

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

By A. J. Challinor^{1,2*}, J. M. Slingo¹, T. R. Wheeler² and F. J. Doblas-Reyes³,
¹*CGAM, Dept. of Meteorology, The University of Reading, Reading RG6 6BB, U.K.*; ²*Department of Agriculture, The University of Reading, Reading RG6 6AT, U.K.*; ³*European Centre for Medium-Range Weather Forecasting (ECMWF) Shinfield Park, RG2 9AX Reading, U.K.*

31 March 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, hindcasts of crop yield (groundnut, *Arachis hypogaea* L.) were produced for ten 2.5 by 2.5 degree grid cells in western India using the DEMETER ensembles and the General Large-Area Model for annual crops, GLAM.

Four key issues are addressed by this study: firstly, crop model calibration methods for use with ensemble data are assessed. 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 ensemble means to predict interannual variability in crop yield is examined and their skill assessed relative to baseline simulations using reanalysis data (the European Centre for Medium Range Weather Forecasts forty-year reanalysis, ERA40). The skill of multi-model yield ensemble means is greater than or comparable to the skill 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 in the ERA40 case than in the multi-model ensemble mean case. Subgrid heterogeneity is shown to impact model skill: where there is no skill on the grid scale, there may be skill on the subgrid scale.

The impact of the results of this study for yield forecasting on timescales from the seasonal to the multi-decadal (climate change) are noted in the conclusions.

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 multi-decadal timescales 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, either by taking an empirical approach (e.g. Landau et al. 2000; Camberlin and Diop 1999; Hsieh et al. 1999) or process-based approach (e.g. Brooks et al. 2001; Jagtap and Jones 2002; Southworth et al. 2000). However, climate on seasonal timescales is inherently unpredictable and recent progress in the use of multi-model ensembles achieved through the DEMETER project (Palmer et al. 2003) provides an excellent opportunity to explore crop yield predictability using probabilistic methods (see also Cantelaube and Terres 2004). This is the topic 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, which uses seasonal timescales, will have relevance

for studies of longer timescales, and climate change, since inherent unpredictability and uncertainties will need to be estimated for these timescales also.

The process-based approach to crop yield prediction has the advantage of potentially capturing changes in the nature of the weather–yield relationship due to changes in climate (e.g. intra-seasonal variability and increased CO₂ levels) and the disadvantage of often having a high input data requirement. For empirical approaches, the converse tends to be true. Challinor et al. (2004a) developed a relatively simple, large area, process-based model (GLAM — the General Large-Area Model for annual crops) which aims to combine the advantages of these two approaches. The model simulates interannual variability in groundnut yield over the Gujarat region of India well when driven with either observed gridded weather data (Challinor et al. 2004a) or the European Centre for Medium Range Weather Forecasts (ECMWF) forty-year reanalysis, ERA40 (Challinor et al. 2004b). GLAM is used in this study and described briefly in section 2.1.

This study aims to develop methods for the use of ensembles with this type of crop model. The chosen crop is

groundnut (i.e. 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). Note, however, that the methods used in this study are not location or crop specific.

The development and assessment of probabilistic yield forecast methods using General Circulation Models (GCMs) raises a range of issues. The first of these is related to the models used: 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 (since the crop model has already been tested in deterministic studies using observed data and reanalysis, crop model skill is a secondary issue in this case). The second issue relates to the output from the system: 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 most appropriate format for processed model output needs to be determined according to where skillful information lies. The third issue is that of uncertainty. The spatial scale on which simulations are carried out is one source of uncertainty. When run over large areas, subgrid heterogeneity may result in low skill. 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 model skill.

The methods used to begin to address the above issues are described in section 2. This section ends with the formulation of four key issues on which this paper focusses. The results are presented in section 3 and section 4 discusses the results in the context of the identified key issues.

2. Method: formulation of crop yield hindcasts

2.1. Crop modelling techniques

The crop model used for this study is the General Large-Area Model for annual crops (GLAM; Challinor et al. 2004a). 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 sub-seasonal variability and retaining validity under unprecedented conditions, such as are likely under future climates). The model, in the configuration employed here, uses daily values of solar radiation, minimum and maximum temperature, and rainfall. It has an intelligent sowing routine, which requires as input a sowing window. Sowing occurs on the first day on which the soil is moist enough or at the end of the window (crisis-sowing) if this does not occur.

GLAM aims to reproduce the impact of weather on observed crop yield. This aim impacts on the level of complexity of the model in two ways: First, complexity at a level far-removed from yield-determining processes is omitted (see Sinclair and Seligman 2000). Second, 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 rest 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. YGP may also be set to simulate zero yield gap (i.e. yield potential). However, it is actual yields with which model output is compared, and this necessitates the calibration of YGP.

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) contains further description, and some analysis of, these data. For the current study, calibration of YGP uses either (i) ECMWF forty-year reanalysis data (ERA40; <http://www.ecmwf.int/research/era/>) prior to the study period (1966–86) to determine a single value of YGP, referred to as GCAL, or (ii) cross-validation using data within the study period (1987–92 data are used to determine YGP for 1993–98, and vice-versa), referred to as NCAL. The source of data for the NCAL calibration was the same source as for the relevant simulations — i.e. ERA40 data was used for ERA40 runs and ensemble data for ensemble runs. Two calibration methods were used for NCAL ensemble runs. In the first, a single pair of YGP values was determined by using the multi-model ensemble mean (multi-model calibration, MMC). In the second method, each single-model ensemble mean was used to determine a pair of YGP values for ensemble members from that model (single-model calibration, SMC). The YGPs resulting from this second method may include a component of weather data bias-correction for each individual model.

The yield data used are either the district-level data (figure 1) or data which have been upscaled to the model grid using an area-weighted mean (this assumes 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 deviation of yields on the simulation grid.

