Climatic Change (2016) 139:551–564
DOI 10.1007/s10584-016-1803-1

Lessons from climate modeling on the design and use of ensembles for crop modeling

Daniel Wallach[1] • Linda O. Mearns[2] • Alex C. Ruane[3] • Reimund P. Rötter[4] • Senthold Asseng[5]

Received: 12 April 2016 / Accepted: 31 August 2016 / Published online: 15 September 2016

Abstract Working with ensembles of crop models is a recent but important development in crop modeling which promises to lead to better uncertainty estimates for model projections and predictions, better predictions using the ensemble mean or median, and closer collaboration within the modeling community. There are numerous open questions about the best way to create and analyze such ensembles. Much can be learned from the field of climate modeling, given its much longer experience with ensembles. We draw on that experience to identify questions and make propositions that should help make ensemble modeling with crop models more rigorous and informative. The propositions include defining criteria for acceptance of models in a crop MME, exploring criteria for evaluating the degree of relatedness of models in a MME, studying the effect of number of models in the ensemble, development of a statistical model of model sampling, creation of a repository for MME results, studies of possible differential weighting of models in an ensemble, creation of single model ensembles based on sampling from the uncertainty distribution of parameter values or inputs specifically

Highlights
?•Ensembles of crop models can improve uncertainty estimates, impact projections, and collaboration.
•The climate modeling community has identified and studied many of the questions related to design and analysis of ensembles.
•Much of the climate experience could be adapted to crop modeling.

Electronic supplementary material The online version of this article (doi:10.1007/s10584-016-1803-1) contains supplementary material, which is available to authorized users.

✉ Daniel Wallach
Daniel.Wallach@toulouse.inra.fr

[1] INRA, UMR AGIR, Castanet Tolosan, France

[2] National Center for Atmospheric Research, Boulder, CO, USA

[3] National Aeronautics and Space Agency Goddard Institute for Space Studies, New York, NY, USA

[4] Georg-August-Universität Göttingen, Göttingen, Germany

[5] University of Florida, Gainesville, FL, USA

oriented toward uncertainty estimation, the creation of super ensembles that sample more than one source of uncertainty, the analysis of super ensemble results to obtain information on total uncertainty and the separate contributions of different sources of uncertainty and finally further investigation of the use of the multi-model mean or median as a predictor.

Keywords Model ensembles · Crop models · Climate models · Model weighting · Super ensembles

1 Introduction

In modeling complex systems, it is important to have a measure of uncertainty in simulated values (Tebaldi and Knutti 2007). A promising way to evaluate this uncertainty is through working with ensembles of models; the variability among the models in the ensemble is a measure of our uncertainty as to how to model the system. Ensembles allow one to obtain a probability distribution instead of a point prediction (Harris et al. 2010). Furthermore, it has been empirically observed in many fields that ensemble averages or medians often better reproduce observations than even the best individual model (Hagedorn et al. 2005; Tebaldi and Knutti 2007; Palosuo et al. 2011; Martre et al. 2015). Additional benefits from working with ensembles of models arise from the closer collaboration between modeling groups.

Climate modeling and crop modeling are two fields in which multiple groups have developed different models to represent the same complex system. In both fields there is major interest in the uncertainty of simulations and the potential of ensemble statistics to improve predictions or projections. (As used here, "projection" is a potential future evolution of some quantity (e.g., future temperature), given assumptions about the future state of the world, while "prediction" (or forecast) is usually a definite statement or statistical estimate of an expected occurrence of an event in the future. Projections are generally more uncertain than predictions). It is therefore not surprising that both fields have major programs related to model intercomparison and the construction of multi-model ensembles (MMEs). However, the climate modeling community began working with MMEs with the Atmospheric Model Intercomparison project in 1989 (Gates et al. 1999), whereas global collaboration to create crop multi-model ensembles began in 2011 with the Agricultural Modeling Intercomparison and Improvement project (Rosenzweig et al. 2013), though more limited intercomparison studies predated that project (e.g., Jamieson et al. 1998; Mearns et al. 1999). Progress in the use of ensembles of both climate and crop models in studies of climate change is discussed in (Challinor et al. 2013). More background material on climate models and crop models is presented in supplementary material.

Many of the methods of working with ensembles, and many of the problems that arise, are common to the fields of climate and process-based crop modeling. Knutti (2010) explicitly suggests that some of the recommendations for working with ensembles of climate models could apply to other types of numerical models. Given the longer experience of the climate modeling community, it seems worthwhile to examine how their experience with ensemble modeling could contribute to crop modeling.

The objective of this paper is to identify questions and approaches related to developing and using model ensembles that have been studied in the climate modeling literature and are relevant to crop models. This should hopefully accelerate the progression of the crop model community in taking advantage of this useful diagnostic approach.

2 Construction of model ensembles

The basic idea behind an ensemble is to carry out simulations using multiple models and/or multiple variants of a single model, with an aim to examine the uncertainty associated with a modeling exercise. In general model structure, model parameters and model inputs are all uncertain, and the ensemble can be created to represent the uncertainty in any of those, either singly or in combination.

