Modelling European Agriculture with Climate Change for Food Security
— a FACCE JPI Knowledge Hub —
FACCEJPI
MACSUR

FACCE-MACSUR

CropM Deliverable C4.2.4: Information to support input data quality and model improvement

Mike Rivington[1]*, Daniel Wallach[2]

[1] The James Hutton Institute, Craigiebuckler, Aberdeen. AB15 8QH. United Kingdom
[2] INRA, Toulouse, France.

*mike.rivington@hutton.ac.uk

Instrument:	Joint Programming Initiative
Topic:	Agriculture, Food Security, and Climate Change
Project:	Modelling European Agriculture with Climate Change for Food Security (FACCE-MACSUR)
Start date of project:	1 June 2012
Duration:	36 months
Theme, Work Package:	CropM 1
Deliverable reference num.:	D-C4.2.4
Deliverable lead partner:	James Hutton Institute
Due date of deliverable:	month 36
Submission date:	2015-06-25
Confidential till:	-

Revision	Changes	Date
1.0	First Release	2015-06-25

Abstract/Executive summary

Data quality is a key factor in determining the quality of model estimates and hence a models' overall utility. Good models run with poor quality explanatory variables and parameters will produce meaningless estimates. Many models are now well developed and have been shown to perform well where and when good quality data is available. Hence a major limitation now to further use of models in new locations and applications is likely to be the availability of good quality data. Improvements in the quality of data may be seen as the starting point of further model improvement, in that better data itself will lead to more accurate model estimates (i.e. through better calibration), and it will facilitate reduction of model residual error by enabling refinements to model equations.

This report sets out why data quality is important as well as the basis for additional investment in improving data quality.

Table of Contents

Introduction

Data quality is a key aspect of improving the utility of models. A model becomes redundant if the quality of the input data leads to unreliable estimates. The lack of location-specific explanatory variable data for spatially and temporally variable entities means that a models' site-specific estimates have potentially large and unquantified uncertainties. Often models are run with data that have a range of quality. The challenge then becomes to discover the consequences of this range in data quality on the utility of the models' estimates. This issue can be further divided into sections: quality of data used to construct the equations used in a model; data used for calibration, and data used for testing and evaluation purposes.

The key message in this report is that model improvement can be achieved through improving the quality of data used to construct, calibrate and run models. Identifying the hierarchy of importance of explanatory variables thus informs of how research efforts can be focused on specific data sources to improve quality (Aggarwal 1995). Methods have been developed to guide selection of data for crop modelling purposes (i.e. Grassini et al 2015).

A crop model can be succinctly written as $f(X;\theta)$ where X are the *explanatory variables* (in general daily weather, soil characteristics, initial soil conditions and management), θ are the parameters in the model and f is the function that translates X and θ into the model outputs (for example yield). Thus another division in data is between the explanatory variables, as either input into a model (X) or parameter values (θ) used. Each of the above terms involves uncertainty.

It is useful to recall some of the key points about explanatory variables, parameters and model uncertainty when consider the role of data quality (see MACSUR CropM deliverable reports C4.1.1 – 'Development of a common set of methods and protocols for assessing and communicating uncertainties', C4.2.2.- 'A framework for assessing the uncertainty in crop model predictions' and C4.2.3 – 'Quantified Evidence of Error Propagation' for further details).

Explanatory variables

There can be uncertainty in the *explanatory variables*, which reflects the range in data quality. The weather data may be from a weather station that is at some distance from the field in question, so those data are only an approximation to the true weather data (Rivington et al 2006). Soil characteristics may just be estimated (for example using a pedotransfer function to estimate water holding capacity from texture). Initial soil conditions are often not available, so are estimated using the model or expert opinion. Management may not be fully recorded, for example one might have only total nitrogen application and not the dates of application.

One can often quantify the uncertainty in *explanatory variables* based on past data (for example, cases where both pedotransfer functions have been used and true water holding capacity has been measured) or expert opinion (for example, it may be known that in the population considered nitrogen is generally applied in three applications, and the range of dates may also be known). Empirical data though may be limited in availability and spatial and temporal representativeness.

There are also cases where it is difficult or impossible to quantify the uncertainty in explanatory variables. When using crop models to evaluate the impact of climate change, for instance, the explanatory variables include future weather, which is unknown and whose uncertainty is difficult to estimate. In this case one may simply replace the uncertain weather by a finite number of possible scenarios.

