Probabilistic Projections of Agro-Climate Indices in North America

Adam Terando*,^, Klaus Keller‡$, and William E. Easterling*

*Department of Geography, The Pennsylvania State University, University Park, Pennsylvania

‡ Department of Geosciences, The Pennsylvania State University, University Park, Pennsylvania

§ Earth and Environmental Systems Institute, The Pennsylvania State University, University Park, Pennsylvania

^ Current affiliation: Department of Biology, North Carolina State University, Raleigh, North Carolina

Corresponding author address: Adam Terando, 127 David Clark Labs, Raleigh, NC 27695-7617.
Email: adam_terando@ncsu.edu
Abstract

We develop probabilistic projections for three agro-climate indices (frost days, thermal time, and a heat stress index) for North America. The selected indices are important for understanding the potential impacts of future anthropogenic climate change on agricultural production. We use Bayesian Model Averaging (BMA) and bootstrapping to quantify the structural uncertainty in an ensemble of downscaled General Circulation Models (GCMs). The prior information contained in the observations and model hindcasts is used to construct physically meaningful temporal comparisons for the period 1961-2010. The comparisons are used to derive model-specific posterior weights to construct probabilistic projections of agro-climate change in the 21st century. A cross validation test covering the most recent 25 years of the observation period indicates considerable overconfidence in the projections when using the calibrated BMA approach. In contrast the probabilistic projections using equally weighted climate models are not overconfident. The strong consensus among the probabilistic projections that shows warming effects for all three agro-climate indices is tempered by the short 50-year calibration period and the small ensemble size. The short calibration period provides a relatively poor observational constraint on estimates of model weights and predictive variance, while the small ensemble size limits the climate sample space. However, the consensus that emerges in spite of the large uncertainties suggests large potential changes in the conditions that farmers will experience over the remainder of the 21st century. Of particular concern is the projected increase in the heat stress index which could lead to large crop damages and associated yield declines.
1. Introduction

Agriculture is vulnerable to anthropogenic climate change (cf. Easterling et al. 2007; Schlenker and Roberts 2009). But the degree of vulnerability and sensitivity is uncertain (e.g. Easterling et al. 2007; Challinor et al. 2009). Because of management inputs into agroecosystems it is often difficult to quantify and disentangle the yield effects of climate change and technology. Nevertheless recent studies provide compelling evidence that there is a causal link between temperature extremes and agricultural production and that this link could result in substantial negative impacts on crop yields in a warming climate (Lobell and Burke 2008; Schlenker and Roberts 2009). These studies highlight the importance of examining the tails of the temperature distribution in climate change impact studies. And it will be particularly important in the context of adaptation efforts to examine climate projections of agriculturally-relevant indices to assess potential changes in crop exposure to stress-inducing temperature extremes.

Complicating such efforts is the difficulty that the current generation of general circulation models (GCMs) have in simulating agro-climate indices. A recent evaluation of GCM output by Terando et al. (2011; hereafter referred to as T11) found large differences between simulations of heat stress and the observed conditions from 1961-2000 in North America. For other agro-climate indices the results are mixed as to which GCMs are better able to reproduce the observed trends and variability. These results suggest that predictions of these impact-relevant variables will be deeply uncertain; and quantifying this uncertainty will be important if decision-makers are to direct more efforts towards developing crop varieties or management practices that are better adapted to a changing climate (e.g. Ortiz et al. 2008). Given that we cannot eliminate this uncertainty it follows that the most useful projections for decision makers will also be probabilistic, calibrated, and sharp (Dawid 1984; Gneiting et al. 2007). Probabilistic
projections are forecasts that take the form of a probability density function (PDF) that describes the uncertainty of the model prediction in contrast to a single-value point prediction which (in effect) assumes no uncertainty for a given model. Projections that are well-calibrated will share the same statistical properties as the underlying distribution produced by nature. For these forecasts to be most useful to decision-makers they should also be sharp, meaning the predictive distribution over all models is not overdispersed with respect to a naive climatological forecast (Gneiting et al. 2007). We note that imposing sharpness constraints on current climate projections may not always be possible because of the risk of producing overconfident forecasts. In other words, given our uncertainty about the climate system, a well calibrated but possibly overdispersed PDF may be all that is possible.