2.2. Input data

Soil hydrological properties were derived from FAO/Unesco (1974) following Challinor et al. (2004a). 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. Crisis sowing is assumed if the sowing criterion is not met thirty days after the first day in the window. Despite having observations of the sowing window, the planting date remains a considerable source of uncertainty. There is some evidence that for Gujarat a sowing window start date which is later than the observed value (early August as opposed

to mid-June) produces more realistic simulations (Challinor et al. 2004b). As a preliminary study of the impact of uncertainty in the sowing window, some simulations with a moderately delayed sowing window (starting on July 9th; denoted by DSW) across the whole region were carried out.

Input weather data ensembles are the DEMETER ensembles (Palmer et al. 2003). Seven models (denoted here as *cnrm*, *crcf*, *lody*, *scnr*, *scwn*, *smpi* and *ukmo*), each with nine ensemble members, each run four times a year (initialised on the first day of February, May, August and November) were used. The study period, 1987–98, is the period for which both input weather data and observed groundnut yield data are available. The study region is ten grid cells in western India (see figure 2).

Each ensemble member is a six-month daily time series. This creates two possibilities for this study: (i) use of the ensembles initialised in May for a three-month groundnut simulation period, followed by use of the ensembles initialised in August (August update, AUP); (ii) use of the simulation initialised in May for the whole of the growing season (no update, NUP). Note that the AUP simulations involve a step change on August 1st to a time series chosen from the nine new ensemble members using ensemble identification number; this is essentially an arbitrary choice.

Two sets of input ensemble weather data were used: one set is the raw DEMETER ensembles (original data, ORI) and the other 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. These simulations differ from those of Challinor et al. (2004b) 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 by 0.5 degrees) and the current study uses data interpolated to the DEMETER grid using the Meteorological Archive and Retrieval System interpolation tool.

Bias-correction (BIC) was applied to daily values of temperature and precipitation for each model 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 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 be estimate the seasonal cycle with the ERA40 data. The bias was defined as the difference between the model and the ERA40 seasonal cycles. Finally, bias-corrected hindcasts were computed as the difference between the hindcasts issued by the coupled models 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.

2.3. Hindcast experiments

Hindcasts of crop yield were created by driving GLAM with individual DEMETER ensemble members. Ensemble mean yields, for either a single model, 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 model grid cells for calibration and evaluation of the model (Area-Averaged Geocode). Table 1 summarises the experiments conducted and the methods used for this set of hindcasts. The second set of hindcasts used the district-level yield data, one district at a time, for calibration and evaluation (One District Geocode). Since these runs were conditional on the Area-Averaged Geocode results, the methods for these runs are described together with the results in section 3.4.

Hindcast 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 (acronyms refer to the hindcast experiments listed in table 1):

1) How is a probabilistic forecasting system best calibrated — is bias correction of input data needed (BIC vs. ORI)? Do estimates of the Yield Gap Parameter need to be current or will estimates based on historical yields suffice (NCAL-ERA40 vs. GCAL-ERA40)? Should calibration be carried out on data from the same source as the data for simulation (NCAL vs. GCAL multi-model ensemble runs)?

2) How can skillful probabilistic forecasts of crop yield be formed from a set of 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?

3) How skillful are the ensemble mean hindcasts (relative to the skill of the benchmark ERA40 simulations)? Do updated forecasts produce increased skill (AUP vs. NUP)? How does the multi-model 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 skill of the simulations? Preliminary analyses described in this paper examine: (i) whether uncertainty in the sowing window affects the skill of both the deterministic and probabilistic simulations equally (GCAL-BIC-AUP[-DSW] vs. GCAL-ERA40[-DSW]); (ii) whether, where there is no skill at the grid-scale, this can be attributed to heterogeneity within the grid cell. Specifically, is there skill on the sub-grid scale (One District Geocode)?

Each of these questions is discussed in turn in the four parts of section 4.

3. Results

3.1. Deterministic performance statistics

Deterministic simulations were formed from yield ensembles by averaging across all the members. The most important performance statistic is the correlation coefficient between observed (detrended) and simulated yields (r_{os}), since bias in the mean and standard deviation can be corrected. Figure 3 shows r_{os} for the control run (GCAL-BIC-AUP) and its ERA40 counterpart (GCAL-ERA40). The multi-model 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).

For one of the grid cells, the Root Mean Square Error (RMSE) of the multi-model ensemble mean varies consider-

ably depending on the calibration and bias correction methods. This variation is presented, together with the RMSE–range of each individual model, and that of ERA40, in figure 4. Corresponding correlation coefficients and relative means and standard deviations are shown in table 2.

Comparison of figures 4b and 4d shows the impact of calibration method where no input data bias correction is used. Calibration on yield data and ensemble input data from the study period (1987–98) via cross-validation (NCAL) results in lower RMSE than calibration on ERA40 data prior to the study period (GCAL), for all single models, and for the multi-model ensemble. Since the use of NCAL in this case implies some form of single-model (weather) bias correction as part of the improved estimate of YGP, this is not surprising. In contrast, comparison of figures 4a and 4c shows that NCAL-BIC simulations produce lower RMSE than GCAL-BIC 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 halves of the time series.

For the case of the multi-model ensemble mean, NCAL provides far more accurate simulations both where bias correction is performed and where it is not (56% and 47% lower RMSE, respectively). This is not true of the corresponding ERA40 simulations, where the RMSE for the two calibrations are within 3% of each other. This is because the YGPs between the two calibrations do not differ greatly (0.25 for both GCAL and NCAL 1987–92; 0.20 for NCAL 1993–98). This implies that cross-validated calibration on 1987–98 ERA40 data would produce ensembles of yield similar to those in the GCAL case. Hence 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 is on ensemble-mean data (as opposed to ERA40 data).