Parameters

Parameter uncertainty is particularly complex. The first question is "what are the true parameter values?". Wallach, D. (2011) suggests that the true parameter values must be defined not in the context of the overall crop model, but in the context of the individual equations within the model. Suppose that an individual equation is $\hat{Y}^{ind} = f^{ind}(X^{ind}, \theta^{ind})$ where the superscript "*ind*" emphasizes that this refers to one of the equations in the model. The true parameter values are then the parameters such that $E(\varepsilon) = 0$ for all X^{ind} , where $\varepsilon = Y^{ind} - \hat{Y}^{ind}$ is the difference between the measured and modelled values. This is just the standard definition of the true parameters in regression. Thus the issue of data quality used to construct the equations is paramount.

In many cases the individual equations in a crop model have been studied individually. An example would be the relation between biomass accumulation (ΔB) and intercepted photosynthetically active radiation (IPAR), usually assumed to be linear: $\Delta B = RUE * IPAR$, where RUE (radiation use efficiency) is a parameter. The uncertainty in parameter values can often be taken from the range of values found in the literature, in studies of the individual equations. This range is often large. For the parameters in the SUCROS87 and LINTUL models, Metselaar (1999) found an average coefficient of variation of 38%. A crop model may also have parameters which are constructs of the model, and for which no values based on studies of the individual equation exist. The uncertainty in these parameters can then only be very approximate.

When the model is calibrated, that is some of the parameters are changed to give a better fit to the data (which has uncertainty associated with it), this changes the uncertainty in the parameters used to calibrate the model. Those now should be considered as *calibration parameters*, and their true values are now the values that minimize overall model error. This is different from the true parameters for the individual equations. These parameters have now become adjustment factors for the overall model. Their uncertainty is related to the data used for calibration, just as in classical regression.

In general we only approximate the relationships in the individual equations in a model. There may be uncertainty in these relationships. In some cases, it may be possible to quantify this uncertainty. If for example, one can chose a particular function to represent a relationship, but this choice was somewhat arbitrary and other functional forms could have been chosen, then the uncertainty could be represented by equal probability on each form. In other cases, it may be very difficult to quantify this uncertainty, because one simply doesn't know what the range of plausible functions is.

Residual error

Residual error measures the difference between the model predictions and the measured value. Even if one has the correct explanatory variables, the correct parameters and the correct functional form, in general a model does not explain all the variability in the response and so residual error will not be zero. In that case, residual error measures how much of the variability in the response is left unexplained by the explanatory variables of the model. If there are errors in the explanatory variables, parameters or functions, residual error will be larger.

Importance of data

A survey of crop modellers around the world showed that improving the mathematical representation within a model was less important than being able to provide more and better calibration data and other means of model improvement (see Fig. 1) (Rivington and Koo 2011). This implies that the formalisation into equations and model code is seen as a lower priority in model improvement than increasing the supply of accurate data.

However, modellers may be limited to refinements within the model, having less capacity to improve data quality (reducing errors in explanatory variables). It is these equations and the structure of the model (the connectivity between model components), that determine how errors present in input data combine with limitations of the formalisation into equations to produce the overall cumulative error.