2.1 Choosing the participants in a multi-model ensemble (MME)

The choice of models clearly affects any conclusions based on a MME. A major question then is what models to include, or more generally what are the criteria for inclusion. (Knutti 2010).

To date, most climate model ensembles are "ensembles of opportunity" (Tebaldi and Knutti 2007). Any modeling group that wishes to participate can do so, assuming the model under consideration has an accepted set of model components. One problem is that some of the candidate models may be very poor determinants for hindcasts, and including them in the ensemble may unrealistically inflate the uncertainty estimations (Knutti 2010). However, an a priori determination of which models are poor is a very complex and difficult task, since this will largely depend on the particular metrics used, the variables being evaluated and the regions being investigated. It is also possible that all the models in an ensemble are 'wrong' in so far as they are missing important components of the climate system (e.g., submodels of glaciers). In fact we do not know how to thoroughly evaluate climate models for the sake of eliminating poor model performers (IPCC 2013a). However, the recently launched CMIP6 will require some verifiable test runs to demonstrate model quality (Meehl et al. 2014).

The relatively short experience with ensembles of crop models has also involved "ensembles of opportunity". The need for quality criteria to exclude poorly performing models has been discussed (Palosuo et al. 2011; Rötter et al. 2012). Asseng et al. (2013) removed those models with the highest and lowest 10 % of simulated values in some of their analyses, but do not discuss that choice. Eliminating poorly performing models may be more important for crop than for climate MMEs, since in general crop models are less thoroughly evaluated. For example, Tubiello and Ewert (2002) found that the models most used for analyzing the effect of increased CO_2 are those that have been evaluated the least using enhanced CO_2 experiments, although this situation has evolved in the interim.

Following the example of the climate modeling community, it seems worthwhile to propose standardized tests for candidate models of crop MMEs. These tests should allow comparison with observations. Going further, it would be of interest to propose and test guidelines, based on standardized tests, for including (or excluding) models in crop MME studies.

2.2 Evaluating the degree of relatedness of the models in a MME

Including closely related models in an ensemble brings the risk of giving undue weight to a single basic modeling approach. A first problem is determining how models are related. One approach with crop models has been to identify models that have similar equations for underlying processes, such as photosynthesis. In general, however, it has not been found that structural similarity leads to similarity in simulated values in crop MMEs (Palosuo et al. 2011; Martre et al. 2015; Li et al. 2015). An alternative approach, proposed for climate models (Bishop and Abramowitz 2013), is to examine the covariance in model errors as the measure of model dependence. High correlation of the model errors indicates that the simulations are not independent. This approach should be explored for crop models, in order to see what insights it brings into the structure of a MME.

2.3 Determining the required number of models in a MME

The number of models in a crop MME is important, because it will affect both the mean and the variability of the ensemble outputs. It has practical implications because it is difficult to organize studies with multiple models. If it is sufficient to have fewer models, then such studies will be that much easier to conduct.

In crop modeling, emphasis has been on how the mean or median of an output variable of interest varies as the number of models is reduced, and in particular on the number of models required for the mean or median to stabilize (Asseng et al. 2013). However, this does not directly address the question of how many ensemble members are necessary to provide a satisfactory estimate of model uncertainty. This should be a topic of further study.

2.4 Proposing a statistical sampling model for model ensembles

In order to better understand ensemble properties, and in particular in order to examine theoretically the effect of number of models in a MME, hypotheses about the population of models being sampled and the sampling process are necessary. One simple hypothesis proposed for climate models is the "truth plus error" paradigm, which assumes that the population of model predictions is correct on the average, but each model has some error drawn from a distribution of errors. A different hypothesis is the "indistinguishable" paradigm, which assumes that both model predictions and truth are drawn from the same distribution (Bishop and Abramowitz 2013). This topic has not been addressed in the crop modeling community. It is important to do so in order to provide a theoretical underpinning for discussion of MME properties. The hypotheses proposed for climate models provide a useful starting point.

2.5 Creating model repositories

Given the effort involved in creating crop MME simulation studies, and their potential usefulness, it is of interest to create data repositories to make the results fully available to the community of researchers and end-users, as is done for climate models (Williams et al. 2011). As the climate modeling experience has shown, this would also spur research into the methodology and practice of using ensembles. The amount of data to be stored is significantly less for crop models than for climate models. On the other hand, a certain fragmentation of the work in crop modeling (for example for different crops) means that it may be harder to create one overall data repository. In any case, the example of the climate modeling community will be very valuable here.

2.6 Assigning different weights to each model in a multi-model ensemble

Model weighting involves giving possibly different weights to results from different models in a MME. The objective is to improve estimates of uncertainty and/or to improve projections or predictions based on the ensemble mean or median.

In climate modeling, there have been a number of approaches testing alternative means of combining models, and this remains an active area of research. Knutti et al. (2010) identified important elements in studies concerning ranking or weighting of models. Among these, is the importance of explicitly explaining both the metric used for ranking and any statistical hypotheses about the models in the ensemble as a sample from a population of models. This clearly applies to crop MMEs as well.