Comparison of figures 4a and 4b shows that for GCAL, the spread of model RMSE is lower with input data bias correction than without. The same is true of NCAL. However, the RMSE of the multi-model ensemble mean is not reduced by bias correction. r_{os} is increased by bias correction in the NCAL case but not the GCAL case (table 2). This may be due to the improved calibration in the NCAL case. Hence bias correction improves results only when the model is well-calibrated.

Table 2 shows that taking an average over a number of ensemble members reduces the interannual standard deviation of yield, in this case from above-observed values for a single run, to values that are below observed. For the multi-model 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. Note that ERA40 rainfall is deficient in both mean and standard deviation when compared to observed gridded data (Challinor et al. 2004b). Hence a bias-correction to ERA40 falls short of being a bias-correction to observations.

Figure 4 can be used to compare the RMSE of individual ensemble members, model ensemble means and the multi-model ensemble mean. Only with the NCAL calibration do ensemble means begin to outperform ensemble mem-

bers. The multi-model ensemble mean has a lower RMSE than six of the seven models. Bias correction in the NCAL case produces either similar or better performance of ensemble means than no bias-correction. This is not true of the GCAL simulations, where bias-correction can degrade skill.

The above analysis suggests that a combination of calibration by cross-validation and input weather data bias correction produces the most skillful deterministic simulations. For this model configuration (NCAL-BIC), the multi-model ensemble mean shows more statistically significant values of r_{os} (three) than any other single model (two, for the scnr, scwn, and ukmo models).

3.2. Probabilistic performance statistics

Two sets of probabilistic analysis have been carried out: (i) 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). This includes use of the Brier score, a mean square error which can be used with either probabilistic or deterministic forecasts. Note that the Brier score can not be compared across different Y_{cf} , since low Brier scores are favoured by low values of Y_{cf} . (ii) the Ranked Probability Score based on climatological terciles (i.e. three categories: below normal, normal and above normal), averaged over all grid cells and years (RPS). 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 of these analysis, the deterministic comparison is the ERA40 yield simulation. The analytical theory behind all the graphs and statistics presented in this section can be found in Staniski et al. (1989) and/or Brown (2001). All analyses use all available grid cells (10) for all available years (12).

Figure 5a compares the control simulation (GCAL-BIC-AUP) Relative Operating Characteristics (ROCs) of the multi-model ensemble and the best single model at two values of Y_{cf} (200 and 500 kg/ha). The lower yield threshold is clearly the most predictable of the two, with some events not being simulated by any ensemble member for the higher threshold. The best single model is more skillful than the multi-model ensemble at low false alarm rate. In terms of the Brier score and the mean Ranked Probability Score (table 3), the multi-model ensemble shows similar or greater skill than ERA40, and similar or worse skill than the best single model. Note, however, that the best single model varies between each case. Figure 5b 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 (figure 5c). The diagrams are included as they provide important information to complement the ROC curves: reliability is an indication of the consistency between observed frequency of occurrence and (probabilistic) forecast frequency of occurrence. The data available suggests that the multi-model ensemble is no less reliable than the best single model.

Figure 6 presents ROC curves for various model calibrations, with and without bias correction of the weather data. The corresponding reliability diagrams are presented in figure 7 and the corresponding values of the Brier score and the mean Ranked Probability Score (RPS) are presented in table 4. Simulations are more skillful and more reliable

for $Y_{cf}=400$ kg/ha than for $Y_{cf}=500$ kg/ha. RPSs for the multi-model ensemble tend to be better than ERA40 but similar to or worse than climatology (33% probability assigned to each category). Given that GLAM tends to more accurate under water-limiting conditions (therefore favouring prediction of crop failure over prediction of high yields), this is 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 (figures 7c&d). The few high probabilities that are predicted do not generally indicate greater certainty (most points in figures 7a&b lie below the 45 degree line).

The least skillful simulation overall is GCAL-ORI-AUP, although this does produce a reliable forecast at low probability (figures 7a&b). This may, however, be due to the greater sample size at low probability for this run (figures 7c&d). Bias correction of input data improves the simulations, although it can also remove the ability of any of the model ensemble members to simulate some of the observed events (figure 6). 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 on ensemble-mean data (NCAL) improves ROC skill further and also produces the lowest values of the Brier score and mean Ranked Probability Score (table 4). These values compare well with ERA40 values and with climatological RPS scores. NCAL multi-model 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 treats the multi-model ensemble as one model or as a sum of separately calibrated models (see section 2.1) 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 (figure 7) and improved (lower) mean Ranked Probability Scores (table 4). This suggests that, for probabilistic information, calibration of single-models may be less effective than calibration of the multi-model ensemble.

Overall, the results suggest that either bias correction of input data (BIC), or calibration via cross-validation on ensemble means (NCAL), or both, are required in order for the hindcasts to be skillful. Further, NCAL alone appears to provide a considerable improvement. The values of YGP calibrated on 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 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 YGP on 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 2.2 with the ERA40 data results in correlations between observed and simulated yields (r_{os}) rising in eight of the ten 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

To establish whether district-level yields could be simulated, runs were performed with one of the districts effectively representing the yield for the whole grid cell. The districts were chosen from grid cells where the inter-annual standard deviation of the area-averaged yield was not simulated well (low correlation coefficient) by ERA40. Using ERA40 (and trying both the sowing windows described in section 2.2) optimal values of YGP were found for the 1989–98 period and four statistically significant ($p < 0.05$) correlations emerged. These districts are marked on figure 1. The values of YGP obtained from the ERA40 analysis were used together with the bias-corrected ensemble weather data to perform the crop simulations.