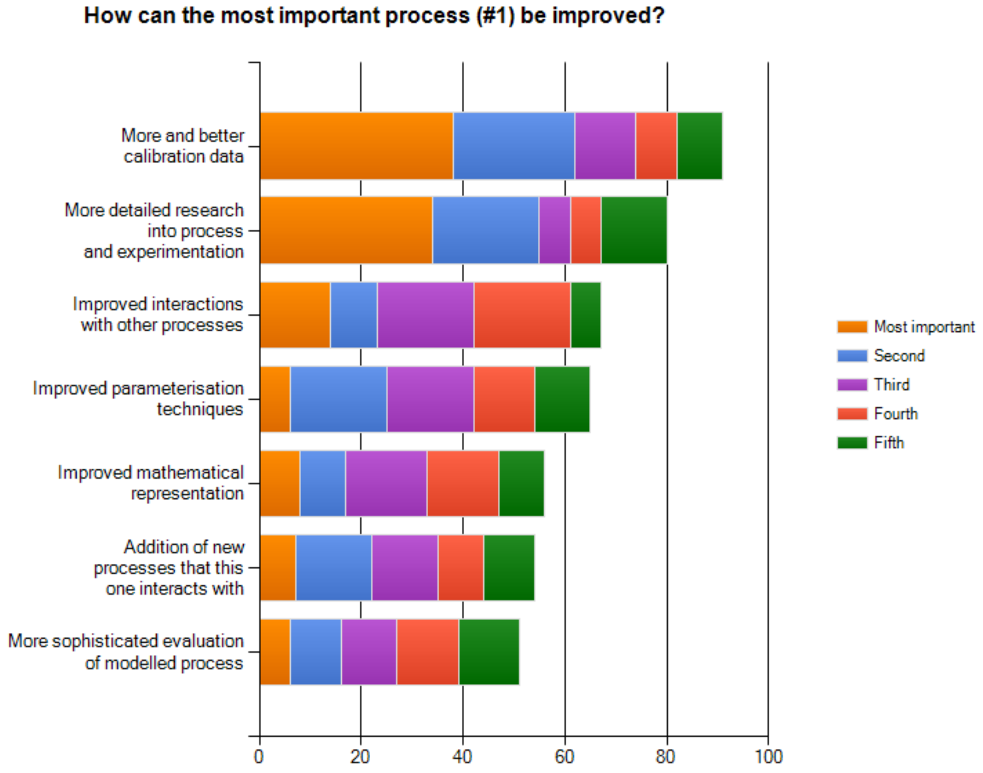

Figure 1. Survey of modellers' views on how the most important modelling processes within a model can be improved (x axis is the count of individual responses) (Rivington and Koo 2011).

Illustration of a range in data quality

The issue of data quality is closely related to that of error propagation. Data that only partially represents a phenomena or entity will introduce errors into a model, which can then be further propagated due to the range of estimates (Y) made by individual equations, constituting model residual error. Figure 2 illustrates this issue:

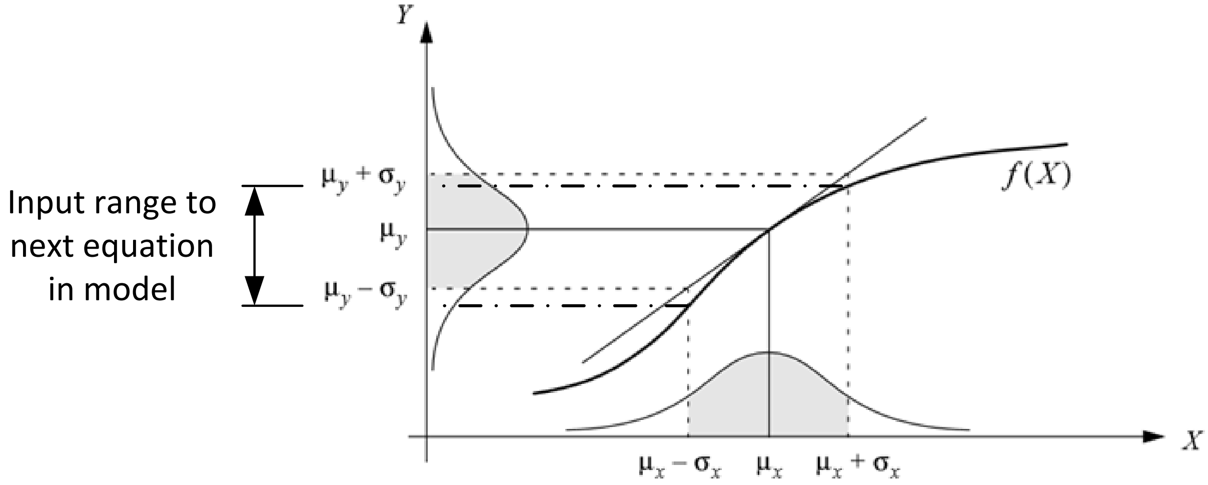

Figure 2. A one-dimensional case of error propagation related to input data quality (adapted from Arras 1998), showing probability interval (68%, shaded area) for *Y* given a probability distribution range and probability area of an explanatory variable *X*. Note however that the non-linear *f(X)* would lead to an asymmetric (non Gaussian) distribution of *Y* (dot-dash line). The estimate *Y* is often then used as input into another equation within a model.

The range of *Y* is the predictive uncertainty. Clearly the closer the data is to μ_x the less the magnitude of error introduced to the model. However, we also know that the distribution of errors in input data is unlikely to be Gaussian, hence the shape of the curve for *X* in Fig 2 is likely to be skewed, again altering the distribution of *Y*. If μ_x - σ_x in Figure 2 represents the limit in acceptable data quality (i.e. in representing an observation), then it can be seen the resulting estimate falls outside of the 68% probability range (μ_y - σ_y), thus introducing an unacceptable error.

