Tellus (2005), 57A, 498–512
Printed in UK. All rights reserved

TELLUS

Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles

By A. J. CHALLINOR[1,2]*, J. M. SLINGO[1], T. R. WHEELER[2] and F. J. DOBLAS–REYES[3]
[1]*CGAM, Department of Meteorology, University of Reading, Reading RG6 6BB, UK;* [2]*Department of Agriculture, University of Reading, Reading RG6 6AT, UK;* [3]*European Centre for Medium-Range Weather Forecasts (ECMWF), Shinfield Park, Reading RG2 9AX, UK*

(Manuscript received 31 March 2004; in final form 28 December 2004)

ABSTRACT

Process-based integrated modelling of weather and crop yield over large areas is becoming an important research topic. The production of the DEMETER ensemble hindcasts of weather allows this work to be carried out in a probabilistic framework. In this study, ensembles of crop yield (groundnut, *Arachis hypogaea* L.) were produced for 10 2.5° × 2.5° grid cells in western India using the DEMETER ensembles and the general large-area model (GLAM) for annual crops.

Four key issues are addressed by this study. First, crop model calibration methods for use with weather ensemble data are assessed. Calibration using yield ensembles was more successful than calibration using reanalysis data (the European Centre for Medium-Range Weather Forecasts 40-yr reanalysis, ERA40). Secondly, the potential for probabilistic forecasting of crop failure is examined. The hindcasts show skill in the prediction of crop failure, with more severe failures being more predictable. Thirdly, the use of yield ensemble means to predict interannual variability in crop yield is examined and their skill assessed relative to baseline simulations using ERA40. The accuracy of multi-model yield ensemble means is equal to or greater than the accuracy using ERA40. Fourthly, the impact of two key uncertainties, sowing window and spatial scale, is briefly examined. The impact of uncertainty in the sowing window is greater with ERA40 than with the multi-model yield ensemble mean. Subgrid heterogeneity affects model accuracy: where correlations are low on the grid scale, they may be significantly positive on the subgrid scale.

The implications of the results of this study for yield forecasting on seasonal time-scales are as follows. (i) There is the potential for probabilistic forecasting of crop failure (defined by a threshold yield value); forecasting of yield terciles shows less potential. (ii) Any improvement in the skill of climate models has the potential to translate into improved deterministic yield prediction. (iii) Whilst model input uncertainties are important, uncertainty in the sowing window may not require specific modelling.

The implications of the results of this study for yield forecasting on multidecadal (climate change) time-scales are as follows. (i) The skill in the ensemble mean suggests that the perturbation, within uncertainty bounds, of crop and climate parameters, could potentially average out some of the errors associated with mean yield prediction. (ii) For a given technology trend, decadal fluctuations in the yield-gap parameter used by GLAM may be relatively small, implying some predictability on those time-scales.

1. Introduction

Numerical crop growth models are increasingly being used to simulate yield over large areas. Seasonal predictability can inform early warning systems (e.g. Rijks et al., 2003) whilst multidecadal time-scales can inform climate change impacts assessments (e.g. Fischer et al., 2002). Most, if not all, studies of yield predictability to date treat crop yield simulation deterministically (i.e. one set of inputs is used to derive one set of outputs). However, climate on seasonal time-scales is inherently unpredictable. Recent progress in the use of multi-model weather ensembles has been achieved through the DEMETER project (Palmer et al., 2004). Such weather ensembles provide an excellent opportunity to explore crop yield predictability using probabilistic methods (see also Cantelaube and Terres, 2005). This is the objective of this paper, which forms part of the methodology for the development of a combined seasonal weather and crop productivity forecasting system outlined by Challinor et al. (2003). This study uses seasonal time-scales although the conclusions

*Corresponding author.
e-mail: ajc@met.rdg.ac.uk

reached will have relevance for studies of longer time-scales, and climate change, because inherent unpredictability and uncertainties have to be estimated for these time-scales also.

Crop modelling approaches are either empirical (e.g. Camberlin and Diop, 1999; Hsieh et al., 1999; Landau et al., 2000) or process-based (e.g. Southworth et al., 2000; Jagtap and Jones, 2002), although in practice the distinction between these can be blurred because empirical regressions are often used to describe processes (e.g. Brooks et al., 2001). The process-based approach has the advantage of potentially capturing changes in the nature of the weather–yield relationship due to changes in climate, such as intraseasonal variability and increased CO_2 levels. However, there is often a high input data requirement for process-based models. For empirical approaches, the converse tends to be true. Challinor et al. (2004) have developed a relatively simple, large-area, process-based model [the general large-area model for annual crops (GLAM)] which aims to combine the advantages of these two approaches. The model simulated the interannual variability in groundnut yield over the Gujarat region of India well, when driven with either observed gridded weather data (Challinor et al., 2004) or the European Centre for Medium-Range Weather Forecasts (ECMWF) 40-yr reanalysis (ERA40; Challinor et al., 2005). GLAM is used in this study and is described briefly in Section 2.2.

The development and assessment of probabilistic yield forecast methods using GCMs raises a number of issues. First, the skill of the GCMs in simulating weather and climate needs to be sufficient. The calibration of the crop model also needs to be sufficiently accurate. Because the GLAM crop model has already been tested in deterministic studies using observed data and reanalysis, crop model skill is a secondary issue in this study. Secondly, given a crop yield ensemble, there may be useful information in the mean, in the spread, or in both. Also, skill may emerge on one or more spatial scales. Hence, the spatial scale for crop model output needs to be determined according to the spatial scale(s) at which the model can skilfully simulate observed yields. The third issue is that of uncertainty. The spatial scale on which simulations are carried out is one source of uncertainty: when a single set of crop model parameters is used to represent crop growth over a large area, subgrid heterogeneity may result in poor agreement between simulated and measured yields (e.g. Hansen and Jones, 2000). There will also be uncertainty associated with crop model inputs, such as soil type and planting date. It is important to understand the impact of these uncertainties on simulation skill.

This study aims to develop methods for the use of weather ensembles with crop models such as GLAM. The chosen crop is groundnut (peanut; *Arachis hypogaea* L.) as this is the crop for which extended records of observed yield are available. The geographical region chosen is in western India, and includes all of Gujarat, the region for which skill has been most effectively demonstrated to date using GLAM. However, the methods used in this study are not location or crop specific.

2. Method: formulation of crop yield hindcasts

2.1. Weather data

The weather data used as input to the crop model in this study are the DEMETER ensembles (Palmer et al., 2004). An ensemble consists of a number of sets of weather variables such as temperature, radiation, rainfall, humidity, each of which provides a complete, and in theory equally probable, meteorological description of the atmosphere. The term 'ensemble member' refers to one of these sets; it is a single realization of weather produced by a GCM. Output from seven GCMs (denoted here as cnrm, crfc, lody, scnr, scwn, smpi and ukmo) each with nine ensemble members (hence 63 ensemble members in total) was used. For each GCM, the nine ensemble members are referred to as a single-model (i.e. single-GCM) ensemble. All 63 ensemble members collectively are referred to as the multi-model (multi-GCM) ensemble.

Each GCM was run four times for each year of the study period. The four start (initialization) dates were the first days of February, May, August and November. Each ensemble member is a six-month daily time series. This creates two possibilities for this study. (i) The use of the ensembles initialized in May for a three-month groundnut simulation period, followed by use of the ensembles initialized in August [August update (AUP)]. These simulations involve a step change on 1 August to a time series chosen from the nine new ensemble members using ensemble identification number; this is essentially an arbitrary choice. (ii) The use of the simulation initialized in May for the whole of the growing season [no update (NUP)].

Two sets of input ensemble weather data have been used: the first is the raw DEMETER ensembles [original data (ORI)] and the second is a bias-corrected set of data (BIC). ERA40 was used for this bias-correction. It was also used to drive the crop model directly, producing benchmark deterministic simulations. The ERA40 simulations here differ from those of Challinor et al. (2005) in that maximum and minimum daily temperatures are used as inputs to the crop model in the current study, whereas mean daily temperature and vapour pressure deficit were used for the previous study. A further difference is that the previous study used the ERA40 grid ($0.5^\circ \times 0.5^\circ$) and the current study uses data interpolated to the DEMETER grid using the Meteorological Archive and Retrieval System interpolation tool.