Bayesian model averaging is particularly attractive, as a standard statistical method of taking into account model uncertainty (Wintle et al. 2003; Clyde and George 2004). In this approach one starts from a priori weights for each model (often equal weights), and then updates the weights based on model agreement with observations. This could improve both uncertainty estimation and prediction. As an example, Robertson et al. (2004) found that their Bayesian optimal weighting scheme for seasonal climate prediction outperformed the ensembles with equal weighting.

Giorgi and Mearns (2002) defined a reliability ensemble average (REA) where the weight assigned to each model depends both on that model's agreement with observed data and its agreement with the ensemble average for projections. This approach has been used in multiple examinations of both global and regional climate model results (Sobolowski and Pavelsky 2012), and as a point of departure for formulating probability density functions (Tebaldi and Knutti 2007). More recently the method has been revised to dispense with the model convergence criterion (Xu et al. 2010), which has been controversial in some quarters.

Several different means of weighting different regional climate models in an ensemble have been suggested by Liu et al. (2010). They compare three weighting methods. The first is equal weights (the simple ensemble). The second weights by the inverse of fractional contribution to squared absolute error. The third method, found to be the best, calculates weights to minimize squared error of the final weighted ensemble.

A relatively recent approach in climate modeling is to weight models based both on performance and on between-model correlations of residual errors (e.g., Sanderson et al. 2015). Weighting based on correlations is supposed to remove much of the dependence between models, and thus make the sample more like a sample of independent models. Studies have shown that such a weighting scheme is superior to simple model averaging, with respect to both evaluating uncertainty and improving predictions (Bishop and Abramowitz 2013; Evans et al. 2013).

Model weighting however is not standard procedure in climate modeling. In the most recent Intergovernmental Panel on Climate Change (IPCC) report, it is acknowledged that the climate community does not know how to weight models to determine the best estimate of future climate change (Flato et al. 2013), though there is an example in that report where models are selected based on a verification protocol for projection of Arctic Sea Ice decrease throughout the twenty-first century.

As far as we know there has been no published work to date on differential weighting of crop models in a MME, although the question has been raised (Martre et al. 2015). Since crop model ensembles are often evaluated using quite limited amounts of data, it might seem that performance weighting of crop models is unlikely to be useful. On the other hand, a major argument against weighting of climate models based on hindcast error is that it assumes that the response to future radiative forcing will be consistent with that in the historical period, which cannot be tested. Crop models on the other hand can at least be tested against field experiments that impose higher atmospheric CO_2 concentrations (Kimball et al. 1995; Ewert et al. 2002) and higher temperatures (Wall et al. 2011) than are observed today.

It would be of interest to explore weighting schemes for crop models, starting with the various approaches tested with climate models. A major decision in performance weighting is the choice of outputs to be considered. The choice is very large for climate models, but more constrained for crop models. Even here, however, it will be necessary to study whether it is preferable to base performance weighting just on the outputs of major interest (often just yield), or to use a larger group of output variables. For example, phenology is essentially always simulated, so part of weighting could be based on agreement with phenology data.

2.7 Creating ensembles based on a single model with multiple parameter vectors

We can create a distribution of outcomes from a single model, by sampling from the probability distribution of uncertain quantities that affect the model outputs. This results in an ensemble of different model configurations, which are effectively different models though all use the same basic equations and structure. For climate models, the two sources of uncertainty within a single model that have been explored are uncertainties in parameter values and uncertainties in the initial conditions.

Ensembles which use the same model but multiple parameter values are referred to as “perturbed physics ensembles” or ‘parameter permutation experiments’ (PPEs). An example is described in Murphy et al. (2004) who estimated climate model uncertainty based on a 53-member ensemble of model versions using different parameter values. The uncertainty ranges of the parameters were determined by expert opinion, and the acceptability of the parameters was based on objective goodness-of-fit criteria for the model with those parameters. Other PPEs are described in (Yokohata et al. 2011) and (Sanderson 2011). It can, however, be difficult to quantify the uncertainty in the parameter values, in particular when these are ad hoc values without any clear theoretical or observational basis. Furthermore, while it is conceptually easy (one uses Monte Carlo sampling) to generate a distribution of outputs by sampling from the distributions of many or all of the parameters, it can be computationally very demanding. The use of model emulators can facilitate the evaluation of a broad number of parameter combinations (Murphy et al. 2007).

Although they are not generally called ensemble studies, there have been many studies where the effect of parameter uncertainty on the predictions of a particular crop model has been studied, including specifically in the study of the consequences of CO_2 enrichment (Challinor and Wheeler 2008). Much of this work has been oriented toward sensitivity analysis (ranking parameters according to their effect on model outputs) rather than toward uncertainty estimates. As for climate models, a major difficulty is quantifying the uncertainty in the parameters. Most commonly, parameter uncertainty is based on the range of values found in the literature (Aggarwal 1995; Richter et al. 2010), and only a fairly small fraction of the total number of parameters is treated as uncertain. In these studies the evaluation of parameter vectors, to eliminate vectors that give unrealistic results, is not usually practiced. This could be considered in future studies.