The situation rapidly becomes even more complex in non-linear systems made up of multiple *X* and *Y* (such as crop growth, i.e. Fig. 3). This places a further emphasis on the need for accuracy of *X* as an input explanatory variable and calibration of parameters. The model above may be a sub-model of a larger set within a model, where Y is carried forward to be used in other calculations within the model.

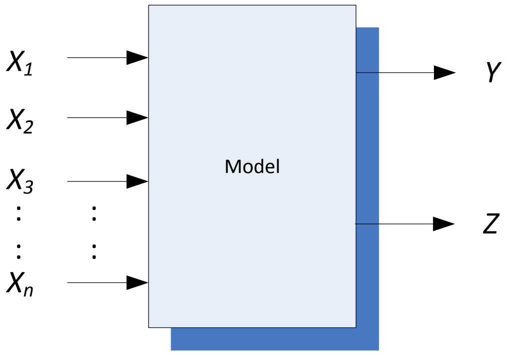

Figure 3. Error propagation in a multi-input multi-output model.

Thus between Figures 2 and 3, it is possible to envisage how data quality issues feed through the model process. However, the behaviour of an error can be either positive, negative, additive or multiplicative or compensatory (cancelling out previous introduced errors), or a combination of all, within the overall modelling process. Where the combination of propagation types amplify the absolute error in an estimate of interest, then identifying the absolute error becomes easier but when propagation leads to compensatory errors, the absolute error is less obvious. This leads to a situation of ‘right result but for the wrong reasons’. Diagnosis of error cause can help identify whether it is a data quality or model residual error issue.

Examples

The following examples focus on the role of weather data as explanatory variables used in crop models. Heinmann et al. (2002) showed that the accuracy of rainfall observations is critical for the simulation of yield and that the variability of simulated estimates is directly correlated to the accuracy of model inputs. Similarly, Nonhebel (1994) found that inaccuracies in solar radiation of 10% and temperature of 1°C resulted in yield estimation errors of up to 1 t ha^{-1}, and up to 10 days difference in vegetative period between emergence and flowering. Aggarwal (1995) tested the relationships between the uncertainty in crop, soil and meteorological inputs with the resulting uncertainties in estimates of yield, evapotranspiration and crop nitrogen uptake, within a deterministic crop growth model. It was then possible to identify the uncertainty importance of an input for a given scenario, concluding that in rain fed environments soil and weather inputs were dominant over crop parameters in introducing uncertainty.

This emphasizes the importance of data quality (accuracy of measurement), as well as site-specific representation. The basic concept of data quality introducing errors to models illustrated here applies to all type of explanatory variables.

Distance to nearest meteorological station.

One illustration of data quality is reflected in the proximity of the nearest meteorological station to the place where a model is applied, which may not have observed weather data, or only partial coverage. Often solar radiation (a key explanatory variable in crop models) is not observed at met stations, and has to be estimated, i.e. using other weather variables such as sunshine duration or temperature. Figure 4 illustrate the range of data quality associated the methods to estimate solar radiation at the place of model application (PoMA), and what the decay in quality is when having to use data from met stations at increasing distance from the PoMA. In this example the nearest alternative met station (Prestwick) has relatively similar precipitation and temperature, but no solar radiation data. The next nearest station (Esk, Esdalemuir) has larger RMSE than stations further away.

