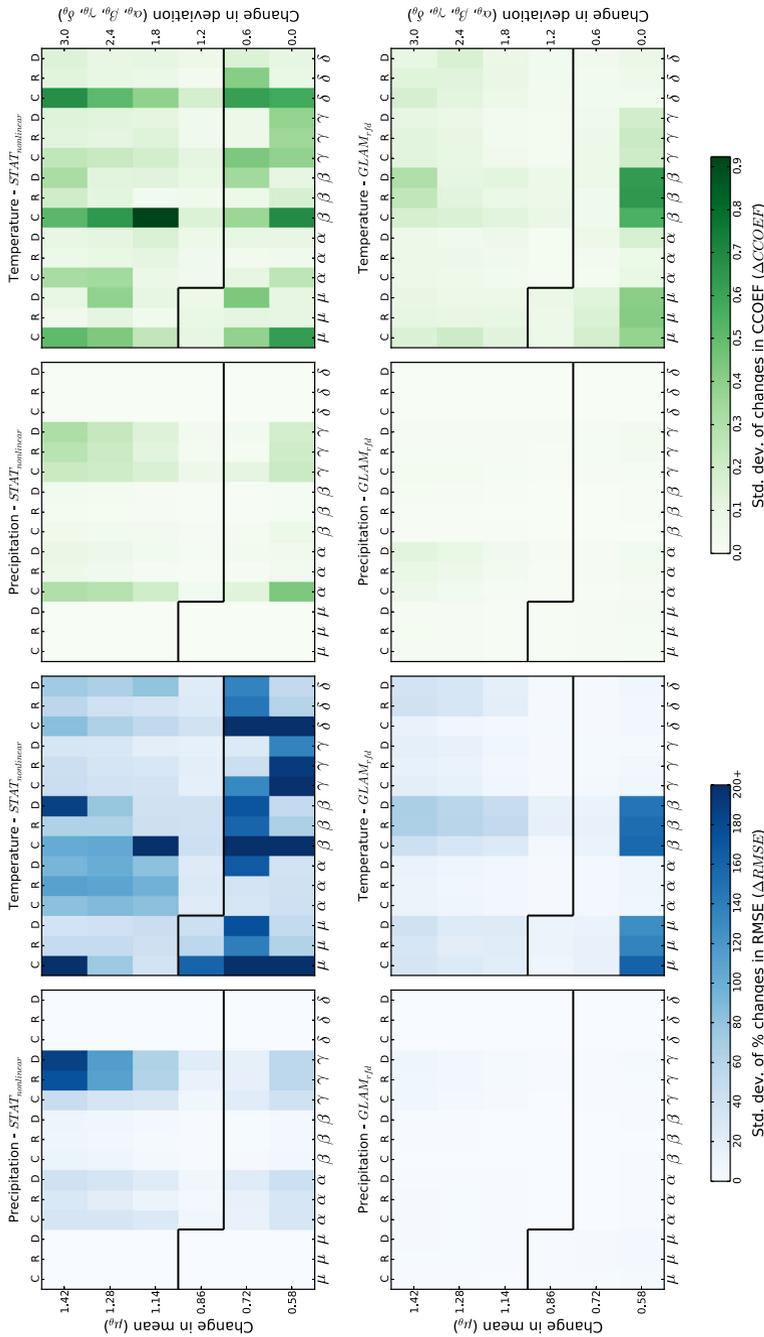
Fig. 7 The standard deviation of the effect of weather transformations on model skill for grid cells with >= 20 % harvest area, using the same layout and labelling scheme as Fig. 6. The sensitivity of $STAT_{nonlinear}$ to precipitation changes according to α_θ and γ_θ is clearly visible. When compared against Fig. 6, it can be seen that $GLAM_{rfd}$'s responses to transformations are generally less varied than those of the statistical model in cases where skill is significantly reduced (with the exception of some reductions in overall mean (μ_θ) and average monthly deviation (β_θ) at regional and departmental yield scales). $\Delta RMSE > 200$ was considered a model failure, so plotted values are capped at this value

4 Conclusions

Care must be used in interpreting these results outside the context of this study. A single crop was analyzed in a single temperate country, and models were not optimized for local conditions beyond the use of their automatic calibration routines. However, particularly where similar results were found using different model structures, some general comments can be made.

The process-based models and the statistical models used in this study were found to be susceptible to different types of input data error. These two model types make different choices of model design and calibration, so this result is not surprising. The contribution here is in the quantification of the effects that different errors have on these model types. The next step is to identify to what extent widely used crop model inputs, such as global climate models, exhibit such errors.

For both regional scale crop model types, and for all three spatial scales of yield calibration data, we found that model skill is most reliable where growing area is above 10-15 %. Thus information on area harvested would appear to be a priority for data collection efforts.

The GLAM process-based model can compensate for some loss in weather information and was found to be resilient to differences in yield data resolution, while mean statistical model response was resilient to overestimation errors. These differing responses to input data error raise the intriguing possibility of using process-based and statistical models in tandem to improve crop yield predictions.

Biases and errors in temperature, precipitation and yield input data influence the results of crop models. Consequently, the management advice given on the basis of these models is subject to influence by these errors. Understanding the detailed impact of different error types helps improve crop yield projections, and can guide efforts to improve models and datasets. The methodology introduced by this study can be applied to a range of impact modelling scenarios that utilize daily weather data. Extending this study to assess the potential impact of data errors on different model types, crops, and locations, would provide modelers with key information for designing and evaluating impacts studies.