The resulting multi-model ensemble mean simulations are compared to their area-averaged counterparts in table 5. Since 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 be useful on the subgrid scale.

4. Summary and discussion

4.1. Optimal calibration and bias correction methods

The analyses of correlations and RMSE presented in section 3.1 show that model skill can be dependent on the method of calibration. Calibration of the multi-model ensemble using ERA40 data was less successful than calibration on ensemble means via cross-validation. Bias correction tended to improve results only for ensemble means and only where the crop model was calibrated on ensemble-mean data. This bias-correction was towards ERA40 data; it is anticipated that bias-correction to observations (Challinor et al. 2004b) could improve results further. These results therefore suggest that the best model configuration for producing output based on the ensemble mean is cross-validation on the ensemble mean, with input data bias correction (i.e. NCAL-BIC).

For probabilistic analyses, the ROC curves presented in section 3.2 showed that bias correction of input weather data removed the ability of the model to simulate some events. The ensemble mean again emerged as the most favourable data on which to calibrate. However, separate calibration of the crop model for each single (GCM) model was shown to have no advantage over calibration using the multi-model ensemble mean.

Overall, the results suggest that estimates of the Yield

Gap Parameter 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 1966–86 and 1987–98. Some of the trends varied considerably between these two periods (e.g. over 60 kg ha^{-1} per year), suggesting that if this change were not accounted for, it would become important to calibrate YGP on data from a period closer to the study period.

4.2. Skillful probabilistic information

The system was successfully used to simulate crop failure, with more severe failures being more predictable. The best single model was not consistent across crop failure thresholds. However, there was no clear advantage to the multi-model approach in the prediction of crop failure. Preliminary efforts with a weighted multi-model approach (not presented) are encouraging; however, the time series is insufficiently long for this method to be used rigorously in this study.

Yield terciles proved harder to simulate, with skill being comparable to the use of climatology. Calibration on ensemble-mean data with no bias-correction of input weather data (NCAL-ORI) produced the lowest mean Ranked Probability Scores and for this simulation the multi-model ensemble produced lower values of RPS than any single model.

4.3. Skill relative to deterministic simulations

The multi-model ensemble was formed from a simple average of the 63 ensemble members. Even this simple configuration, with no weights, produced good correlations with observed yields — better than the ERA40 simulations overall. Averaging over ensemble members did tend to reduce the interannual standard deviation, however. The multi-model ensemble also showed more statistically significant correlations with observations than any other single model. The use of the August forecast to update the model did not significantly affect the results.

4.4. The impact of uncertainties

Two preliminary studies of uncertainty were made. A delayed sowing window significantly (in the statistical sense) improved correlations for one of the grid cells using ERA40. The corresponding multi-model 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, as in this case, skillful simulations result, then this is an advantage over the deterministic approach.

For two of the grid cells, significant correlations were found at the sub-grid scale where there were none at the grid-scale. 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 output shows skill would be a worthwhile endeavour.

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 ensemble mean information has been shown to perform better than calibration on reanalysis.

The implications for forecasting on short-to-medium timescales (a season to a decade) are as follows. Firstly, probabilistic forecasting: there is the potential for the probabilistic prediction of crop failure, defined by a given threshold yield value. Tercile forecasts may also become feasible as the skill of General Circulation Models (GCMs) increases. However, longer time series are needed in order to test the robustness of these results, and in particular the forecast reliability. Secondly, ensemble means showed skill in predicting interannual variability. Since bias correction to ERA40 showed the potential to increase this skill, improved GCM skill may well translate into improved deterministic yield prediction. Thirdly, uncertainties in model inputs are important, particularly for a model 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 multi-decadal (climate change) timescales are as follows. Firstly, 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 YGP on decadal timescales may be small. Hence climate change impacts studies may be formed from a number of plausible technology scenarios whilst keeping 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 visited, since this study has not provided a long enough time series for this. It is possible that some of the earliest and most severe impacts of climate change will come from the exceedance 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. Acknowledgements

The authors are grateful to the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) for the crop productivity data. AJC would like to thank Dr. Christopher Ferro for discussions on statistical inference. FJDR received

support from the EU-funded DEMETER project (EVK2-1999-00024).