Bias-correction was applied to daily values of maximum temperature, minimum temperature and precipitation for each GCM separately. First, an estimate of the seasonal cycle at each grid point was obtained. The seasonal cycle with daily resolution was computed by averaging, for a given start date and lead time, all the weather ensemble members and hindcasts available. This estimate was smoothed out by retaining the three first harmonics in a Fourier decomposition of the time series. The same method was used to estimate the seasonal cycle with the ERA40 data. The bias was defined as the difference between the GCM and the

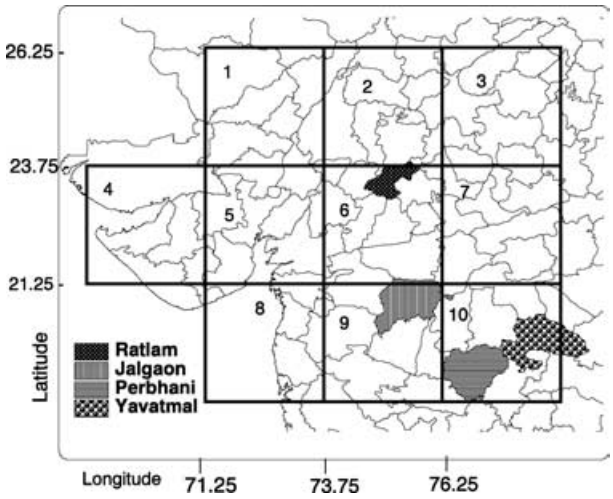


Fig. 1. Map of the crop model grid (with corresponding cell numbers) and districts. The four districts referred to in Section 3.4 are highlighted.

ERA40 seasonal cycles. Finally, bias-corrected hindcasts were computed as the difference between the hindcasts issued by the coupled GCMs minus the estimated bias. Negative precipitation values were removed under the constraint that the total precipitation of the hindcasts is equal to that of ERA40.

The study region is 10 grid cells in western India (Fig. 1). The study period, 1987–1998, is the period for which both input weather data and observed groundnut yield data (see Section 2.2) are available. The GCMs varied in their simulations of weather for this region over this period. For example, August rainfall for the ORI inputs showed a wide range of values: the difference between the two extreme-valued GCMs varied (spatially) between 90 and 170 mm. A similar analysis of mean temperature gives a range of 1.5° to 5°. Bias-correction greatly reduced these ranges, to 30–40 mm and 0°–0.5°.

2.2. *Crop modelling techniques*

The crop model used for this study is GLAM (Challinor et al., 2004). GLAM seeks to combine the benefits of empirical modelling (validity over large areas, low input data requirement) with the benefits of process-based modelling (capturing the impacts of subseasonal variability and retaining validity under unprecedented conditions, such as are likely under future climates).

Crop development is determined by accumulating daily mean values of temperature above a base temperature (thermal time) with development stages occurring at specific thermal times. The leaf area index (LAI) is modelled using a maximum growth rate modified by an indicator of water stress. LAI and solar radiation are used to calculate the energy-limited evapotranspiration. Actual transpiration is water-limited, and will depend on the available water as given by the soil/roots submodel. The ratio of actual to energy-limited evapotranspiration is the indicator of water stress. Use of a transpiration efficiency (which is a function of ambient vapour pressure deficit) then allows the calculation of biomass, which through a harvest index allows the calculation of yield.

The model, as it is used here, uses daily values of solar radiation, minimum and maximum temperature, and rainfall. Radiation is used to determine evapotranspirative demand and rainfall is used as the input to the uppermost soil layer. Maximum and minimum temperatures are averaged to produce daily mean temperature, and they are also used to calculate the vapour pressure deficit if those data are not available. GLAM has an intelligent sowing routine, which requires as input a sowing window. Sowing occurs on the first day on which the uppermost soil layer is moist enough, or at the end of the window (crisis-sowing) if this does not occur. The soil water model was initialized at the start of the sowing window with zero available soil moisture.

Of the impacts on yield due to factors other than weather (pests, diseases, management factors, etc., which act to reduce yields by an amount referred to as the yield gap) only two are modelled explicitly: planting date and soil type. The key soil attribute is the water storage capacity. This is simulated using a lower limit, drained upper limit and saturated limit. Other soil influences are not simulated. All remaining influences on yield are modelled using a single yield-gap parameter (YGP), which acts to decrease the leaf area available for transpiration. This allows the model to focus on the impact of weather and climate on the spatio-temporal variability of crop yield. The YGP may also be set to simulate zero yield gap (i.e. yield potential, which is limited only by water, radiation, humidity and temperature). However, it is observed yields with which model output is compared, and this necessitates the calibration of the YGP.

The YGP takes values between zero and unity, in steps of 0.05, and it is calibrated using observed yields. Hence, calibration is a form of mean bias-correction, which may incorporate the impact of biases additional to the yield gap, such as input data bias and crop model error. Yearly district-level groundnut yield data for calibration and evaluation were provided by the Socio-economic Policy Division of the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India, from yearly agricultural bulletins (Agricultural Situation in India, Department of Agriculture, Government of India). Challinor et al. (2003) provide further description, and some analyses, of these data. For the current study, calibration of the YGP uses either (i) ERA40 data (http://www.ecmwf.int/research/era/) prior to the study period (1966–1986) to determine a single value of the YGP, referred to as GCAL, or (ii) cross-validation using data within the study period (1987–1992 data are used to determine the YGP for 1993–1998, and vice versa), referred to as NCAL. The source of data for the NCAL calibration was the same source as for the relevant simulations: ERA40 data were used for ERA40 runs and weather ensemble data for ensemble runs. Two calibration methods were used for NCAL ensemble runs (Table 1). In the first, a single pair of YGP values was determined by using the

Table 1. Naming convention for the crop model configurations, together with a list of the simulation experiments performed. Each of the simulation experiments refers to a particular choice of configuration (calibration, input weather data, sowing window and bias-correction). Each simulation experiment was carried out using all 63 multi-model ensemble members individually, with the exception of WIA and ERA40 runs, as explained below

Runcode	Description
NUP	Weather data: May hindcast used throughout (no August hindcast update)
AUP	Weather data: May hindcast used for three months, then with August hindcast used
GCAL	Calibration of the YGP on 1966–1986 yield data using ERA40 input data
NCAL	Calibration of the YGP by cross-validation on 1987–1998 yield data (1987–1992 data used to determine the 1993–1998 YGP, and vice versa). All NCAL runs use the August hindcast update (AUP)
MMC	Calibration using the NCAL method on the multi-model yield ensemble mean
SMC	Calibration using the NCAL method on the respective single-model yield ensemble mean
BIC	Bias-correction of the input weather data to ERA40 has been performed
ORI	Bias-correction has not been performed (original input weather data)
DSW	Delayed sowing window
WIA	Weather inputs averaged: ensemble-averaged weather values are used as input
ERA40	ERA40 weather data are used as input
Run	**Comments**
GCAL–BIC–NUP	This was the only NUP run performed
GCAL–BIC–AUP	The control run: independent calibration with bias-corrected input weather data
GCAL–BIC–AUP–DSW	Control run with delayed sowing window (9 July–7 August)
GCAL–BIC–AUP–WIA	As control run, but with a single simulation using weather inputs averaged across the multi-model ensemble
NCAL–ORI–MMC	Calibration using yield ensemble means and yield data from the study period, using a single pair of YGP values for all GCMs
NCAL–ORI–SMC	Model calibration using yield ensemble means and yield data from the study period, using a GCM-specific pair of YGP values
GCAL–ORI–AUP	True hindcast: no 1987–1998 data used
NCAL–BIC–MMC	Full use of available 1987–1998 data
NCAL–BIC–SMC	–
GCAL–ERA40	Benchmark comparison run with independent calibration
NCAL–ERA40	Benchmark comparison run with calibration on current yields
GCAL–ERA40-DSW	Allows assessment of relative impact of delayed sowing window on control and benchmark runs

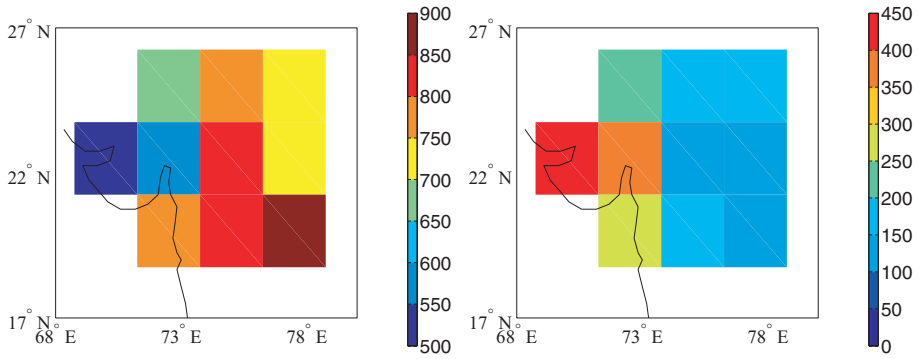
multi-model (i.e. multiGCM) yield ensemble mean [multi-model calibration (MMC)]. In the second method, each single-model (i.e. single-GCM) yield ensemble mean was used to determine a pair of YGP values for yield ensemble members from that GCM [single-model calibration (SMC)]. The YGPs resulting from this second method may include a component of weather data bias-correction for each individual GCM. In all cases the calibrated value of the YGP is that for which the root mean square error (RMSE) in yield is minimized.