There have also been a few parameter uncertainty estimates based on a Bayesian approach to model calibration (Iizumi et al. 2009; Wallach et al. 2012)). A simpler alternative that has been applied to crop models is the GLUE algorithm, which explores the space of possible parameter values, calculates a likelihood and eliminates parameter vectors whose likelihood is below a threshold (Wang et al. 2005). Controversy concerning GLUE due to its subjective aspects is summarized in (Beven and Binley 2014).

There is a need to pursue studies that quantify the contribution of parameter uncertainty to crop model prediction uncertainty. This will require improved approaches to quantifying parameter uncertainty. It will be useful to distinguish between those parameters estimated by calibration using data common to all the models in a MME, and the other model parameters based on data specific to each model.

2.8 Creating ensembles based on a single model with multiple input values

Another type of climate model ensemble based on a single model results from varying the initial conditions that are used to start the simulations (e.g. Deser et al. 2012; Deser et al. 2014), thus exploring the internal variability that results. For example, Deser et al. (2014) examined the ensemble of results of the NCAR CCSM3 climate model with each member beginning from a slightly different initial atmospheric state.

For crop models, typical input variables are initial conditions, daily weather, soil properties and crop management. Here initial conditions have no special importance; many of the input variables are difficult to estimate and may have quite large uncertainties. Though not usually referred to as ensemble studies, there have been multiple studies of the effect of input uncertainty on crop model simulations (Bouman 1994; Aggarwal 1995; Moeller et al. 2009; Roux et al. 2014). Two specific types of input uncertainty have received particular attention recently. First, one method of upscaling crop model outputs from field to region or beyond is to execute a model in every grid cell, using a representative field for each cell (Rosenzweig et al. 2014). The uncertainty in the choice of representative field is in fact uncertainty in the input variables. Uncertainty due to scale change is explicitly studied in (Zhao et al. 2015). Secondly, in impact assessment studies, the uncertainty in the climate projections must be taken into account. This has led to running crop models with multiple future climates (some recent examples are Asseng et al. 2013; Li et al. 2015).

One analogy between climate and crop model inputs may be instructive. It has been argued that when studying the impact of climate change on crops and soil, it is important to do long-term simulations without reinitializing soil conditions each year (Basso et al. 2015). It would be worthwhile to investigate whether the uncertainty due to uncertain initial conditions increases with time, as in climate models.

More information concerning the importance of uncertainty in explanatory variables for crop models, in both absolute terms and relative to other sources of uncertainty, would be valuable.

2.9 Super ensembles

The core concept of super ensembles is combining different types of ensembles, for example, MMEs and multiple initial conditions or multiple global climate models coupled with multiple regional climate models (Kendon et al. 2010). In the context of seasonal climate forecasts a common combination is multi-model experiments along with single model initial conditions realizations (Robertson et al. 2004). There have not been, to this point, efforts to combine climate models that represent the three different types of ensembles.

In crop modeling, there have been few studies that combine multiple sources of uncertainty. These include multiple crop models with multiple climate models (Mearns et al. 1999; Asseng et al. 2013; Li et al. 2015), multiple crop and climate models with multiple parameters (Tao et al. 2009) and multiple parameter values with multiple values for inputs (Aggarwal 1995).

Given that computational considerations are less limiting for crop models than for climate models, it would probably be feasible to do full factorial simulation experiments for crop models (multiple models, multiple inputs, multiple parameterizations for each model). It would be important to consider all of these sources of uncertainty in a common framework, in order to obtain better estimates of overall uncertainty and the relative importance of the different contributions to uncertainty.

3 Analyzing the results of ensembles

3.1 Quantifying and displaying uncertainty

The major motivation for working with ensembles is the information provided about uncertainty. Many of the difficulties of assessing and reporting uncertainty are described in (Wesselink et al. 2015).

Examining the variability within a multi-model ensemble (MME) is the central approach in IPCC reports to quantifying uncertainty in climate projections due to structural uncertainties. Results concern different output variables at various spatial scales. An atlas in the most recent report displays results for different scenarios of radiative forcing (IPCC 2013b), showing the 25th, 50th, and 75th percentiles of the model results.

Primarily Bayesian probabilistic methods have been used in quantifying uncertainty in PPEs whether they are used for quantifying uncertainty in climate sensitivity (Murphy et al. 2004) or in regional climate changes (Murphy et al. 2007). In examining a 57 member ensemble of PPEs, (Murphy et al. 2014) used fans of uncertainty (10th to 90th percentiles) to communicate uncertainty. In contrast, in analysis of the results of initial condition experiments, more analytic displays of uncertainty have been used. For example, (Deser et al. 2014) display individual results for the suite of initial condition simulations performed with the NCAR CCSM3.

As for climate models, crop MME studies have focused on inter-model variability to quantify uncertainty. In considering the effect of changed conditions, the uncertainty information is presented as an inter-model coefficient of variation (standard deviation of simulated values/mean simulated value), the inter-quartile range (Pirttioja et al. 2015) or as the percentage of models agreeing in the sign of change (Rosenzweig et al. 2014).