References

- Brooks, R. J., M. A. Semanov, and P. D. Jamieson (2001). Simplifying Sirius: 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 online at <http://www.chmi.cz/meteo/ov/wmo/>.
- Camberlin, P. and M. Diop (1999). Inter-relationships between groundnut yield in Senegal, Interannual rainfall variability and sea-surface temperatures. *Theor. Appl. Climatol.* 63, 163–181.
- Cantelaube, P. and J. M. Terres (2004). Use of seasonal weather forecasts in crop yield modelling. *Tellus A*. DEMETER special issue.
- Challinor, A. J., J. M. Slingo, T. R. Wheeler, P. Q. Craufurd, and D. I. F. Grimes (2003, February). 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., T. R. Wheeler, J. M. Slingo, P. Q. Craufurd, and D. I. F. Grimes (2004a). Design and optimisation of a large-area process-based model for annual crops. *Agric. For. Meteorol.*. In press.
- Challinor, A. J., T. R. Wheeler, J. M. Slingo, P. Q. Craufurd, and D. I. F. Grimes (2004b). Simulation of crop yields using the era40 re-analysis: limits to skill and non-stationarity in weather–yield relationships. *J. Appl. Meteorol.*. Submitted.
- FAO/Unesco (1974). Fao/unesco soil map of the world, 1:5,000,000, ten volumes.
- Fischer, G., M. Shah, and H. van Velthuizen (2002). Climate change and agricultural vulnerability. Technical report, International Institute for Applied Systems Analysis. Available at <http://www.iiasa.ac.at/Research/LUC/>.
- Hsieh, W. W., B. Y. Tang, and E. R. Garnett (1999). Teleconnections between Pacific sea surface temperatures and Canadian prairie wheat yield. *Agric. For. Meteorol.* 96(4), 209–217.
- Jagtap, S. S. and J. W. Jones (2002). Adaptation and evaluation of the cropgro–soybean model to predict regional yield and production. *Agric. Ecosyst. Environ.* 93, 73–85.
- Landau, S., R. A. C. Mitchell, V. Barnett, J. J. Colls, J. Craigon, and R. W. Payne (2000). A parsimonious, multiple-regression model of wheat yield response to environment. *Agric. For. Meteorol.* 101, 151–166.
- Palmer, T. N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, M. Déqué, E. Díaz, F.-J. Doblas-Reyes, H. Feddersen, R. Graham, S. G. and J.-F. Gu/’er/’emy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and M. C. Thomson (2003). Development of a european multi-model ensemble system for seasonal to inter-annual prediction (demeter). *Bull. Am. Meteorol. Soc.*. submitted.
- Reddy, P. S. (Ed.) (1988). *Groundnut*. Krishi Anusandhan Bhavan, Pusa, New Delhi, India: Indian Council of Agricultural Research.
- Rijks, D., F. Rembold, T. N/’egre, R. Gomme, and M. Cherlet (Eds.) (2003). *Crop and Rangeland Monitoring in Eatern Africa for early warning and food security*. Joint Research Centre — Food and Agriculture Organisation. Proceedings of an International Workshop organised by JRC—FAO, Nairobi, 28–30 January 2003.
- Sinclair, T. R. and N. Seligman (2000). Criteria for publishing papers on crop modelling. *Field Crops Research* 68, 165–172.
- Southworth, J., J. C. Randolph, M. Habeck, O. C. Doering, R. A. Pfeifer, D. G. Rai, and J. J. Johnston (2000). Consequences of future climate change and changing climate variability on maize yields in the midwestern united states. *Agric. Ecosyst. Environ.* 82, 139–158.
- Stanski, H. R., L. J. Wilson, and W. R. Burrows (1989, July). Survey of common verification methods in meteorology. Research Report MSRB 89-5, Atmospheric Environment Service, Forecast Research Division., 4905 Differin Street, Downsview, Ontario, Canada M3H5T4. Second Edition. Available online at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Stanski_et_al/Stanski_et_al.html.
- Wheeler, T. R., P. Q. Craufurd, R. H. Ellis, J. R. Porter, and P. V. V. Prasad (2000). Temperature variability and the annual yield of crops. *Agric. Ecosyst. Environ.* 82, 159–167.

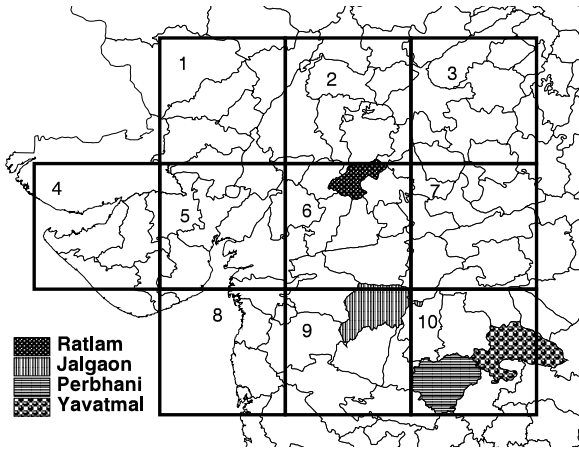


Fig 1. Map of the model grid (with corresponding cell numbers) and districts. The four districts referred to in section 3.4 are highlighted.

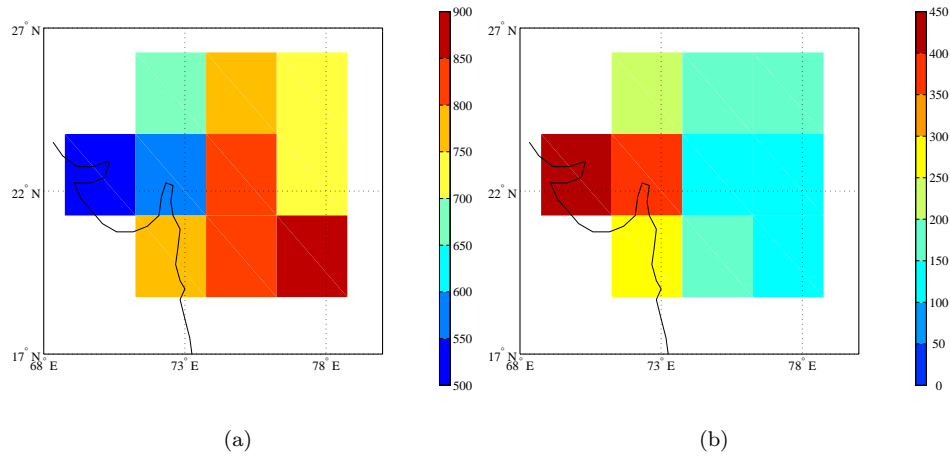


Fig 2. Observed mean (a) and standard deviation (b) of (linearly) detrended groundnut yields (kg/ha) in India, for the period 1987–1998, on the DEMETER grid.