The yield data used are either the district-level data (Fig. 1) or data that have been upscaled to the crop model grid using an area-weighted mean. Upscaling is carried out by assuming that the area under cultivation is spread evenly throughout each district. Yield often shows a monotonically increasing trend over time, which is attributable to improvements in management and crop variety. Hence, for this study, all yield data have been linearly detrended to 1987 levels. Figure 2 shows the mean and standard deviations of yields on the simulation grid.

Soil hydrological properties were derived from FAO/Unesco (1974) following Challinor et al. (2004). The input sowing window used (Reddy, 1988) varies geographically, with the earliest sowing across the region being the last day in May. The latest sowing window starts in the last week in July. All sowing windows last 30 d. Despite having observations of the sowing window, the planting date remains a considerable source of uncertainty. There is some evidence that choosing a sowing window start date that is later than the observed value in Gujarat produces more realistic simulations (Challinor et al., 2005). As a preliminary study of the impact of uncertainty in the sowing window, some simulations with a moderately delayed sowing window (starting on 9 July; denoted by DSW) across the whole region were carried out.

2.3. Hindcast simulation experiments

Hindcasts of crop yield were created by driving GLAM with individual DEMETER ensemble members. Ensemble mean yields,



(a) (b)

Fig. 2. Observed mean (a) and standard deviation (b) of (linearly) detrended groundnut yields (kg ha^{-1}) in India, for the period 1987–1998, on the DEMETER grid.

for either a single model (GCM) or the multi-model ensemble, were then created by averaging output yields. Two sets of hindcasts were carried out. The first of these used yield averaged over crop model grid cells for calibration and evaluation of the crop model (area-averaged geocode). This set of hindcasts consists of a number of configurations of the crop model, each performed across the whole study period and region. The configuration determines the choice of calibration (GCAL or NCAL; MMC or SMC; see Section 2.2), input weather data (NUP or AUP; see Section 2.1), sowing window (standard or delayed, DSW; see Section 2.2) and bias-correction method (BIC or ORI; see Section 2.1). Each configuration is performed on either ensemble average weather [weather inputs averaged (WIA)], ERA40 weather (denoted by ERA40) or individual ensemble members (all remaining simulations). Table 1 summarizes the configurations used in this study.

The second set of hindcasts used the district-level yield data, one district at a time, for calibration and evaluation [one district geocode (ODG)]. Hence, the ODG simulations sought to establish whether district-level yields could be simulated. Runs were performed with one of the districts representing the yield for the whole grid cell. The districts were chosen from grid cells where the interannual standard deviation of the area-averaged yield was not simulated well (low correlation coefficient) by ERA40. The GCAL–BIC–AUP configuration was used for these simulations.

Hindcast simulation experiments are aimed at beginning to address the issues highlighted at the end of Section 1. In particular they seek to address the following key questions.

(1) How is a probabilistic forecasting system best calibrated – is bias-correction of input weather data needed (BIC versus ORI)? Do estimates of the YGP need to be current or will estimates based on historical yields suffice (NCAL–ERA40 versus GCAL–ERA40)? Should calibration be carried out on yields from the same source/period as the yield data being simulated (NCAL versus GCAL multi-model ensemble runs)?

(2) How can skilful probabilistic forecasts of crop yield be formed from a set of yield ensembles? Is information based on a dichotomous analysis of crop failure more accurate than more highly resolved information such as indicators of high/medium/low yields? Are there any benefits specific to the multi-model, as opposed to the single-model approach? Do updated forecasts produce increased probabilistic skill (GCAL–BIC–AUP versus GCAL–BIC–NUP)?

(3) How skilful are the yield ensemble mean hindcasts when measured relative to the accuracy of the benchmark ERA40 simulations? Do updated forecasts produce increased deterministic skill (GCAL–BIC–AUP versus GCAL–BIC–NUP)? How does the multi-model yield ensemble mean perform relative to individual models?

(4) What are the impacts of two of the key uncertainties – sowing window and spatial scale – on the accuracy of the simulations? Preliminary analyses described in this paper examine: (i) whether uncertainty in the sowing window affects the accuracy of both the deterministic and probabilistic simulations equally [GCAL–BIC–AUP(–DSW) versus GCAL–ERA40(–DSW)]; (ii) whether, where there is little accuracy at the grid-scale, this can be attributed to heterogeneity within the grid cell. Specifically, is there accuracy on the subgrid scale (ODG)?

For all of the above questions, ensembles of yield are used in the analysis. However, for a deterministic simulation of crop yield, only a single set of weather inputs is necessary. Hence, a further simulation was carried out, using the multi-model ensemble mean weather variables (maximum and minimum temperature, solar radiation and rainfall) as input to GLAM. This simulation used the control run configuration (i.e. GCAL–BIC–AUP). It was aimed at one specific further issue which forms part of the third question above: for a deterministic simulation, is greater skill achieved when averaging at the input (weather) stage, or at the output (yield) stage (GCAL–BIC–AUP–WIA

versus GCAL–BIC–AUP)? Each of the questions above is discussed in turn in the four parts of Section 4.

2.4. Analysis methods

Deterministic simulations were formed from yield ensembles by averaging across all the members. An important performance statistic is the correlation coefficient between observed (detrended) and simulated yields (r_{os}). A second measure of simulation accuracy is the RMSE, which includes both simulation bias (which can be corrected based on observations) and correlation (which cannot). The term 'skill', in both deterministic and probabilistic analyses, is reserved for the description of simulation accuracy relative to the accuracy of some baseline forecast method; a simulation may be accurate and still show low skill if similar accuracy can be achieved using the baseline method. For deterministic simulations, the ERA40 yield simulations are the baseline. This definition of deterministic skill is a stringent one because the ERA40 data contain (assimilated) observed weather data whereas the ensemble hindcasts do not.

Two sets of probabilistic analyses have been carried out. The first of these is a dichotomous analysis of crop performance based on an a priori crop failure yield threshold (Y_{cf}) and an a priori detection threshold in probability (P_t). Predictive error is measured using the Brier score, a mean square error which can be used with either probabilistic or deterministic forecasts. Error in this case is defined as the difference between the forecast probability of crop failure (i.e. the fraction of ensemble members predicting failure) and the observed probability (zero or unity). Note that the Brier score cannot be compared across different Y_{cf}, because low Brier scores are favoured by low values of Y_{cf}.

Relative operating characteristics (ROCs) describe the skill of the crop failure hindcasts. ROCs are presented as plots of the hit rate (fraction of crop failure observations which were correctly forecast) against the false alarm rate (fraction of no-failure events that were forecast as events). Zero skill (measured relative to a random forecast) on a ROC curve is represented by the 1 : 1 line. Skill is given by the area between the ROC curve and the 1 : 1 line, with skill being negative if the curve lies below this line. Reliability diagrams provide important information to compliment the ROC curve. They indicate the consistency between the forecast probability of occurrence (plotted on the x-axis) and the observed frequency of occurrence (plotted on the y-axis). The latter is calculated across the subset of observations determined by the forecast probability value (i.e. location on the x-axis). A reliable forecast has points along the 1 : 1 line, where crop failure is predicted with the same frequency with which it is observed.

The second type of probabilistic analysis is based on the ability of the hindcasts to simulate climatological terciles: below normal, normal and above normal. One measure of this ability is the ranked probability score (RPS), averaged over all grid cells and years. The RPS is an extension of the Brier score to multiple categories, and can also be used with deterministic forecasts, by assigning a probability of one to the forecast category.

For both the crop failure analysis and the tercile analysis, two comparisons are made. The first is a comparison with the deterministic ERA40 yield simulation. This is done by assigning a probability of one to the relevant category (failure/no failure or tercile) in which the ERA40 result falls. Note that lower (more skilful) values of the RPS could be achieved by selecting a non-zero probability for all three terciles, with a weighting towards the ERA40 tercile. The second comparison is with climatology: observed frequency of crop failure across all grid cells and all time for the dichotomous analysis, and a probability of 33.3% for each category in the tercile analysis.

Simulations show positive skill with respect to ERA40 when Brier scores (crop failure) or RPS values (terciles) are lower than those of the ERA40 yield simulation. A similar comparison can be made to assess skill relative to climatological forecasts. In addition, a measure of skill relative to random forecasts is afforded for the crop failure case by the ROC analyses described above. The analytical theory behind all the graphs and statistics presented in Section 3.2 can be found in Stanski et al. (1989) and/or Brown (2001). All analyses use all available grid cells (10) for all available years (12). All references to statistical significance are for the 5% level.