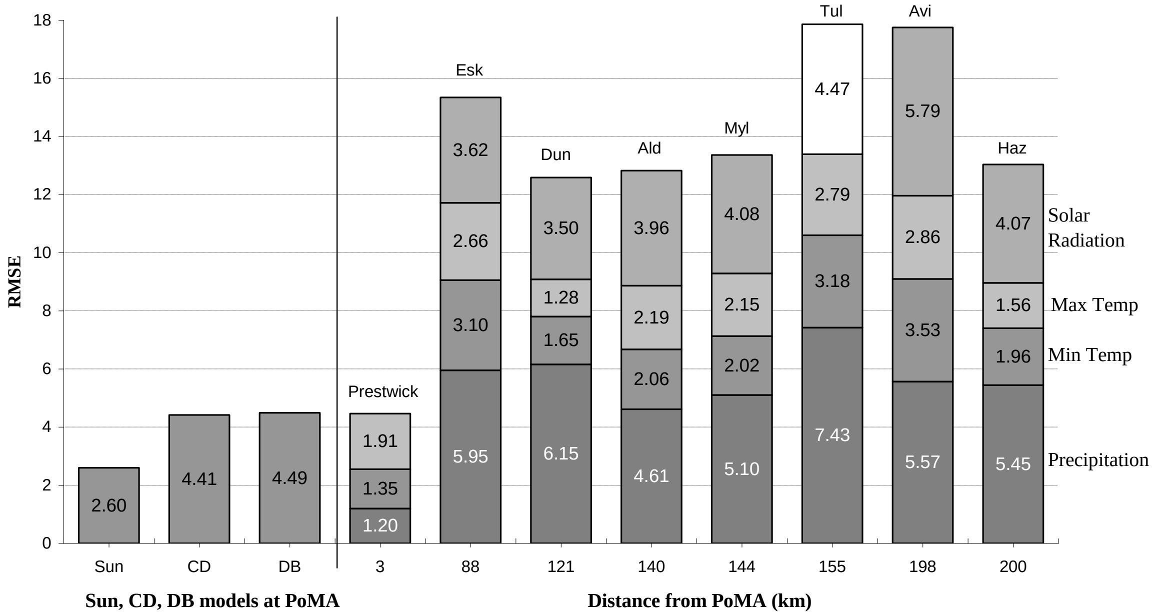

Figure 4. Accumulated total RMSE for precipitation (mm), max and min air temperature (°C), solar radiation (MJ^{-1} m^{-2} day^{-1}) at increasing distance (km) from place of model application (PoMA = Auchencruive, Scotland), compared with solar radiation estimated from: observed sunshine duration (Sun), temperature using Campbell-Donatelli method (CD) and temperature using Donatelli-Bellocchi method (DB) model solar radiation output RMSE (MJ^{-1} m^{-2} day^{-1}). Unpublished – see Rivington et al 2006.

This variation in weather data quality (in this example, how well the alternative met station data represents the place of model application) has an impact on the estimates made aby a crop model.

Table 1. Consequences of alternative meteorological station data on estimates of spring barley yield made by the CropSyst model compared to yield estimates made with observed weather data at Auchencruive. This shows that the nearest alternative station is not necessarily the best in terms of explanatory variable quality and impact on model estimates.

Site (PoMA)	Auchencruive		
Substitute	Esk	Dun	Ald
Distance (km)	88	121	140
Total Yield Diff / n (t/ha)	1.88	0.48	0.24
Mean Yield Diff (t/ha)	1.88	0.45	0.24
Absolute Diff (t/ha)	1.88	0.61	0.33
Max over est. (t/ha)	2.97	1.31	0.77
Max under est. (t/ha)	1.34	-0.49	-0.33
SE (t/ha)	0.27	0.35	0.32
n (years)	6	18	17

Same weather source in three models

A second example shows how using the same weather data source in three different models (DSSAT, APSIM and CropSyst), produces a range of estimates. This examples illustrates both the differences in model estimates due to different weather data sources (explanatory variable quality) and model residual error, as divergence is seen in the model estimates when using the same weather data (Cammarano et al 2015).

Figure 5 shows results from running the three models using Regional Climate Model (RCM) hindcast (1960-1990) and future (2030 -2060) estimated data. This RCM hindcast and future data was also bias corrected against observed weather data, to give four weather data sources. Results are compared against the three models estimates made using observed weather data. All other variables are kept the same.

What Fig. 5 shows is that the three models produce very similar estimates of anthesis and maturity dates using the same weather data source, but have a range of estimates responses for yield and cumulative evapotranspiration.

The challenge then becomes in conducting a diagnosis of why each model produces different estimates with the same weather data. A graphical representation of why differences occur in given in Figure 6.