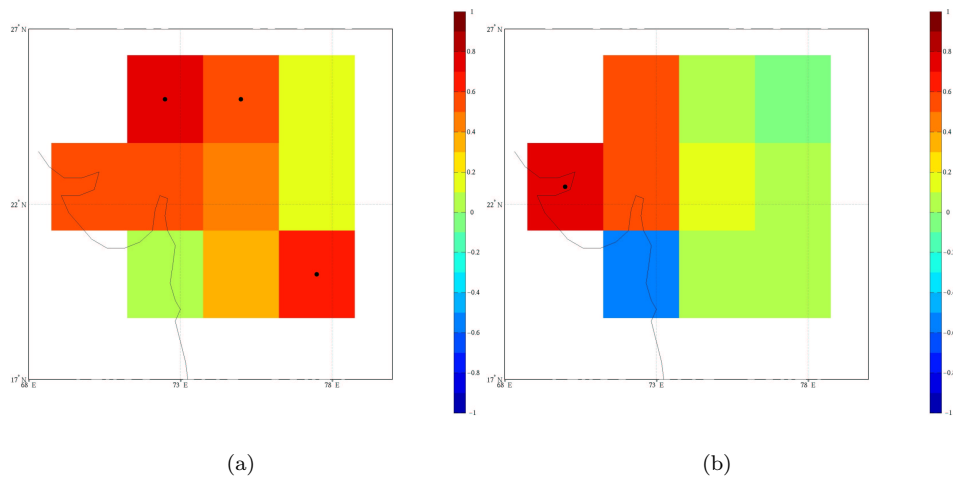


Fig 3. Correlation coefficients for observed and simulated yields (r_{os}) for the period 1987–1998, for a) the control run, GCAL-BIC-AUP and b) GCAL-ERA40. Statistically significant correlations are marked with a dot.

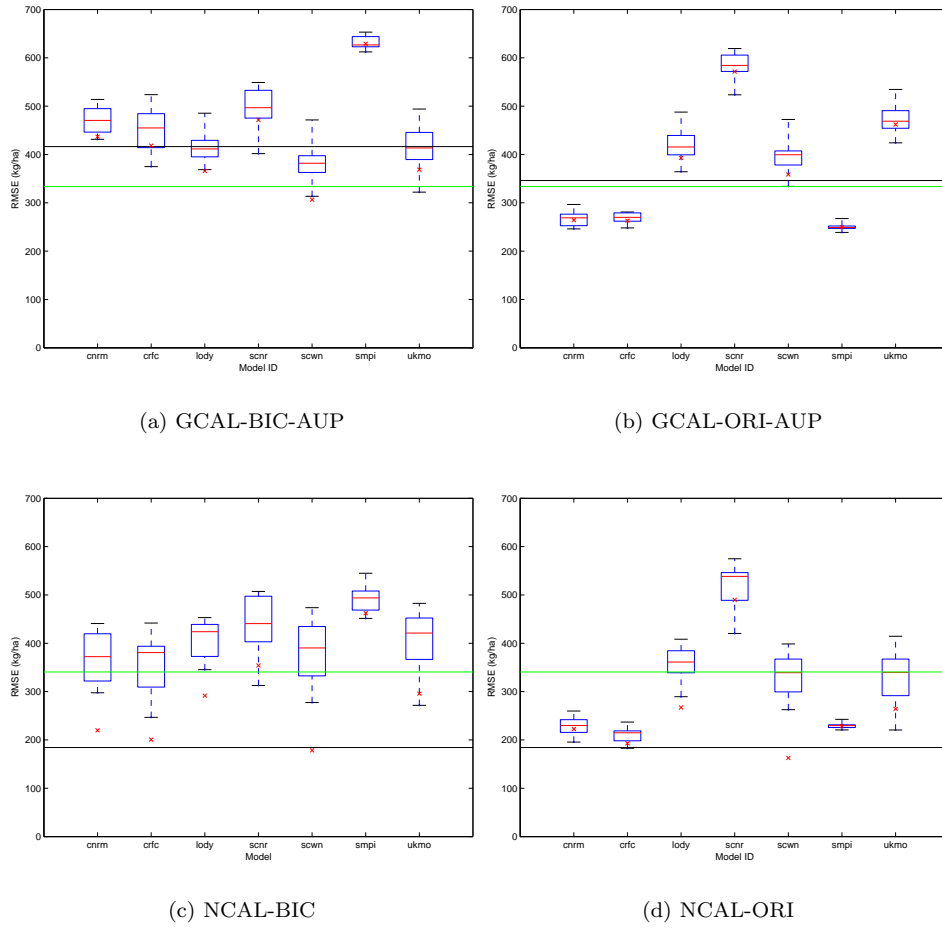
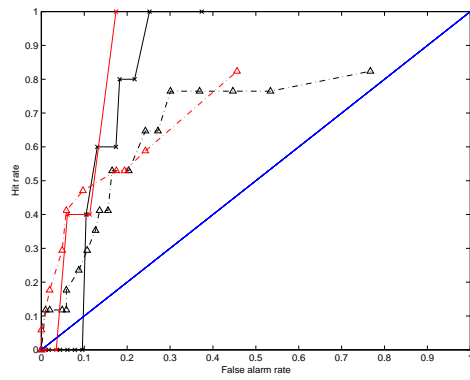
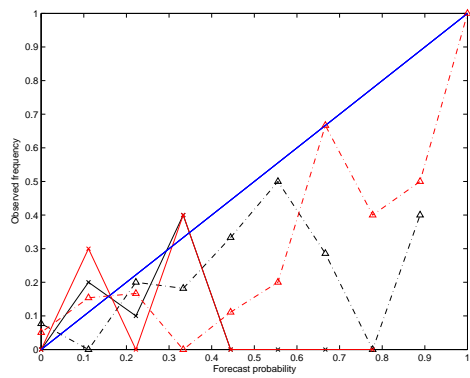