3. Results

3.1. Deterministic performance statistics

Figure 3 shows the correlation coefficient between observed (detrended) and simulated yields (r_{os}) for the control run (GCAL–BIC–AUP) and its ERA40 counterpart (GCAL–ERA40). The multi-model yield ensemble mean shows higher correlations than the ERA40 run, and both show high correlations for the north-west of the region (where the climate signal is known to be strong from observations; Challinor et al., 2003). In terms of RMSE, the control run and its ERA40 counterpart perform similarly (Fig. 4).

Simulation accuracy shows some dependence on configuration (the choice of calibration and bias-correction method). The control run has three statistically significant (at the 5% level) values of r_{os}, GCAL–ORI–AUP and NCAL–BIC–MMC both have two, and NCAL–ORI–MMC has one. Hence, bias-correction produces one more significant correlation than the raw data (for both GCAL and ORI) and GCAL produces one more significant correlation than NCAL (for both ORI and BIC). It is possible that the step change in the YGP resulting from cross-validation (Section 2.2) reduces correlations in the NCAL case. Note, however, that these differences in the number of statistically significant correlations may not be statistically significant.

The RMSE of the multi-model yield ensemble mean shows a clearer dependence on configuration than r_{os} (Fig. 4). NCAL tends to produce lower RMSE than GCAL. Bias-correction also tends to reduce the RMSE. NCAL–BIC–MMC performs best

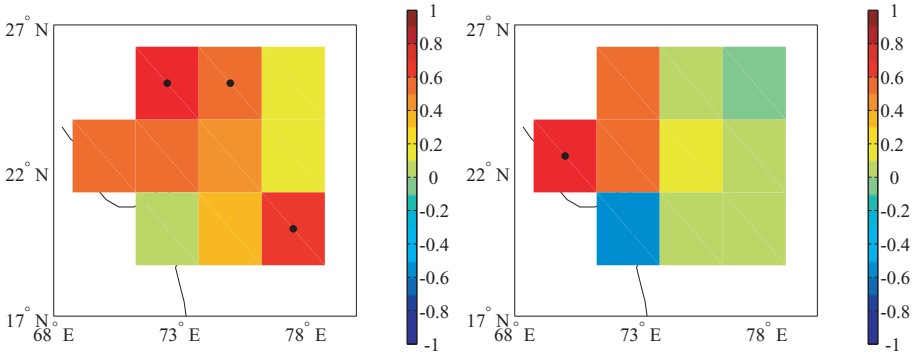


Fig. 3. Correlation coefficients for observed and simulated yields (r_{os}) for the period 1987–1998, for (a) the control run, GCAL–BIC–AUP and (b) GCAL–ERA40. Statistically significant correlations are marked with a dot.

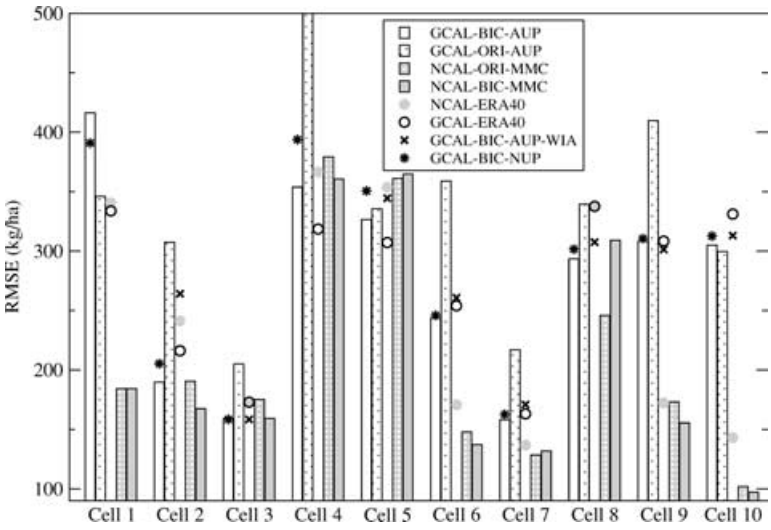


Fig. 4. RMSE of the multi-model yield ensemble mean for eight hindcast simulation experiment and all 10 grid cells. Note that GCAL–ORI–AUP in grid cell 4 has a RMSE of 1021 kg ha^{-1}; this point is not included on the graph. Two points from the GCAL–BIC–AUP–WIA hindcast are also outside the range of the graph: grid cell 1 (647 kg ha^{-1}) and grid cell 4 (543 kg ha^{-1}).

overall (in terms of RMSE): in six out of the 10 grid cells NCAL–BIC–MMC has the lowest (or joint lowest) RMSE. In contrast to the GCAL calibration, NCAL–BIC–MMC has RMSE comparable to or lower than that of its ERA40 counterpart (NCAL–ERA40). For the NCAL–BIC–MMC configuration, the multi-model yield ensemble mean shows more statistically significant values of r_{os} (three) than any other single model (two, for the scnr, scwn and ukmo models).

The simulation using (weather) ensemble averaging at the input stage (GCAL–BIC–AUP–WIA) showed the same number of significant correlations as the control simulation, although the RMSE was higher in eight of the 10 cases (Fig. 4). The simulation without the August forecast update (GCAL–BIC–NUP) produced only one statistically significant correlation and RMSEs which were (slightly) higher than the control run in nine out the 10 cases (Fig. 4).

An analysis of calibrated YGP values for grid cell 1 reveals the reason for the NCAL results generally showing greater accuracy in the multi-model yield ensemble mean than the GCAL results. First, note that NCAL–ERA40 does not outperform GCAL–ERA40 (Fig. 4). This is because the YGPs between the two calibrations do not differ greatly (0.25 for both GCAL and NCAL 1987–1992; 0.20 for NCAL 1993–1998). This implies that cross-validated calibration using 1987–1998 ERA40 data would produce ensembles of yield similar to those in the GCAL case. Hence, estimates of the YGP using data prior to the study period are adequate; the improved performance of NCAL over GCAL in the single-model and multi-model ensemble cases is due primarily to the YGP values being more optimal when calibration uses yield ensemble mean data (as opposed to ERA40 data).

The RMSE of each individual GCM and for each ensemble member for grid cell 1 is presented in Fig. 5. For both NCAL and GCAL, the spread of the RMSE is lower with input data bias-correction than without. NCAL–BIC simulations produce lower RMSE than GCAL–BIC–AUP for all but three models (lody, scwn and ukmo). These are the only three models that show a change in the calibration parameter, YGP, between the two