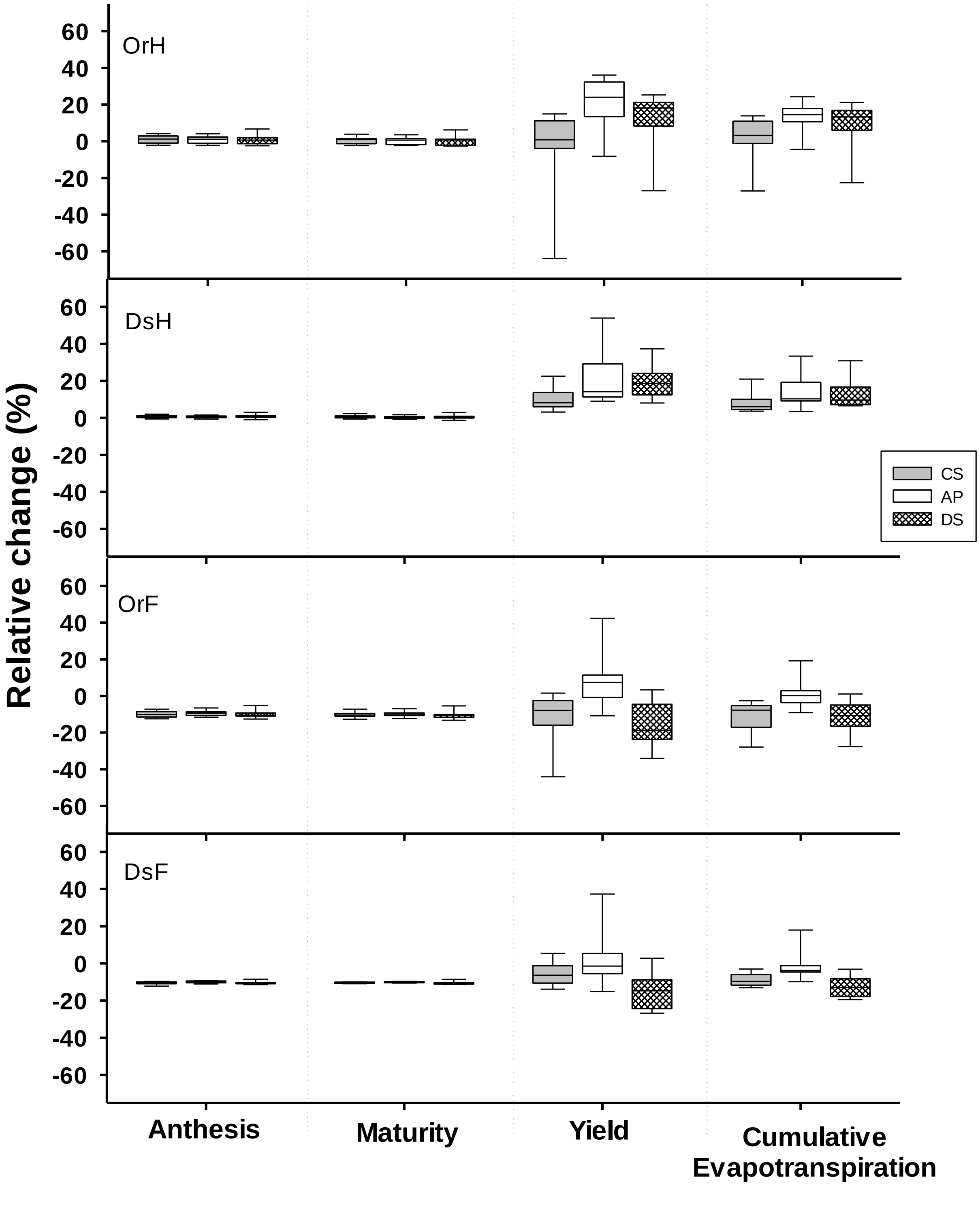

Figure 5. Relative change (%) for simulated anthesis, maturity, yield, and cumulative evapotranspiration between observed weather data and original Regional Climate Model hindcast (OrH, downscaled RCM hindcast (DsH), original RCM future projection (OrF), and downscaled RCM future projection (DsF) for CropSyst (CS, grey bars), APSIM (AP, white bars), and DSSAT (DS, grey patterns bars). For each boxplot horizontal lines represent, from the bottom to the top, the 10th percentile, 25th percentile, median, 75th percentile and 90th percentile of simulations.