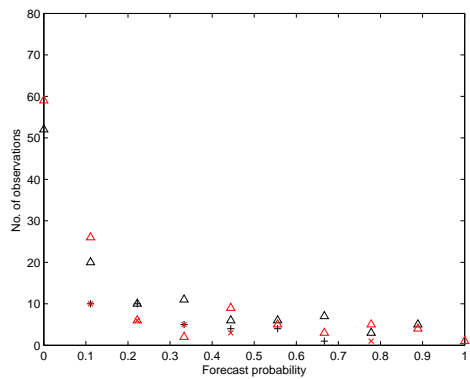
Fig. 4. Boxplots (the box shows the upper and lower quartiles, the red line shows the median and the whiskers show the full extent of the data), from four configurations of calibration (GCAL/NCAL) and bias-correction (BIC/ORI), of Root Mean Square Error (RMSE) for the seven models, and RMSE for ERA40 (green line) and the multi-model ensemble mean (black line). Red crosses show the RMSE of the mean of the individual models. All results shown are for grid cell 1 (see figure 1). For the NCAL runs, individual models are calibrated individually using the ensemble mean and the multi-model ensemble is calibrated using the multi-model ensemble mean. Table 1 describes the runs performed.



(a)



(b)



(c)

Fig 5. Evaluation of control simulation (GCAL-BIC-AUP) for the multi-model ensemble (black curves) and best single model (red curves) for crop failure thresholds of 200 (crosses) and 500 (triangles) kg/ha. (a) Relative Operating Characteristic (ROC) curves (skill is proportional to area bounded by the blue 1:1 line, the ROC curve, and the horizontal hit-rate=1 line), (b) reliability diagrams (skill is indicated by proximity to the blue 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.

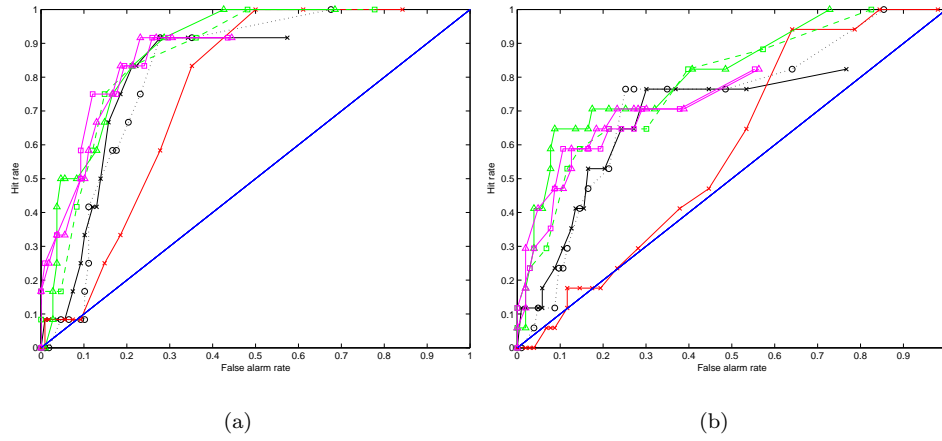


Fig 6. Relative Operating Characteristic (ROC) curves for the period 1987–89 for crop failure, defined as yield below (a) 400 kg/ha (12 observed events out of 120 data points) and (b) 500 kg/ha (17 observed events out of 120 data points). Black lines are for GCAL-BIC runs, with circles denoting the NUP run and crosses denoting the AUP run. Green lines show NCAL-ORI runs, magenta lines show NCAL-BIC runs; for both of these cases, triangles denote calibration using the single-model ensemble mean, and squares denote calibration using the multi-model ensemble mean. The red line is for the GCAL-ORI-AUP run. The blue lines shows the zero-skill baseline.