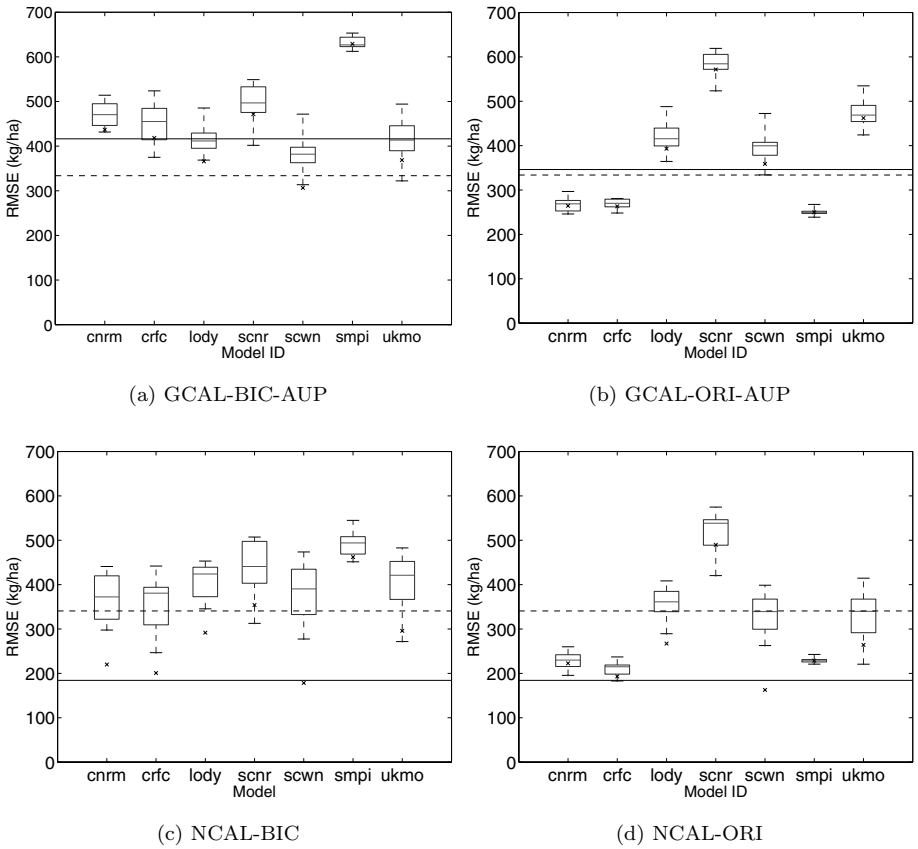


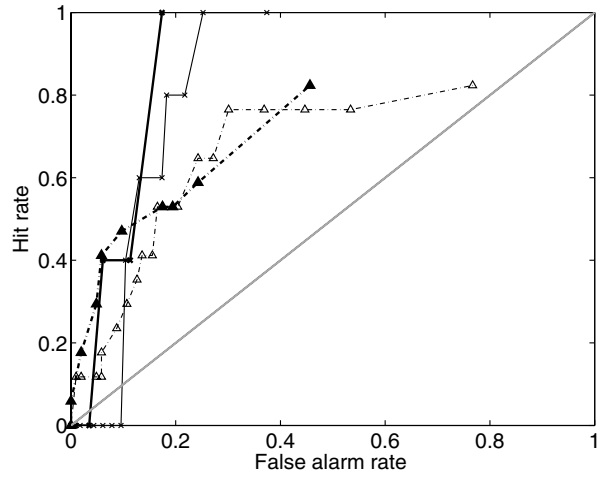
Fig. 5. Box plots (the box shows the upper and lower quartiles, the line within shows the median and the whiskers show the full extent of the data) from four configurations of calibration (GCAL/NCAL) and bias-correction (BIC/ORI), of RMSE in yield for the seven models (GCMs), and RMSE for ERA40 (dashed line) and the multi-model ensemble mean (continuous line). Crosses show the RMSE of the mean of the individual models. All results shown are for grid cell 1 (see Fig. 1). For the NCAL runs, individual models are calibrated individually on the yield ensemble mean (SMC) and the multi-model yield ensemble is calibrated on the multi-model yield ensemble mean (MMC). Table 1 describes the runs performed.

halves of the time series. Figure 5 can also be used to compare the RMSE in yield of individual ensemble members, model (GCM) ensemble means and the multi-model ensemble mean. Only with the NCAL calibration do ensemble means begin to outperform ensemble members. The multi-model yield ensemble mean has a lower RMSE than six of the seven models.

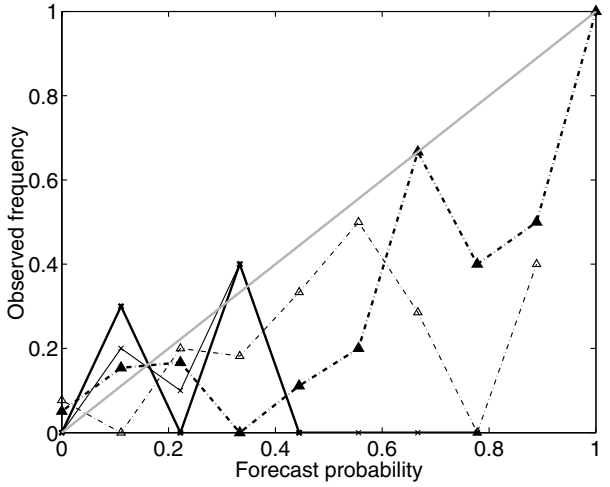
The correlation coefficients and relative means and standard deviations for both the multi-model yield ensemble and the ERA40 simulations in grid cell 1 are shown in Table 2. It is clear that averaging over a number of yield ensemble members reduces the interannual standard deviation of yield. For the multi-model yield ensemble, both the standard deviations and the means are more deficient in the GCAL cases than in the corresponding NCAL cases. For the corresponding ERA40 runs, NCAL improves the mean slightly, but the standard deviation remains too high, so that NCAL is no better than GCAL.

Table 2. Correlation coefficient for observed and simulated yields for the multi-model yield ensemble mean (MME), and for the corresponding ERA40 run, for the four runs used in Fig. 5. Bold indicates significance at the 1% level. Brackets indicate repeated values. Also shown is the ratio of observed and simulated values of (i) standard deviation of yield (σ_y) and (ii) mean yield ($\overline{y}$), for each case. Taking an ensemble average results in a lower interannual standard deviation than that of individual ensemble members. For instance, the nine members of the GCAL–BIC–AUP run have $\sigma_y^{sim}/\sigma_y^{obs} = [1.50, 0.94, 0.91, 1.25, 0.94, 1.33, 1.37, 1.08, 1.30]$

	Correlation		$\sigma_y^{sim}/\sigma_y^{obs}$		$\overline{y}^{sim}/\overline{y}^{obs}$	
Run	MME	ERA40	MME	ERA40	MME	ERA40
GCAL–BIC–AUP	**0.73**	0.56	0.42	1.37	0.45	0.70
GCAL–ORI–AUP	**0.76**	(0.56)	0.25	(1.37)	0.58	(0.70)
NCAL–BIC–SMC	**0.73**	0.50	0.81	1.47	0.85	0.75
NCAL–ORI–SMC	0.57	(0.50)	0.54	(1.47)	1.00	(0.75)



(a)



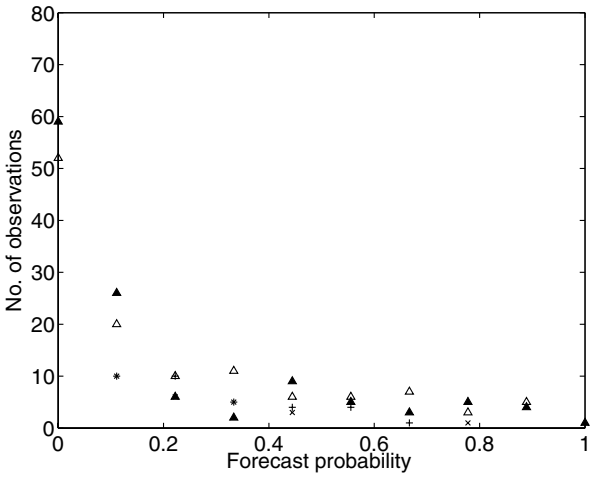
(b)

Fig. 6. Evaluation of control simulation (GCAL–BIC–AUP) for the multi-model yield ensemble (thin lines, open symbols) and best single model (thick lines, filled symbols) for crop failure thresholds of 200 kg ha^{-1} (crosses) and 500 kg ha^{-1} (triangles). (a) ROC curves (skill is proportional to area bounded by the grey 1 : 1 line, the ROC curve, and the horizontal 'hit rate = 1' line); (b) reliability diagrams (reliability is indicated by proximity to the grey line); (c) the number of observations used in each point plotted in (b). In this last plot, black pluses are used in place of black crosses to avoid masking of points. The best single model is defined here as that which, for false alarm rate increasing from zero to one, achieves the highest hit rate at the lowest false alarm rate.

3.2. *Probabilistic performance statistics*

Figure 6a compares the control simulation (GCAL–BIC–AUP) ROCs of the multi-model yield ensemble and the best single model (GCM) at two values of Y_{cf}: 200 and 500 kg ha^{-1}. The latter of these has been identified by Rao et al. (2000) as the point at which costs exceed the value of the crop. All events at the lower yield threshold are simulated, whereas some events are not simulated by any ensemble member for the higher threshold. The best single model is more skilful than the multi-model ensemble at low false alarm rate. In terms of the Brier score and the mean RPS (Table 3), the multi-model yield ensemble shows accuracy similar to or greater than ERA40, and similar or worse skill than the best single model. Note, however, that the best single model varies between each case. Figure 6b compares the same simulations as above using a reliability diagram. The accuracy of this diagram is limited by a low number of observations, particularly at high forecast probabilities (Fig. 6c). The data available suggest that the multi-model ensemble is no less reliable than the best single model.

Figure 7 presents ROC curves for various crop model configurations. The corresponding reliability diagrams are presented in Fig. 8 and the corresponding values of the Brier score and the



(c)

Fig. 6. (cont'd).