Here we examine probabilistic projections of three agro-climate indices, frost days, thermal time, and heat stress index for North America. We particularly focus on characterizing, quantifying, and potentially reducing projection uncertainties. The hope is that this improved projection can help farmers and other decision-makers to better prepare for a changing climate.

The rest of the paper is structured as follows. Section 2 provides additional background on the simulation of agro-climate indices, especially in regards to model performance relative to other climate variables. We then discuss questions of model uncertainty and review methods to characterize uncertainty. Section 3 discusses our methodological choices used to develop probabilistic projections given the information in the observations and hindcasts. Section 4 describes the data sources and methods. Section 5 presents the projections and the uncertainty estimates. We discuss the potential implications of our results in Section 6 along with caveats in Section 7. Finally we review our findings and suggest areas for future work in Section 8.
2. Background

a. The link between temperature and field crops

Temperature changes affect crops. Schlenker and Roberts (2009) provide a damage function linking temperature and corn, soybeans, and cotton yields in the US. They demonstrate that temperature extremes are non-linearly related to yields, especially for temperatures greater than ~30 °C and that more recent crop varieties (tested over the period 1978-2005) are only marginally less sensitive extreme heat compared to earlier varieties. Using this established relationship, they predict that yields could decrease between 30-82% by the end of the century under various warming scenarios. The magnitude of these results has not only sparked a vigorous discussion about the robustness of these results for other regions or time periods (Meerburg et al. 2009), but also highlights the importance of examining the tails of the temperature distribution in climate change impact studies. Others have also argued for a focus on more impact-relevant measures of anthropogenic climate change at the regional level in addition to mean surface temperature (e.g. Easterling 2000; Feng and Hu 2004; Tebaldi et al. 2006, Meehl et al. 2007a; Terando et al. 2011). These measures typically include temperature and precipitation extremes based on some frequency percentile, event duration (e.g. dry or wet period, cf. Frich et al. 2002), and measures of heat accumulation or heat thresholds such as frost days, thermal time, and heat stress. This latter category of heuristics is particularly important in agricultural impact studies since these measures are better correlated with biomass production than mean temperatures.

The results of Schlenker and Roberts (2009) indicating a lack of agricultural adaptation to temperature extremes in the U.S. are not surprising. In T11 observed spatio-temporal patterns of frost days and thermal time (growing degree days) show statistically significant warming trends across broad swaths of North America but the same is not true for a measure of heat stress. In fact, negative heat stress trends are much more common over the period 1951-2010, especially in
the eastern part of the continent where crops are typically more vulnerable to high temperatures because of the reliance on rain-fed as opposed to irrigated agriculture. Thus the persistence of crop vulnerability to extreme temperatures may be a consequence of market forces and government policy reacting to a lack of warming in the recent climatological record.

Furthermore, Lobell and Burke (2008) show that temperature change is likely to be the greater determinant of future impacts on agriculture compared to precipitation changes. The reason suggested by Lobell and Burke (2008) is that the magnitude of projected temperature increases is much greater than the interannual variability. This is not the case for precipitation where the projected changes and the projection uncertainty are within the range of variability of the system. And as such, projections of temperature-based agro-climate variables are a key component of adaptation efforts.

Here we construct projections of three temperature-based agro-climate indices: frost days, thermal time and heat stress index. All three indices capture some aspect of the relationship between temperature and biomass production in the most commonly grown cereal crops in North America. Frost days, defined as the annual number of days when the minimum temperature is below 0 °C, capture the effects of both early and late season cold stress on crops as well as the total growing season length. Thermal time, or growing degree days, is a cumulative measure of heat units within a pre-defined range of daily temperatures over the growing season. The accumulation of heat units is correlated with phenological stages of crop development. The calculation of thermal time varies depending on the growing season of the crop and the lower and upper temperature bounds that define the respective boundaries of a zero growth rate, linear growth with temperature, and declining growth rate commensurate with heat stress (Tollenarr et al. 1979). We use a definition most commonly applied to maize (*Zea mays* L.) as it is the most
widespread crop grown in our study area. Similarly, we define a heat stress index that captures maize exposure to temperatures that retard growth and yields (Tollenarr et al. 1979).