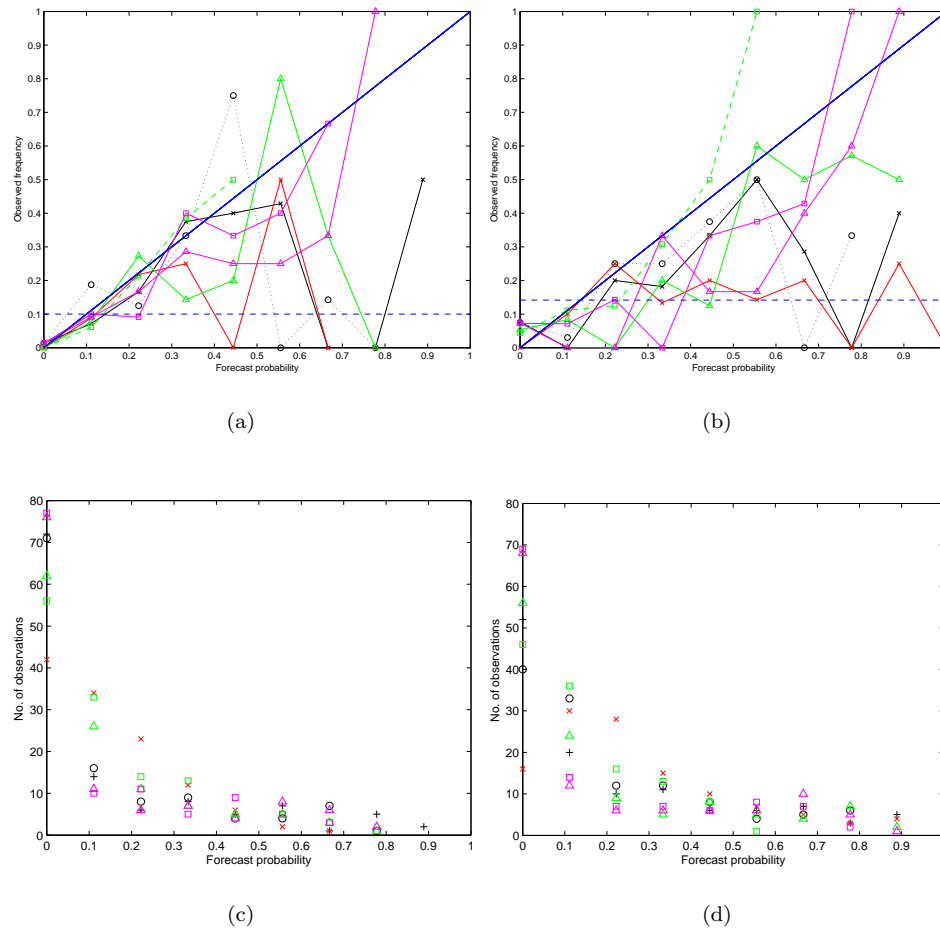


Fig 7. Reliability diagram for for the period 1987–89 for crop failure, defined as yield below (a) 400 kg/ha and (b) 500 kg/ha. The climatology is plotted as a dotted blue line and a perfectly reliable forecast as a solid blue line. The corresponding number of observations used for each point plotted are shown in (c) for 400 kg/ha and (d) for 500 kg/ha. The legend is exactly as figure 6: black lines are for GCAL-BIC runs, with circles denoting the NUP run and crosses denoting the AUP run. Green lines show NCAL-ORI runs, magenta lines show NCAL-BIC runs; for both of these cases, triangles denote calibration using the single-model ensemble mean, and squares denote calibration using the multi-model ensemble mean. The red 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.

[1]

Table 1. Naming convention for numerical experiments, together with a list of the experiments performed.

Runcode	Description
NUP	May hindcast used throughout (no August hindcast update).
AUP	May hindcast used for three months, then with August hindcast used.
GCAL	Calibration of YGP on 1966–86 yield data using ERA40 input data.
NCAL	Calibration of YGP by cross-validation on 1987–98 yield data (1987–92 data used to determine the 1993–98 YGP, and vice-versa). All NCAL runs use the August hindcast update (AUP).
MMC	NCAL method using the multi-model ensemble mean for calibration.
SMC	NCAL method using the respective single-model ensemble mean.
BIC	Input weather data has been bias-corrected to ERA40.
ORI	Original input weather data (no bias correction).
DSW	Delayed Sowing Window

Run	Comments
GCAL-BIC-NUP	This was the only NUP run performed.
GCAL-BIC-AUP	The control run: independent model calibration with bias-corrected input weather data.
GCAL-BIC-AUP-DSW	Control run with delayed sowing window (July 9th — August 7th).
NCAL-ORI-MMC	Model calibration based on ensemble means and yield data from the study period, using a single pair of YGP values for each model.
NCAL-ORI-SMC	Model calibration based on ensemble means and yield data from the study period, using a model-specific pair of YGP values.
GCAL-ORI-AUP	True hindcast: no 1987–98 data used.
NCAL-BIC-SMC	Full use of available 1987–98 data.
GCAL-ERA40	Benchmark comparison run with independent calibration.
NCAL-ERA40	Benchmark comparison run with calibration based on current yields.
GCAL-ERA40-DSW	Allows assessment of relative impact of delayed sowing window on control and benchmark runs.

Table 2. Correlation coefficient for observed and simulated yields for the multi-model ensemble mean yield (MME), and for the corresponding ERA40 run, for the four runs used in figure 4. 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 (\bar{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 $\sigma_y^{sim}/\sigma_y^{obs}=[1.50, 0.94, 0.91, 1.25, 0.94, 1.33, 1.37, 1.08, 1.30]$

Run	Correlation		$\sigma_y^{sim}/\sigma_y^{obs}$		$\bar{y}^{sim}/\bar{y}^{obs}$	
	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)

Table 3. The Brier score for two threshold crop–failure yield values (Y_{cf}) for the GCAL-BIC-AUP run. Values calculate using the multi–model ensemble (MME), the best single model (BSM) and ERA40 are shown. Four of these correspond to the four runs from figure 5. Also shown is the mean Ranked Probability Score (RPS; averaged over all available years and grid cells) for climatological yield terciles (bounded at 661 and 827 kg/ha).

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)

Table 4. The Brier score for two threshold crop–failure yield values (Y_{cf}) for the five runs used in figure 6 and for two ERA40 runs. Note that the Brier score is by definition lower for rarer events, so that values across different Y_{cf} should not be compared. Also shown is the mean Ranked Probability Score (RPS; averaged over all available years and grid cells) for climatological yield terciles (bounded at 661 and 827 kg/ha). The RPS of climatology (33.3% probability for each tercile) is 0.28.

Run	Y_{cf}	Brier	RPS
GCAL-BIC-NUP	400	0.101	0.51
GCAL-BIC-AUP	400	0.100	0.50
NCAL-ORI-MMC	400	0.073	0.26
NCAL-ORI-SMC	400	0.075	0.30
GCAL-ORI-AUP	400	0.093	0.63
NCAL-BIC-MMC	400	0.071	0.28
NCAL-BIC-SMC	400	0.077	0.31
GCAL-ERA40	400	0.092	0.60
NCAL-ERA40	400	0.083	0.44
GCAL-BIC-NUP	500	0.134	(0.51)
GCAL-BIC-AUP	500	0.143	(0.50)
NCAL-ORI-MMC	500	0.101	(0.26)
NCAL-ORI-SMC	500	0.106	(0.30)
GCAL-ORI-AUP	500	0.190	(0.63)
NCAL-BIC-MMC	500	0.110	(0.28)
NCAL-BIC-SMC	500	0.119	(0.31)
GCAL-ERA40	500	0.167	(0.60)
NCAL-ERA40	500	0.125	(0.60)

Table 5. Performance of multi–model ensemble mean yields 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 (One District Geocode, ODG) runs and the corresponding Area–Averaged Geocode runs (AAG).

Grid cell	District	Correlation		$\sigma_y^{sim}/\sigma_y^{obs}$	
		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
	Yavatmal	0.66			