Table 3. The Brier score for two threshold crop-failure yield values (Y_{cf}) for the GCAL–BIC–AUP run. Values calculated using the multi-model yield ensemble (MME), the best single model (BSM) and ERA40 are shown. Four of these correspond to the four runs from Fig. 6. Also shown is the mean RPS (averaged over all available years and grid cells) for climatological yield terciles (bounded at 661 and 827 kg ha^{-1})

Run	Y_{cf}	Brier	RPS
MME	200	0.053	0.50
BSM (crfc)	200	0.043	0.56
ERA40	200	0.050	0.60
MME	500	0.143	(0.50)
BSM (lody)	500	0.125	0.53
ERA40	500	0.167	(0.60)

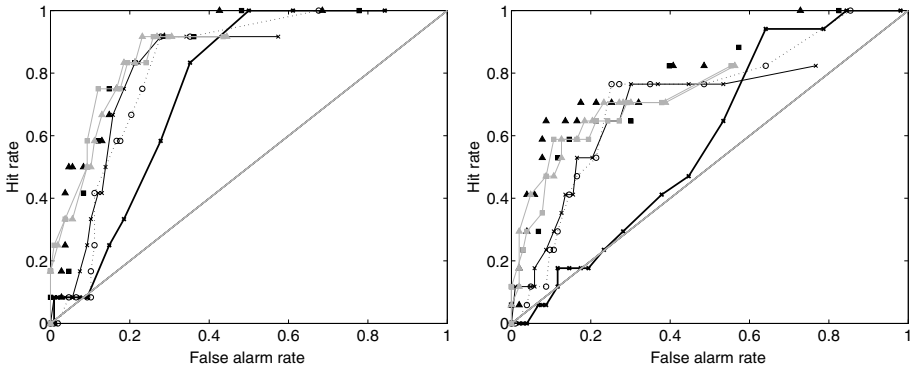


Fig. 7. ROC curves for the period 1987–1989 for crop failure, defined as yield below (a) 400 kg ha^{-1} (12 observed events out of 120 data points) and (b) 500 kg ha^{-1} (17 observed events out of 120 data points). Black lines are for GCAL–BIC runs, with circles denoting the NUP run and crosses denoting the AUP run. Filled symbols (no lines) show NCAL–ORI runs, grey lines show NCAL–BIC runs; for both of these cases, triangles denote calibration using the single-model yield ensemble mean, and squares denote calibration using the multi-model yield ensemble mean. The thick black line is for the GCAL–ORI–AUP run. The thick grey lines shows the zero-skill baseline.

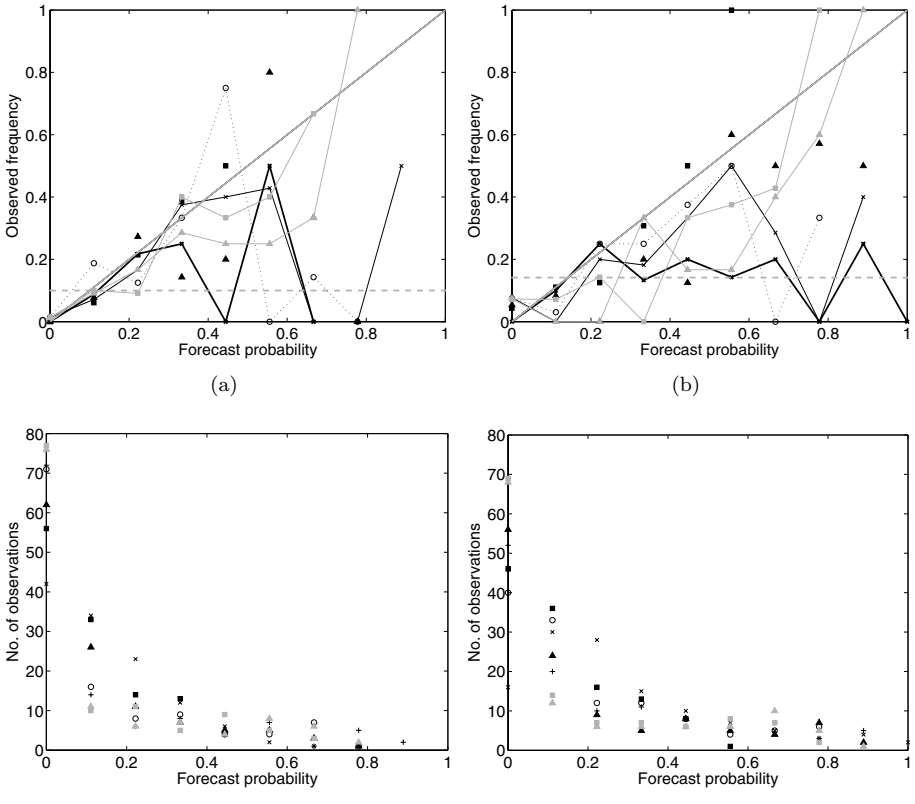


Fig. 8. Reliability diagram for for the period 1987–1989 for crop failure, defined as yield below (a) 400 kg ha^{-1} and (b) 500 kg ha^{-1}. The climatology is plotted as a dotted grey line and a perfectly reliable forecast as a solid grey line. The corresponding number of observations used for each point plotted are shown in (c) for 400 kg ha^{-1} and (d) for 500 kg ha^{-1}. The legend is exactly as Fig. 7: black lines are for GCAL–BIC runs, with circles denoting the NUP run and crosses denoting the AUP run. Filled symbols (no line) show NCAL–ORI runs, grey lines show NCAL–BIC runs; for both of these cases, triangles denote calibration using the single-model yield ensemble mean, and squares denote calibration using the multi-model yield ensemble mean. The thick black line is for the GCAL–ORI–AUP run. In (c) and (d), black pluses are used in place of black crosses to avoid masking of points.

Table 4. The Brier score for two threshold crop-failure yield values ($Y_{cf} = 400$ and 500 kg ha^{-1}) for the five runs used in Fig. 7 and for two ERA40 runs. Note that the Brier score is by definition lower for rarer events, so that values for different Y_{cf} should not be compared. The Brier scores of climatology (probability of crop failure equals observed climatological occurrence across all years and grid cells) are shown for comparison. Also shown is the mean RPS (averaged over all available years and grid cells) for climatological yield terciles (RPS–CLI, bounded at 661 and 827 kg ha^{-1}) and for model yield terciles (RPS–MOD, with configuration-dependent yield boundaries). The RPS of climatology (33.3% probability for each tercile in the observations) is shown for comparison

Run	Brier-400	Brier-500	RPS–CLI	RPS–MOD
GCAL–BIC–NUP	0.101	0.134	0.51	0.33
GCAL–BIC–AUP	0.100	0.143	0.50	0.32
NCAL–ORI–MMC	0.073	0.101	0.26	0.28
NCAL–ORI–SMC	0.075	0.106	0.30	0.28
GCAL–ORI–AUP	0.093	0.190	0.63	0.36
NCAL–BIC–MMC	0.071	0.110	0.28	0.29
NCAL–BIC–SMC	0.077	0.119	0.31	0.27
GCAL–ERA40	0.092	0.167	0.60	0.49
NCAL–ERA40	0.083	0.125	0.44	0.42
Climatology	0.090	0.122	0.28	–

mean RPS are presented in Table 4. Simulations are more reliable for a crop failure yield threshold (Y_{cf}) of 400 kg ha^{-1} than for $Y_{cf} = 500$ kg ha^{-1}. RPS values for the multi-model yield ensemble tend to be better (lower) than ERA40, but similar to or worse than climatology. Given that GLAM tends to be more accurate under water-limiting conditions, therefore favouring prediction of crop failure over prediction of high yields, these results are not surprising.

On the whole, the hindcasts are not particularly sharp for crop failure prediction: there are few occasions when the forecast probability is high (Figs. 8c and 8d). The few high probabilities that are predicted do not generally indicate greater certainty: most points in Figs. 8a and 8b lie below the 45° line.

The least skilful simulation overall is GCAL–ORI–AUP, although this does produce a reliable forecast at low probability (Figs. 8a and 8b). This may however be due to the greater sample size at low probability for this run (Figs. 8c and 8d). Bias-correction of input weather data improves the simulations, although it can also remove the ability of any of the ensemble members to simulate some of the observed crop failures (Fig. 7). The GCAL–BIC–NUP and GCAL–BIC–AUP simulations produce similar results to each other, indicating that the use of the August update does not impact significantly on the results. Calibration by cross-validation using yield ensemble mean data (NCAL) improves ROC skill further and also produces the lowest values of the Brier score and mean RPS (Table 4). Brier scores compare well with both ERA40 and climatology values showing some skill in the prediction of crop failure. RPS values show skill relative to the deterministic ERA40 hindcast but not relative to climatology.

NCAL multi-model yield ensembles (both with and without bias-correction) produced lower values of RPS than any of their single-model counterparts: values were 10% and 15%, respectively, lower than the lowest single-model value.

Whether calibration of the crop model treats the multi-model ensemble as one model or as a sum of separately calibrated models (see Section 2.2) makes little difference to the ROC curves. However, there is some indication that multi-model calibrations may produce more reliable forecasts at high (30–60%) probability thresholds (Fig. 8). This suggests that, for probabilistic information, calibration of the crop model using single-model yield ensembles may be less effective than calibration using the multi-model yield ensemble.

Overall, the results suggest that either bias-correction of input weather data (BIC), or calibration via cross-validation with yield ensemble means (NCAL), or both, are required in order for the hindcasts to be skilful. Further, NCAL alone appears to provide a considerable improvement. The values of YGP calibrated using ERA40 differ across NCAL and GCAL by 0.05 or less for seven grid cells, and 0.10 or less for nine grid cells, suggesting that it is the calibration on yield ensemble means, as opposed to the use of data within the study period, that is the source of the increased skill in the NCAL case. This suggests that for probabilistic information, crop model calibration of the YGP on yield ensemble mean data can effectively be used as a bias-correction.