b. Simulating Agro-climate Indices

GCMs can reproduce many aspects of regional and global climate fields (e.g. spatio-temporal trends and variability) when the post-industrial anthropogenic forcing is included (Meehl et al. 2007b; Randall et al. 2007). There is confidence that these models will project with some accuracy future climatic changes at global and continental scales (Randall et al. 2007). This is based on their ability to simulate the observed climate and past climatic changes. However, there is less evidence that GCMs can successfully reproduce other more impact-relevant measures of climate change at local scales at which adaptation activities are likely to occur (Hayhoe et al. 2008). Precipitation is one obvious example where extremes and multi-day events are generally poorly simulated due to the difficulty in resolving small-scale processes and the potential break-down of parameterization schemes at high resolution (Kain 2004; Liang et al. 2008). However even simulations of many temperature-based indices are characterized by significant model biases and errors. Recent studies examining agricultural climate indices and temperature extremes indicate that GCMs less accurately reproduce the observed inter-annual variability of these variables compared to the mean temperature (although they generally correctly simulate the sign of the observed trend; cf. Tebaldi et al. 2006). T11 use 17 GCMs from the Coupled Model Inter-comparison Project (CMIP III; Meehl et al. 2007c) to assess model performance for three agro-climate indices (frost days, thermal time, and heat stress days). T11 shows that GCM skill scores (Taylor 2001) for historic trends in North America are highest for frost days and lowest for heat stress days. This differing model performance is correlated to how
much of the temperature distribution is captured by the agro-climate index. For example the heat
stress index used in T11 is defined for a small portion of the temperature distribution in North
America since only temperatures above 30 °C are included. Conversely, the frost day threshold
of 0 °C is much closer to the central tendency of the (minimum) temperature distribution and
model skill is higher. T11 also found differences in model skill that are specific to the diurnal
temperature cycle which they suggest are related to the difficulty of simulating the effects of
atmospheric aerosols and clouds on maximum daily temperature.

c. Structural Uncertainty Between GCMs

In addition to the issue of model fidelity to the observational record there are broader
concerns about structural and parametric uncertainty between and within GCMs. Structural
uncertainty arises from the different representations and parameterizations of the physical
processes governing the dynamics and forcings of the climate system, reflecting our incomplete
knowledge about these forcings, or our inability to explicitly resolve them. Parametric
uncertainty reflects the incomplete knowledge about the true value of the parameter values.
While an exhaustive cataloguing and quantification of these sources of uncertainty is at this time
an outstanding challenge (Gneiting et al. 2007; Knutti et al. 2010), they should be acknowledged
and recognized in the decision-making process (Keller et al. 2007).

From a decision-making perspective, we can describe a hierarchy of methods for
representing climate model output in impact assessments based on the treatment of projection
uncertainty and the attendant risk of overconfident predictions. By overconfident we refer to an
ensemble prediction where the probability of forecasting a given value or event is too small
compared to the true predictive probability density function (Wilks 2001; Raftery et al. 2005). At
one level (and representing the greatest risk of overconfident predictions) is the use of point projections from a single model. Such an approach was most common before multi-model ensembles became common-place (e.g. the first US National Assessment; National Assessment Synthesis Team 2001) although this single-model approach is still sometimes seen, especially in studies where the actual GCM results are not necessarily the focus of the analysis (e.g. Schlenker and Roberts 2009).

The next level can be described as an ensemble approach where output from multiple GCMs are reported (e.g. Schmittner et al. 2005, Meehl et al. 2007b, Hay and McCabe 2010, Knutti et al. 2010, Cook et al. 2011). This method has grown in popularity over the last decade as the number of available models has increased. An often implicit assumption in this approach is that the true realized climate state will be within this sample of models. There is in fact, evidence that using an ensemble of models reduces the probability of comparing spurious trends due to internal climate variability (both simulated and observed) and lowers mean-model error (Weigel et al. 2008; Pierce et al. 2009). However there is still a potentially non-trivial probability that the true climate state will lie outside of the sample space of the model ensemble (e.g. Draper 1995). Such a result is particularly undesirable in cases where the resulting climate change is associated with high impact events, such as more frequent and intense heat waves (Rahmstorf and Coumou 2011) or climate threshold responses such as a persistent weakening of the Meridional Overturning Current (Alley et al. 2005; Urban and Keller 2010).