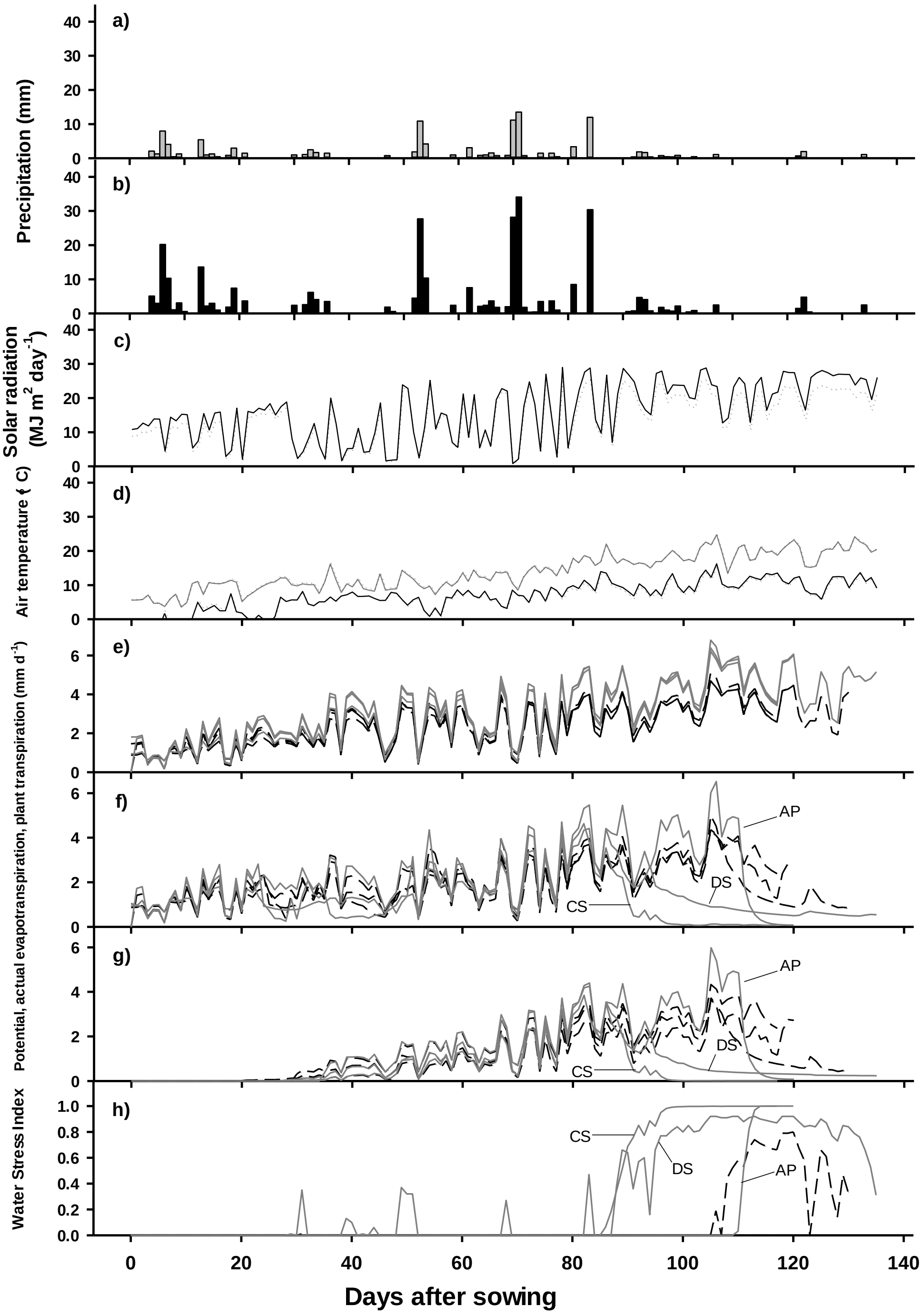

Figure 6. Comparison of Regional Climate Model original hindcast (OrH, grey bars and grey lines) and downscaled hindcast (DsH, black bars and black lines) estimates of *a, b*) precipitation; *c*) solar radiation (S_o); *d*) maximum and minimum temperatures; for one simulated growing season at Mylnefield, Scotland (1973) and their impact on CropSyst (CS), APSIM (AP), and DSSAT (DS) *e*) potential evapotranspiration; *f*) actual evapotranspiration; *g*) plant transpiration; *h*) water stress index effects on growth.

Figure 6 shows that each model starts to diverge in key estimates such as potential and actual evapotranspiration, thus affecting water stress (and other calculations). CropSyst and DSSAT showed higher values of water stress index about 90 days after sowing while APSIM about 111 days after sowing. That caused a different pattern of simulated plant transpiration (Fig. 6g, grey lines) between the models, with AP showing higher daily values of transpiration decreasing sharply when its water stress index becomes severe.

Discussion

An improvement in data quality is a clear route to improving the utility and overall quality of models and as a way of reducing uncertainty associated with a primary source. Improving data quality will not directly reduce model residual error, but would instead enable better equation formulation, calibration and testing, leading to a reduction in overall error.