3.3. Delayed sowing window

Use of the delayed sowing window described in Section 2.1 with the ERA40 data results in correlations between observed and simulated yields (r_{os}) strengthening in eight of the 10 grid cells. The largest change is significant at the 5% level and occurs in grid cell 10, between GCAL–ERA40 ($r_{os} = 0.01$) and GCAL–ERA40–DSW ($r_{os} = 0.74$). The corresponding values for the mean yield from the multi-model ensemble runs are $r_{os} = 0.62$ (GCAL–BIC–AUP) and $r_{os} = 0.58$ (GCAL–BIC–AUP–DSW). Hence, the representation of forecast uncertainty, in this case at least, makes the representation of uncertainty in the sowing window redundant.

3.4. District-level analysis

In this section we describe the results of the ODG simulations (see Section 2.3). Using ERA40 data with both (in turn) of the sowing windows described in Section 2.1, optimal values of the YGP were found for the 1989–1998 period, and four districts with statistically significant ($p < 0.05$) correlations

Table 5. Performance of multi-model yield ensemble means for four individual districts in the three grid cells shown. Correlation coefficients (with values statistically significant at the 5% level in bold) and the ratio of observed and simulated standard deviation are shown for the single-district (ODG) runs and the corresponding area-averaged geocode (AAG) runs

		Correlation		$\sigma_y^{sim}/\sigma_y^{obs}$	
Grid cell	District	ODG	AAG	ODG	AAG
6	Ratlam	**0.66**	0.20	0.31	0.55
9	Jalgaon	0.50	0.12	0.29	0.54
10	Parbhani	0.52	0.54	0.37	0.35
10	Yavatmal	**0.66**	0.54	0.30	0.35

emerged. Simulations and analysis focused on these districts, which are marked in Fig. 1. The values of the YGP obtained from the ERA40 analysis using the standard sowing window were used to perform the crop simulations (i.e. GCAL calibration). The simulations were performed using the BIC–AUP configuration.

The resulting multi-model yield ensemble means are compared to their area-averaged counterparts (GCAL–BIC–AUP) in Table 5. Because standard deviations of yield (σ_y) at smaller spatial scales are often higher than those at larger spatial scales, the disparity in σ_y is in some cases greater for the district-level case. Correlation coefficients however are generally an improvement on the area-averaged case. Hence, where forecasts are not useful on the grid scale because of low skill, they may useful on the subgrid scale.

4. Summary and discussion

4.1. Optimal calibration and bias-correction methods

A number of crop model calibration methods exist. Most, if not all, of these are applicable to deterministic simulations, where one set of inputs determines one set of outputs. The crop model parameters may be adjusted in order that yields and/or phenology match observations (Travasso and Delécolle, 1995; Kaur and Hundal, 1999). Alternatively, a yield correction factor can be applied to the predicted yields in order to minimize the RMSE (Jagtap and Jones, 2002). Calibration methods for probabilistic studies are not well developed; Marletto et al. (2005) applied no formal calibration to their crop model. The results presented here are a first step to developing crop model calibration procedures for probabilistic studies.

The analyses of correlations and RMSE presented in Section 3.1 show that simulation accuracy is dependent on the method of calibration. For multi-model yield ensemble means, calibration of the crop model using ERA40 data often showed higher RMSE than calibration via cross-validation (Figs. 4 and 5). However, the number of significant correlations between observed and simulated yields was not greater when cross-validation was used (Section 3.1). Bias-correction tended to improve results for yield ensemble means (Section 3.1) but not for prediction of crop failure (Section 3.2). This bias-correction was towards ERA40 data. Because the ERA40 rainfall is deficient in both mean and standard deviation when compared to observed gridded data (Challinor et al., 2005) it is anticipated that bias-correction to observations could improve results further. These results therefore suggest that the best configuration for producing output based on the yield ensemble mean is cross-validation on the ensemble mean, with input weather data bias-correction (i.e. NCAL–BIC).

For probabilistic analyses, Figs. 6a and 7 show that bias-correction of input weather data removed the ability to simulate some events. The yield ensemble means again emerge as the most favourable data on which to calibrate. However, for yield ensemble means, separate calibration of the crop model for each single model (GCM) was shown to have no advantage over calibration using the multi-model ensemble (Fig. 7 and Table 4).

Both the probabilistic and deterministic analyses suggest that estimates of the YGP do not need to be based on yields from within the study period. However, it is important to base calibrations on data properly adjusted for technology trend. In this study, all yields were adjusted to 1987 levels, using a linear regression over the periods 1966–1986 and 1987–1998. Some of the trends varied considerably between these two periods (e.g. over 60 kg ha^{-1} yr^{-1}). Therefore, if this change were not accounted for, it would become important to calibrate the YGP on data from a period closer to the study period.

4.2. Skilful probabilistic information

The importance of using probabilistic information has been noted for both hydrological (Krzysztofowicz, 2001) and crop modelling (Hansen and Indeje, 2004) studies. Probabilistic information on seasonal time-scales currently available to the agricultural sector consists typically of climate terciles – an indication of the probability of low, average or high rainfall (e.g. Barnston et al., 2003). The results presented here suggest that an extension of such forecasts to yield can produce predictions comparable to the use of climatological tertiaries (i.e. low, or even zero, skill). Calibration using yield ensemble mean data with no bias-correction of input weather data (NCAL–ORI) produced the lowest (i.e. most skilful) mean RPS. For this simulation, the multi-model yield ensemble produced lower values of RPS than any single model.

Positive skill was seen in the simulation of crop failure, with more severe failures being more reliable (Fig. 8). The best single model was not consistent across crop failure thresholds (Table 3). There was no clear advantage to the multi-model approach in the prediction of crop failure. A crop-failure prediction system based

on the principles outlined in this study could potentially be used as part of a hybrid (modelling plus descriptive) decision support system, such as that described by Hansen (2002).

The use of the August forecast to update the crop model did not appear to have a significant impact on the results (Figs. 7 and 8 and Table 3).

4.3. Skill relative to deterministic simulations

The multi-model yield ensemble was formed from a simple average of the 63 ensemble members. Even this simple configuration, with no weights, showed evidence of positive skill: a greater number of significant correlations with observed yields than the ERA40 simulations in the GCAL case, and lower RMSE overall in the NCAL case. However, averaging over ensemble members did tend to reduce the interannual standard deviation (Table 2). The multi-model yield ensemble also showed more statistically significant correlations with observations than any single model (Section 3.1). Averaging of the weather data (all 63 ensembles) prior to crop simulation produced yields with equally high correlations but higher RMSE than the control run (Section 3.1). In a similar (but not identical) comparison, Trnka et al. (2004) found that the use of scenarios averaged across several GCMs with the CERES crop model resulted in yields with similar characteristics to single-GCM scenarios. Because GLAM is computationally cheap to run, there is no great advantage in this scenario-averaged approach.

The use of the August forecast had a small positive impact on the RMSE (up to 10%, but mostly less; Fig. 4) and a positive (but not necessarily statistically significant) impact on the correlation between observed and simulated yields (Section 3.1).

4.4. Impact of uncertainties

Two preliminary studies of uncertainty were made. A delayed sowing window (Section 3.3) significantly (in the statistical sense) improved correlations for one of the grid cells using ERA40. The corresponding multi-model yield ensemble mean correlations were both statistically significant but not significantly different from each other. Sowing window uncertainty, then, may be of secondary importance when ensemble mean forecasts are used. If accurate simulations result, then this is an advantage over the deterministic approach. Note that the choice of sowing window made by the farmer remains important (e.g. Rao et al., 2000); it is only with respect to prediction that it may be secondary.

For two of the grid cells, significant correlations were found at the subgrid scale where there were none at the grid scale (Table 5). This implies that subgrid heterogeneity can impact on skill over large areas, making it a potentially important source of uncertainty. It also implies that as part of the development of a forecasting system, a study of the spatial scale(s) on which the yield simulations show skill would be worthwhile. This study has not taken account of another scale-related issue: that the yield scenarios themselves depend upon the spatial scale of the climate and soils information (e.g. Mearns et al., 1999, 2001).

5. Conclusions: implications for yield forecasting

The ensembles of yield developed in this study using DEMETER hindcast weather ensembles and the GLAM crop model have shown predictive skill in both the ensemble mean and the ensemble spread. In both cases, calibration using yield ensemble mean information has lower RMSE than calibration using reanalysis. However, there is some evidence that cross-validation of the YGP reduces the correlation between observed and simulated yield (Section 4.1). An important caveat to any conclusions drawn is the length of the time series: 12 yr represents a small sample over which to estimate the predictability of crop yields. Further study using a longer time series of ensemble hindcasts and yield data would be needed to enable conclusions to be drawn more firmly.