As discussed above, variability between models is not the only source of projection uncertainty. There is also parameter and input uncertainty to consider, as well as common biases that may cause the truth to fall outside of the ensemble spread. An important future activity in crop modeling will be to evaluate overall uncertainty and error, both for projections under uncertain climate and for predictions under testable conditions. Also, attention to the communication of uncertainty information is important.

An additional important topic that we do not explore in detail here is the role of data in determining estimations of uncertainty, whether it be the data used for model development, for calibration or the data used for comparison with simulated values.

3.2 Evaluating the separate contributions to overall uncertainty

Another important step is to disaggregate total uncertainty among the various contributions, commonly referred to as sensitivity analysis (Saltelli et al. 2000). One approach that has been applied to climate models is that of analysis of variance, as suggested by Yip et al. (2011) and used in the context of regional and global climate models by Mearns et al. (2013). (Wallach et al. 2016) propose a random effects analysis of variance for crop model uncertainty, with analytical expressions for the contributions from model structure, parameters and inputs. (Asseng et al. 2013) use a simple comparison of variances to separate the contributions of multiple crop and multiple climate models to overall uncertainty. Further studies that implement these or other approaches for crop model simulation experiments are needed.

3.3 Evaluating uncertainty estimates

An important question is the extent to which uncertainties estimated on the basis of one data set represent true uncertainty of new predictions or projections. If the uncertainty is represented by a probability distribution, then the true response for the new data should be within the estimated x% confidence intervals, x% of the time. This cannot be checked for climate projections, but could be checked for crop models under conditions such as increased CO_2 and temperature, that partially imitate future climate.

3.4 Using the ensemble average as estimator or predictor

Tebaldi and Knutti (2007) cite literature that shows that though the mean ensemble prediction may not be significantly better than the best model for each particular variable, the ensemble prediction usually does do better when performance over several variables is considered. However, the ensemble mean may not be best if common biases are present across all modeling groups or if a particular model features a unique aspect of crucial importance (Bukovsky et al. 2015).

Many MME studies with crop models have remarked that the mean or median of the simulated values seems to give good agreement with observed yields (Asseng et al. 2014; Bassu et al. 2014; Li et al. 2015; Palosuo et al. 2011). Martre et al. (2015) found that when multiple outputs are considered, the mean and median were both better predictors than even the best individual model, with the median being slightly better than the mean.

The conclusion that the mean or median is really a better estimator than the best model would imply that even without improving present-day models, one could obtain better predictions by using ensembles. Building off an expectation that ensemble medians are stronger predictors of future conditions than any individual model, Asseng et al. (2014) base extrapolations of future temperature impacts on world wheat production on the median of a MME. Given the potential importance of the mean or median in providing improved predictions, and the relatively limited supporting evidence to date, it seems important to test this hypothesis more thoroughly.

There has been only limited discussion of why the crop MME mean or median is a good predictor. In addition to the same arguments as for climate models one could add that different crop models are developed and tested using different data (except usually for a relatively small shared data set used for calibration) so in this sense crop model ensembles are based on more data than individual models (Martre et al. 2015). It is critical to better understand why the ensemble mean or median is a good predictor, as the basis for identifying the situations where this result will hold and where it will not.

Further exploration is also needed into how the choice of models, the number of models, and possible weighting of models affect the quality of the mean or median as a

Table 1 Proposed actions to improve creation and use of crop MMEs and associated benefits

For details, see section	Proposed action (see section indicated for details)	Benefits
2.1	Develop standardized tests for candidate models of crop MMEs. Propose and test guidelines for including (or excluding) models in crop MME studies.	Clearer rules for creating crop MMEs, and better appreciation of consequences of those rules.
2.2	Test the use of covariances of errors as a way of quantifying the degree of relatedness of models in a MME	Improved insights into the structure of a MME. Basis for down-weighting related models.
2.3	Investigate the effect of number of models in a crop model MME on estimates of uncertainty and on the quality of the mean or median as a predictor	Information on minimum required number of models in a MME.
2.4	Develop and evaluate statistical assumptions about the sampling process underlying MMEs	Provide theoretical basis for studies of the properties of MMEs.
2.5	Create repositories for storing multi model results	Make crop MME results available for further studies.
2.6	Develop and test model weighting approaches for crop MMEs	Model weighting could improve estimate of uncertainty. Weighted mean could be better predictor than simple mean.
2.7	Carry out further studies of the uncertainty in crop model simulations due to parameter uncertainty. Develop improved methods of estimating parameter uncertainty	Evaluate importance of parameter uncertainty per se and in relation to other sources of uncertainty.
2.8	Carry out further studies of the uncertainty in crop simulations due to input uncertainty, including how uncertainty increases in multi-year simulations	Evaluate importance of input uncertainty per se and in relation to other sources of uncertainty.
2.9	Carry out simulation studies involving all of structure, parameter and input uncertainty	Create a common framework for estimating uncertainty from all of these sources of uncertainty.
3.1	Produce estimates of total uncertainty, including common biases of models	Produce information for evaluating overall confidence in simulated values.
3.2	Estimate the separate contributions to overall uncertainty	Prioritize future work on evaluating and possibly reducing uncertainty from some of the different sources.
3.3	Test whether ensemble based uncertainty estimates are realistic	Determine level of confidence in uncertainty estimates.
3.4	Carry out further studies on the mean or median of MMEs versus the best model. Develop a theoretical framework for evaluating the MME mean or median	Identify when the mean or median is likely to be a useful predictor.

predictor. Once again, it is important to consider the output variables that are taken into account, since improvement may not concern all outputs equally.