References

Asseng S et al (2013) Uncertainty in simulating wheat yields under climate change. Nat Clim Chang 3:827–832. doi:10.1038/nclimate1916

Berg A, Sultan B, de Noblet-Ducoudré N (2010) What are the dominant features of rainfall leading to realistic large-scale crop yield simulations in West Africa? Geophys Res Lett 37:L05,405

Birch C, Vos J, van der Putten P (2003) Plant development and leaf area production in contrasting cultivars of maize grown in a cool temperate environment in the field. Europ J Agronomy 19:173–188

Challinor A, Wheeler T (2008) Use of a crop model ensemble to quantify co2 stimulation of water-stressed and well-watered crops. Agric For Meteorol 148:1062–1077

Challinor A, Wheeler T, Craufurd P, Slingo J, Grimes D (2004) Design and optimisation of a large-area process-based model for annual crops. Agr Forest Meteorol 124:99–120. doi:10.1016/j.agrformet.2004.01.002

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

Challinor A, Watson J, Lobell D, Howden M, Smith D, Chhetri N (2014) A meta-analysis of crop yield under climate change and adaptation. Nat Clim Chang 4:287–291. doi:10.1038/nclimate2153

de Wit A, van Diepen C (2007) Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. Agric For Meteorol 146:38–56

Estes LD, Beukes H, Bradley BA, Debats SR, Oppenheimer M, Ruane AC, Schulze R, Tadross M (2013) Projected climate impacts to South African maize and wheat production in 2055: a comparison of empirical and mechanistic modeling approaches. Glob Chang Bio 19:3762–3774. doi:10.1111/gcb.12325

Greatrex H (2012) The application of seasonal rainfall forecasts and satellite rainfall estimates to seasonal crop yield forecasting for Africa. PhD thesis, University of Reading

Hansen J, Jones J (2000) Scaling-up crop models for climate variability applications. Agric Syst 65:43–72

Hawkins E, Fricker T, Challinor A, Ferro C, Ho C, Osborne T (2013a) Increasing influence of heat stress on french maize yields from the 1960s to the 2030s. Glob Chang Biol 19:937–947. doi:10.1111/gcb.12069

Hawkins E, Osborne TM, Ho CK, Challinor AJ (2013b) Calibration and bias correction of climate projections for crop modelling: An idealised case study over Europe. Agric For Meteorol 170:19–31. doi:10.1016/j.agrformet.2012.04.007

Haylock M, Hofstra N, Klein Tank A, Klok E, Jones P, New M (2008) A European daily high-resolution gridded dataset of surface temperature and precipitation. J Geophys Res Atmos 113:D20,119. doi:10.1029/2008JD10201

Knutti R, Sedláček J (2013) Robustness and uncertainties in the new CMIP5 climate model projections. Nat Clim Chang 3:369–373. doi:10.1038/nclimate1716

Koehler AK, Challinor AJ, Hawkins E, Asseng S (2013) Influences of increasing temperature on Indian wheat: quantifying limits to predictability. Environ Res Lett 8:034,016. doi:10.1088/1748-9326/8/3/034016

Lobell DB (2013) Errors in climate datasets and their effects on statistical crop models. Agric For Meteorol 170:58–66. doi:10.1016/j.agrformet.2012.05.013

Lobell DB, Asner GP (2003) Climate and management contributions to recent trends in US agricultural yields. Science 299:1032

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. Nat Clim Chang 3:497–501. doi:10.1038/nclimate1832

Meehl G, Covey C, Delworth T, Latif M, McAvaney B, Mitchell J, Stouffer R, Taylor K (2007) The WCRP CMIP3 multi-model dataset: A new era in climate change research. Bull Am Meteorol Soc 88:1383–1394. doi:10.1175/BAMS-88-9-1383

Monfreda C, Ramankutty N, Foley J (2008) Farming the planet. part 2: geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Glob Biogeochem Cycles 22:GB1022. doi:10.1029/2007GB002947

Ramirez-Villegas J, Challinor A, Thornton P, Jarvis A (2013) Implications of regional improvement in global climate models for agricultural impact research. Environ Res Lett 8:024,018. doi:10.1088/1748-9326/8/2/024018

Romero CC, Hoogenboom G, Baigorria GA, Koo J, Gijsman AJ, Wood S (2012) Reanalysis of a global soil database for crop and environmental modeling. Environ Model Softw 35:163–170. doi:10.1016/j.envsoft.2012.02.018

Sacks W, Deryng D, Foley J, Ramankutty N (2010) Crop planting dates: an analysis of global patterns. Glob Ecol Biogeogr 19:607–620. doi:10.1111/j.1466-8238.2010.00551.x

Tallec T, Béziat P, Jarosz N, Rivalland V, Ceschia E (2013) Crops water use efficiencies in temperate climate: comparison of stand, ecosystem and agronomical approaches. Agric For Meteorol 168:69–81

Tans P, Keeling R (data retrieved 2013) http://www.esrl.noaa.gov/gmd/ccgg/trends/. NOAA/ESRL and Scripps Institution of Oceanography

Taylor K, Stouffer R, Meehl G (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. doi:10.1175/BAMS-D-11-00094.1

Watson J, Challinor A (2013) The relative importance of rainfall, temperature and yield data for a regional-scale crop model. Agric For Meteorol 170:47–57. doi:10.1016/j.agrformet.2012.08.001

White JW, Hoogenboom G, Kimball BA, Wall GW (2011) Methodologies for simulating impacts of climate change on crop production. Field Crop Res 124:357–368