These results of this study suggest four implications for forecasting on short-to-medium time-scales (a season to a decade). First, calibration on yield ensemble means prior to the study period would give the greatest predictive skill for studies of this kind. Secondly, there is the potential for the probabilistic prediction of crop failure, defined by a given threshold yield value. Tercile forecasts may also become feasible if the skill of GCMs increases. Thirdly, ensemble means can show skill in predicting interannual variability in yield. Because bias-correction to ERA40 showed the potential to increase skill further, improved GCM skill has the potential to translate into improved deterministic yield prediction. Fourthly, uncertainties in crop model inputs are important, particularly when operating on large spatial scales. However, the results presented here suggest that one of these uncertainties, the sowing window, may not require explicit modelling.

The implications for forecasting on multidecadal (climate change) time-scales are as follows. First, yield ensembles based on the perturbation of uncertain parameters in both crop and climate models could be used to create forecasts of mean yields, in the same way as the ensembles in this study. This would smooth out any information on the interannual variability but may average out errors associated with the prediction of mean yields in future climates. Secondly, the results suggest that, as long as the technology trend is known to some degree of accuracy (using, for example, a linear regression), changes in the YGP on decadal time-scales may be small. Hence, climate change impacts studies may be formed from a number of plausible technology scenarios whilst keeping the YGP constant. Alternatively, actual yield values may be ignored, and the focus placed on spatial patterns of yield.

Finally, it is worth noting that the issue of extreme events, which may become more important in future climates, has not been addressed in this study because the time series was not long enough. It is possible that some of the earliest and most severe impacts of climate change will come from the exceeding of climate thresholds, such as temperature, over short periods during critical crop development stages (e.g. Wheeler et al., 2000). Climate change impacts studies clearly need to take account of this.

6. Acknowledgments

The authors are grateful to the ICRISAT for the crop productivity data. The reviewer's comments helped to improve the clarity of the paper. AJC would also like to thank Dr Christopher Ferro for discussions on statistical inference. FJDR received support from the European Union funded DEMETER project (EVK2-1999-00024).

References

Barnston, A. G., Mason, S. J., Goddard, L., DeWitt, D. G. and Zebiak, S. E. 2003. Multi-model ensembling in seasonal climate forecasting at IRI. *Bull. Am. Meteorol. Soc.* **84**(12), 1783–1796.

Brooks, R. J., Semenov, M. A. and Jamieson, P. D. 2001. Simplifying sirus: sensitivity analysis and development of a meta-model for wheat yield prediction. *Eur. J. Agron.* **14**, 43–60.

Brown, B. G. 2001. Verification of precipitation forecasts: a survey of methodology Part II: verification of probability forecasts at points. In: *Proceedings of the WWRP/WMO Workshop on the Verification of Quantitative Precipitation Forecasts*, Prague, 14–16 May 2001, NCAR, Boulder, CO, USA (available on-line at http://www.chmi.cz/meteo/ov/wmo/).

Camberlin, P. and Diop, M. 1999. Inter-relationships between groundnut yield in Senegal. Interannual rainfall variability and sea-surface temperatures. *Theor. Appl. Climatol.* **63**, 163–181.

Cantelaube, P. and Terres, J. M. 2005. Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A* **57A**, 476–487.

Challinor, A. J., Slingo, J. M., Wheeler, T. R., Craufurd, P. Q. and Grimes, D. I. F. 2003. Towards a combined seasonal weather and crop productivity forecasting system: determination of the spatial correlation scale. *J. Appl. Meteorol.* **42**, 175–192.

Challinor, A. J., Wheeler, T. R., Slingo, J. M., Craufurd, P. Q. and Grimes, D. I. F. 2004. Design and optimization of a large-area process-based model for annual crops. *Agric. For. Meteorol.* **124**, 99–120.

Challinor, A. J., Wheeler, T. R., Slingo, J. M., Craufurd, P. Q. and Grimes, D. I. F. 2005. Simulation of crop yields using the ERA40 reanalysis: limits to skill and non-stationarity in weather–yield relationships. *J. Appl. Meteorol.* in press.

FAO/Unesco. 1974. *FAO/Unesco Soil Map of the World, 1:5,000,000*, 10 volumes.

Fischer, G., Shah, M. and van Velthuizen, H. 2002. Climate change and agricultural vulnerability, Technical Report, International Institute for Applied Systems Analysis, available at http://www.iiasa.ac.at/Research/LUC/.

Hansen, J. W. 2002. Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges. *Agric. Syst.* **74**, 309–330.

Hansen, J. W. and Indeje, M. 2004. Linking dynamic seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. *Agric. For. Meteorol.* **125**, 143–157.

Hansen, J. W. and Jones, J. W. 2000. Scaling-up crop models for climatic variability applications. *Agric. Syst.* **65**, 43–72.

Hsieh, W. W., Tang, B. Y. and Garnett, E. R. 1999. Teleconnections between Pacific sea surface temperatures and Canadian prairie wheat yield. *Agric. For. Meteorol.* **96**(4), 209–217.

Jagtap, S. S. and Jones, J. W. 2002. Adaptation and evaluation of the cropgro–soybean model to predict regional yield and production. *Agric. Ecosyst. Environ.* **93**, 73–85.

Kaur, P. and Hundal, S. S. 1999. Forecasting growth and yield of groundnut (Arachis hypogaea) with a dynamic simulation model 'PNUTGRO' under Punjab conditions. *J. Agric. Sci.* **133**, 167–173.

Krzysztofowicz, R. 2001. The case for probabilistic forecasting in hydrology. *J. Hydrology* **249**(1–4), 2–9.

Landau, S., Mitchell, R. A. C., Barnett, V., Colls, J. J., Craigon, J. and Payne, R. W. 2000. A parsimonious, multiple-regression model of wheat yield response to environment. *Agric. For. Meteorol.* **101**, 151–166.

Marletto, V., Zinoni, F., Criscuolo, L., Fontana, G., Marchesi, S. and co-authors. 2005. Evaluation of downscaled DEMETER multi-model ensemble seasonal hindcasts in a northern Italy location by means of a model of wheat growth and soil water balance. *Tellus A* **57A**, 488–497.

Mearns, L. O., Mavromatis, T. and Tsvetsinskaya, E. 1999. Comparative response of EPIC and CERES crop models to high and low spatial resolution climate change scenarios. *J. Geophys. Res.* **104**(D6), 6623–6646.

Mearns, L. O., Easterling, W., Hays, C. and Marx, D. 2001. Comparison of agricultural impacts of climate change calculated from the high and low resolution climate change scenarios: Part I. The uncertainty due to spatial scale. *Climate Change* **51**, 131–172.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004. Development of a European multi-model ensemble system for seasonal to interannual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872.

Rao, K. N., Gadgil, S., Rao, P. R. S. and Savithri, K. 2000. Tailoring strategies to rainfall variability—the choice of the sowing window. *Current Science* **78**(10), 1216–1230.

Reddy, P. S., ed. 1988. *Groundnut*. Indian Council of Agricultural Research, Krishi Anusandhan Bhavan, Pusa, New Delhi, India.

Rijks, D., Rembold, F., Négre, T., Gommes, R. and Cherlet, M., eds. 2003. *Crop and rangeland monitoring in Eastern Africa for early warning and food security*. Joint Research Centre–Food and Agriculture Organization, Proceedings of an International Workshop ogranized by JRC–FAO, 28–30 January 2003, Nairobi.

Southworth, J., Randolph, J. C., Habeck, M., Doering, O. C., Pfeifer, R. A. and co-authors. 2000. Consequences of future climate change and changing climate variability on maize yields in the mid-western United States. *Agric. Ecosyst. Environ.* **82**, 139–158.

Stanski, H. R., Wilson, L. J. and Burrows, W. R. 1989. Survey of common verification methods in meteorology, 2nd edition. Research Report MSRB 89-5, Atmospheric Environment Service, Forecast Research Division, 4905 Differin Street, Downsview, Ontario, Canada M3H 5T4. Available on-line at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Stanski_et_al/Stanski_et_al.html.

Travasso, M. I. and Delécolle, R. 1995. Adaptation of the CERES–wheat model for large area yield estimation in Argentina. *Eur. J. Agron.* **4**(3), 347–353.

Trnka, M., Dubrovský, M., Semerádová, D. and alud, Z. 2004. Projections of uncertainties in climate change scenarios into expected winter wheat yields. *Theor. Appl. Climatol.* **77**, 229–249.

Wheeler, T. R., Craufurd, P. Q., Ellis, R. H., Porter, J. R. and Prasad, P. V. V. 2000. Temperature variability and the annual yield of crops. *Agric. Ecosyst. Environ.* **82**, 159–167.