4 Conclusions

Working with ensembles of models is a recent, important development in crop modeling. The major advantage of doing so is that MMEs provide information on uncertainty in crop model estimations and projections related to model structure, which has been shown to be a major source of uncertainty. It has also been shown that the mean or median of multiple crop models can be a better estimator than even the best model. A third advantage of working with multi-model ensembles is the collaboration and exchanges that it fosters within the crop modeling community.

These important advantages, and the experience from the climate modeling community, strongly suggest that studies based on crop model ensembles may expand in the future to become a common way of working with crop models. Advancing the methodology of working with crop MMEs is necessary to realize the full potential of such studies.

This work will progress much faster to the extent that it builds on the large amount of research in this area already accomplished by the climate modeling community. Of course, given the differences between climate and crop modeling, not all of the lessons from climate modeling are applicable to crop models.

Table 1 summarizes the proposals that have been made here for ways forward for crop modeling based on ensembles. They are largely inspired by the work in the climate modeling community. The propositions include defining criteria for acceptance of models in a crop MME, exploring criteria for evaluating the degree of relatedness of models in a MME, studying the effect of number of models in the ensemble, development of a statistical model of model sampling, creation of a repository for MME results, studies of possible differential weighting of models in an ensemble, creation of single model ensembles based on sampling from the uncertainty distribution of parameter values or inputs specifically oriented toward uncertainty estimation, the creation of super ensembles that sample more than one source of uncertainty, the analysis of super ensemble results to obtain information on total uncertainty and the separate contributions of different sources of uncertainty and finally further investigation of the use of the multi-model mean or median as a predictor.

Acknowledgments The authors gratefully acknowledge the role of the AgMIP project, which has encouraged the close collaboration between climate and crop modelers of which this study is one example. Funding from the MACSUR project for two of the authors is also gratefully acknowledged. Dr. Mearns' participation was funded by NCAR, which is funded by the U.S. National Science Foundation. Dr. Ruane's work was supported by funding from NASA Earth Sciences (WBS: 281945.02.03.03.96).

References

Aggarwal PK (1995) Uncertainties in crop, soil and weather inputs used in growth models: implications for simulated outputs and their applications. Agric Syst 48:361–384. doi:10.1016/0308-521X(94)00018-M

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

Asseng S, Ewert F, Martre P, et al. (2014) Rising temperatures reduce global wheat production. Nat Clim Chang 5:143–147. doi:10.1038/nclimate2470

Basso B, Hyndman DW, Kendall AD, et al. (2015) Can impacts of climate change and agricultural adaptation strategies Be accurately quantified if crop models are annually Re-initialized? PLoS One 10:e0127333. doi:10.1371/journal.pone.0127333

Bassu S, Brisson N, Durand J-L, et al. (2014) How do various maize crop models vary in their responses to climate change factors? Glob Chang Biol 20:2301–2320. doi:10.1111/gcb.12520

Beven K, Binley A (2014) GLUE: 20 years on. Hydrol Process 28:5897–5918. doi:10.1002/hyp.10082

Bishop C, Abramowitz G (2013) Climate model dependence and the replicate earth paradigm. Clim Dyn 41:885–900. doi:10.1007/s00382-012-1610-y

Bouman BAM (1994) A framework to deal with uncertainty in soil and management parameters in crop yield simulation: a case study for rice. Agric Syst 46:1–17

Bukovsky MS, Carrillo CM, Gochis DJ, et al. (2015) Toward assessing NARCCAP regional climate model credibility for the north American monsoon: future climate simulations. J Clim 28:6707–6728. doi:10.1175/JCLI-D-14-00695.1

Challinor AJ, Wheeler TR (2008) Use of a crop model ensemble to quantify CO2 stimulation of water-stressed and well-watered crops. Agric For Meteorol 148:1062–1077

Challinor AJ, Smith MS, Thornton P (2013) Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. Agric For Meteorol 170:2–7. doi:10.1016/j.agrformet.2012.09.007

Clyde M, George EI (2004) Model uncertainty. Stat Sci 19:81–94

Deser C, Phillips A, Bourdette V, Teng H (2012) Uncertainty in climate change projections: the role of internal variability. Clim Dyn 38:527–546. doi:10.1007/s00382-010-0977-x

Deser C, Phillips AS, Alexander MA, Smoliak BV (2014) Projecting north American climate over the next 50 Years: uncertainty due to internal variability. J Clim 27:2271–2296. doi:10.1175/JCLI-D-13-00451.1

Evans JP, Ji F, Abramowitz G, Ekström M (2013) Optimally choosing small ensemble members to produce robust climate simulations. Environ Res Lett 8:44050