Many parameter values are estimates based on literature searches or expert opinion, hence there is a need to increase the quality of measurement and range of variables measured during experimentation and increasing their availability. However, this may be beyond the control of modellers, hence the issue is more about how we integrate modelling with experimentation.

More immediate gains may be made in better understanding the hierarchy of importance of the different explanatory variables (X) and parameters (θ), so that research efforts can be focused on them, and at the same time develop error propagation protocols to better understand how errors behaviour changes during their passage through a model. This also indicates the priority for improving calibration parameters. Sensitivity analysis can inform us of how the model responds to different X and θ, and can help identify what a value range may be of either one. This information is useful to guide the hierarchy of importance of X and θ, again helping to target collections efforts.

Improving access to data is also a key step in better calibrating models. A considerable part of a modelling project is in accessing and organising data for calibration and testing purposes. Similarly, synchronising data, in space and time, will help construct simulations more easily. In the past modellers have rarely shared data used in simulations, but with MACSUR and AgMIP, this is changing. A structured approach to archive explanatory variables and parameters within a data base would facilitate easier.

Next Steps

There needs to be better integration between experimentation scientists and modellers. Better dialogue and funding can help facilitate appropriate data collection and organisation to meet the needs of experimentalists and modellers. Consideration also needs to be given to the spatial extent to which data is observed, as there are large gaps where little or no data exists.

MACSUR phase 2 should aim to develop a set of proposals to link with experimental researchers better, and to make better use of existing data. The original aims set out in the MACSUR phase 1 proposal are still valid: that of developing protocols to test equations, responses to extreme weather variables, responses to elevated CO_2 levels etc.

A key lesson learned in MACSUR phase 1 is that research to better understand uncertainty and develop methods to reduce it often come second place to applications of models to address more immediate questions. The overheads of developing and testing methods for uncertainty reduction can be high, particularly in terms of researcher time, hence there is a specific need for targeted resourcing of uncertainty based research.

Acknowledgements
This paper is a contribution to the FACCE MACSUR knowledge hub.

The work was funded by The Scottish Government Rural and Environment Science and Analytical Services Division (RESAS), and the French National Agricultural Research Institute (INRA).

References

Aggarwal, P.K., 1995. Uncertainties in crop, soil and weather inputs used in growth models - implications for simulated outputs and their applications. Agricultural Systems 48, 36–384.

Arras, K.O. 1998. An Introduction to Error Propagation: Derivation, Meaning and Examples of Equation *Cy* = $F_xC_xF_xT$. Technical Report No. EPRL-ASL-TR-98-01 R3. Autonomous Systems Lab, Institute of Robotic Systems, Swiss Federal Institute of Technology Lausanne. http://www.nada.kth.se/~kai-a/papers/arrasTR-9801-R3.pdf

Cammarano, D. Rivington, M. Matthews, KB, Miller, DG and Bellocchi, G. 2015. Implications of climate model biases and downscaling on crop models simulated climate change impacts. To be submitted.

Grassini, P. van Bussel, JGV. Van Wart, J. Wolf, J. Claessens, L. Yang, H. Boogaard, H. de Groot, H. van Ittersum, MK. Cassman, KG. 2015. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. Field Crops Research 177, 49-63.

Heinmann, A.B., Hoogenboom, G., Chojnicki, B., 2002. The impact of potential errors in rainfall observations on the simulation of crop growth, development and yield. Ecological Modelling 157, 1–21.

Metselaar, K. (1999, February 2). Auditing predictive models: a case study in crop growth. WAU Dissertation no. 2570.

Nonhebel, S., 1994. Inaccuracies in weather data and their effects on crop growth simulation results. I. Potential production. Climate Research 4, 47–60.

Rivington, M. and Koo, J. 2011. Report on the Meta-Analysis of Crop Modelling for Climate Change and Food Security Survey. Consultative Group on International Agricultural Research / Earth Systems Science Partnership Climate Change, Agriculture and Food Security Challenge Program.
http://www.macaulay.ac.uk/climatechange/CC_CCAFS.php
http://ccafs.cgiar.org/content/publications
http://labs.harvestchoice.org/2011/02/meta-analysis-of-crop-modeling-for-climate-change-and-food-security/

Rivington, M., Matthews, K.B., Bellocchi, G. and Buchan, K. (2006). Evaluating uncertainty introduced to process-based simulation model estimates by alternative sources of meteorological data. *Agricultural Systems* **88**, 451-471.

Wallach, D. (2011). Crop model calibration: A statistical perspective. Agronomy Journal, 103, 1144–1151.