Ewert F, Rodriguez D, Jamieson P, et al. (2002) Effects of elevated CO2 and drought on wheat: testing crop simulation models for different experimental and climatic conditions. Agric Ecosyst Environ 93:249–266. doi:10.1016/S0167-8809(01)00352-8

Flato G, Marotzke J, Abiodun B, et al. (2013) Evaluation of Climate Models. In: Stocker T, Qin D, Plattner G-K, et al. (eds) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambride University Press, Cambridge United Kingdom

Gates WL, Boyle JS, Covey C, et al. (1999) An overview of the results of the atmospheric model intercomparison project (AMIP I. Bull Am Meteorol Soc 80:29–55. doi:10.1175/1520-0477(1999)080 < 0029:AOOTRO > 2.0.CO;2

Giorgi F, Mearns LO (2002) Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the `Reliability Ensemble Averaging' (REA) Method. J Clim 15:1141–1158. doi:10.1175/1520-0442(2002)015&lt;1141:COAURA&gt;2.0.CO;2

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept Tellus A 219–233

Harris GR, Collins M, Sexton DMH, et al. (2010) Probabilistic projections for twenty-first century European climate. Nat Hazards Earth Syst Sci 10:2009–2020. doi:10.5194/nhess-10-2009-2010

Iizumi T, Yokozawa M, Nishimori M (2009) Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: application of a Bayesian approach. Agric For Meteorol 149:333–348

IPCC (2013a) Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change

IPCC (2013b) Annex I: Atlas of global and regional climate projections. In: Stocker TF, Qin D, Plattner G-K, et al. (eds) Climate Change 2013: The Physical Sci - ence Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [van Oldenborgh, G.J., M. Collins, J. Arblaster, J.H. Christensen, J. Marotzke, S.B. Power,. Cambridge University Press, Cambridge, United Kingdom and New York, USA

Jamieson PD, Porter JR, Goudriaan J, et al. (1998) A comparison of the models AFRCWHEAT2, CERES-wheat, Sirius, SUCROS2 and SWHEAT with measurements from wheat grown under drought. F. Crop Res 55:23–44. doi:10.1016/S0378-4290(97)00060-9

Kendon EJ, Jones RG, Kjellström E, et al. (2010) Using and designing GCM–RCM Ensemble regional climate projections. J Clim 23:6485–6503. doi:10.1175/2010JCLI3502.1

Kimball B, Pinter P, Garcia R, et al. (1995) Productivity and water use of wheat under free-air CO2 enrichment. Glob Chang Biol 1:429–442. doi:10.1111/j.1365-2486.1995.tb00041.x

Knutti R (2010) The end of model democracy? Clim Chang 102:395–404. doi:10.1007/s10584-010-9800-2

Knutti R, Abramowitz G, Collins M, et al (2010) IPCC expert meeting on assessing and combining multi-model climate projections. Good practice guidance paper on assessing and combining multi model climate projections

Li T, Hasegawa T, Yin X, et al. (2015) Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob Chang Biol 21:1328–1341. doi:10.1111/gcb.12758

Liu C-M, Wu M-C, Paul S, et al. (2010) Super-ensemble of three RCMs for climate projection over East Asia and Taiwan. Theor Appl Climatol 103:265–278. doi:10.1007/s00704-010-0275-x

Martre P, Wallach D, Asseng S, et al. (2015) Multimodel ensembles of wheat growth: many models are better than one. Glob Chang Biol 21:911–925. doi:10.1111/gcb.12768

Mearns LO, Mavromatis T, Tsvetsinskaya E, et al. (1999) Comparative responses of EPIC and CERES crop models to high and low spatial resolution climate change scenarios. J Geophys Res Atmos 104:6623–6646. doi:10.1029/1998JD200061

Mearns LO, Sain S, Leung LR, et al. (2013) Climate change projections of the north American regional climate change assessment program (NARCCAP. Clim Chang 120:965–975. doi:10.1007/s10584-013-0831-3

Meehl GA, Moss R, Taylor KE, et al. (2014) Climate model Intercomparisons: preparing for the next phase. Eos. Trans Am Geophys Union 95:77–78. doi:10.1002/2014EO090001

Moeller C, Asseng S, Berger J, Milroy SP (2009) Plant available soil water at sowing in Mediterranean environments—is it a useful criterion to aid nitrogen fertiliser and sowing decisions? F. Crop Res 114: 127–136. doi:10.1016/j.fcr.2009.07.012

Murphy JM, Sexton DMH, Barnett DN, et al. (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430:768–772. doi:10.1038/nature02771

Murphy JM, Booth BBB, Collins M, et al. (2007) A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. Philos trans R Soc London a math. Phys Eng Sci 365:1993–2028

Murphy JM, Booth BBB, Boulton CA, et al. (2014) Transient climate changes in a perturbed parameter ensemble of emissions-driven earth system model simulations. Clim Dyn 43:2855–2885. doi:10.1007/s00382-014-2097-5

Palosuo T, Kersebaum KC, Angulo C, et al. (2011) Simulation of winter wheat yield and its variability in different climates of Europe: a comparison of eight crop growth models. Eur J Agron 35:103–114. doi:10.1016/j.eja.2011.05.001

Pirttioja N, TR C, Fronzek S, et al. (2015) Temperature and precipitation effects on wheat yield across a European transect: a crop model ensemble analysis using impact response surfaces. Clim Res 65:87–105

Richter GM, Acutis M, Trevisiol P, et al. (2010) Sensitivity analysis for a complex crop model applied to durum wheat in the Mediterranean. Eur J Agron 32:127–136. doi:10.1016/j.eja.2009.09.002

Robertson AW, Lall U, Zebiak SE, Goddard L (2004) Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. Mon Weather Rev 132:2732–2744. doi:10.1175/MWR2818.1

Rosenzweig C, Jones JW, Hatfield JL, et al. (2013) The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies. Agric For Meteorol 170:166–182. doi:10.1016/j.agrformet.2012.09.011

Rosenzweig C, Elliott J, Deryng D, et al. (2014) Assessing agricultural risks of climate change in the twenty-first century in a global gridded crop model intercomparison. Proc Natl Acad Sci U S A 111:3268–3273. doi:10.1073/pnas.1222463110

Rötter RP, Palosuo T, Kersebaum KC, et al. (2012) Simulation of spring barley yield in different climatic zones of northern and Central Europe: a comparison of nine crop models. F. Crop Res 133:23–36. doi:10.1016/j.fcr.2012.03.016

Roux S, Brun F, Wallach D (2014) Combining input uncertainty and residual error in crop model predictions: a case study on vineyards. Eur J Agron 52:191–197. doi:10.1016/j.eja.2013.09.008

Saltelli A, Chan K, Scott EM (2000) Sensitivity analysis. Wiley, New York

Sanderson BM (2011) A Multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. J Clim 24:1362–1377

Sanderson BM, Knutti R, Caldwell P (2015) Addressing interdependency in a multimodel ensemble by interpolation of model properties. J Clim 28:5150–5170. doi:10.1175/JCLI-D-14-00361.1

Sobolowski S, Pavelsky T (2012) Evaluation of present and future North American Regional Climate Change Assessment Program (NARCCAP) regional climate simulations over the southeast United States

Tao F, Zhang Z, Liu J, Yokozawa M (2009) Modelling the impacts of weather and climate variability on crop productivity over a large area: a new super-ensemble-based probabilistic projection. Agric For Meteorol 149: 1266–1278. doi:10.1016/j.agrformet.2009.02.015

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans A Math Phys Eng Sci 365:2053–2075. doi:10.1098/rsta.2007.2076

Tubiello FN, Ewert F (2002) Simulating the effects of elevated CO2 on crops: approaches and applications for climate change. Eur J Agron 18:57–74. doi:10.1016/S1161-0301(02)00097-7

Wall GW, Kimball BA, White JW, Ottman MJ (2011) Gas exchange and water relations of spring wheat under full-season infrared warming. Glob Chang Biol 17:2113–2133. doi:10.1111/j.1365-2486.2011.02399.x

Wallach D, Keussayan N, Brun F, et al. (2012) Assessing the uncertainty when using a model to compare irrigation strategies. Agron J 104:1274–1283. doi:10.2134/agronj2012.0038

Wallach D, Thorburn P, Asseng S, et al (2016) Estimating model prediction error: Should you treat predictions as fixed or random? Environ. Model. Softw.in press

Wang X, He X, Williams JR, et al. (2005) Sensitivity and uncertainty analyses of crop yields and soil organic carbon simulated with EPIC. Trans ASAE 48:1041–1054

Wesselink A, Challinor AJ, Watson J, et al. (2015) Equipped to deal with uncertainty in climate and impacts predictions: lessons from internal peer review. Clim Chang 132:1–14. doi:10.1007/s10584-014-1213-1

Williams DN, Lawrence BN, Lautenschlager M, et al. (2011) Earth system grid federation: delivering globally accessible Petascale data for CMIP5. In: proceedings of the Asia-Pacific advanced. Network 32:121–130

Wintle BA, McCarthy MA, Volinsky CT, Kavanagh RP (2003) The use of Bayesian model averaging to better represent uncertainty in ecological models. Conserv Biol 17:1579–1590. doi:10.1111/j.1523-1739.2003.00614.x

Xu Y, Gao X, Giorgi F (2010) Upgrades to the reliability ensemble averaging method for producing probabilistic climate-change projections. Clim Res 41:61–81. doi:10.3354/cr00835

Yip S, Stephenson DB, Hawkins E (2011) A simple, coherent framework for partitioning uncertainty in climate predictions. J Clim 24:4634–4643

Yokohata T, Annan JD, Collins M, et al. (2011) Reliability of multi-model and structurally different single-model ensembles. Clim Dyn 39:599–616. doi:10.1007/s00382-011-1203-1

Zhao G, Hoffmann H, van Bussel LGJ, et al. (2015) Effect of weather data aggregation on regional crop simulation for different crops, production conditions, and response variables. Clim Res 65:141–157