This last category of studies uses probabilistic projections to represent model uncertainty. These involve more explicit treatments of the structural uncertainty in numerical climate and weather models, mainly through Bayesian techniques that derive a posterior probabilistic prediction based on some measure of model fidelity to observations (Gneiting et al. 2007).
Raftery et al. (2005; referred to as R05) used Bayesian Model Averaging (BMA) to derive posterior model-specific weights and weighted PDFs for operational weather forecasting models. The weights correspond to the probability of a model being correct given a set of prior observations that are compared to corresponding model hindcasts. The sharpness of the model forecast PDF is also based on model fidelity to observations and reflects both between-forecast variability and within-forecast variability (Raftery et al. 2005). The EM algorithm (Expectation Maximization; Dempster et al. 1977) is used to derive the set of weights that maximizes the likelihood of observing the training data. The overall BMA PDF that is conditioned on the set of model forecasts is a weighted PDF based on past model performance over the training period. A more recent study by Bhat et al. (2011) extends the BMA to include GCMs, taking into account spatial and temporal dependency and using hindcast output from the ensemble to estimate the model weights.

Tebaldi et al. (2005) and Smith et al. (2009) (referred to as TS09) used a different approach to quantify structural uncertainty with a hierarchical Bayes model that treats the parameters of interest (projected temperature change and climatic variability) as random variables whose joint density is estimated through Markov Chain Monte Carlo (MCMC) simulation. The weights are derived based on model bias (compared to the observations) and model convergence towards some measure of central tendency estimated from model forecasts (calculated using the “reliability ensemble averaging” as described in Giorgi and Mearns 2002). Tebaldi and Lobell (2008) apply this method by developing probabilistic yield projections based on an ensemble of GCMs and uncertainties in crop responses to temperature, precipitation, and CO₂ fertilization. All these methods that follow either the R05 or TS09 approaches strive for
calibrated and sharp forecasts (as defined by Gneiting et al. 2007), and depending on how the model expansion is implemented can control to some degree for underdispersive predictions (but rely, of course, on strong statistical assumptions).

However, by focusing on projections that are calibrated and sharp, there is still a risk of overconfident predictions if the estimated model uncertainty is too small compared to the actual model uncertainty (Draper 1995). In particular the convergence term used by TS09 will penalize ‘outlier’ models that do not project climatic changes similar to the majority of the ensemble. This may or may not be a valid assumption in practice. Smith et al. (2009) point out that accuracy could suffer if the GCMs from the different modeling groups have the same sources of error arising from our incomplete knowledge of the climate system, or if similar (and similarly flawed) methods are used to parameterize processes that cannot be resolved at each model time-step. This speaks to broader issues of model weighting that result from the calibration process and is the subject of vigorous debate in the climate literature (Knutti 2010; Mote et al. 2011). Given our imperfect knowledge about the true sensitivity of the climate system to anthropogenic climate forcings, it has been argued that projections should be calibrated to reflect equal model weights rather than deriving an optimal set of weights based on model fidelity to observations (Brekke et al. 2008). The risk of this approach is that the projections will be underdispersed and diffuse and therefore not as useful for decision-makers compared to a sharper PDF that is able to fully incorporate the information provided by the hindcasts and the observations into the predictive distribution.

Clearly there are many unresolved issues when using multi-model ensembles to provide decision-relevant information. However, we can at a minimum establish that a necessary (but not sufficient) condition for a multi-model ensemble is that it is at least not overconfident in out-of-
sample cross validation tests. This serves as a useful first test of ensemble skill that can also be easily operationalized.

Our approach is to use Bayesian Model Averaging to derive probabilistic projections of frost days, thermal time, and heat stress index for North America. We improve on the standard method of representing projection uncertainty with an ensemble of point projections (e.g. Hegerl et al. 2007) by deriving a predictive distribution for the ensemble conditional on the observations and hindcasts. By using BMA, we seek to avoid overconfident predictions by quantifying the structural uncertainty between models while also making no assumptions as to whether the central tendency of the ensemble is an optimal predictor. We report results from the probabilistic projections using weighted and un-weighted (equivalent to equal weights) ensembles and discuss some of the implications of the results in terms of the ability to improve decision-making.

3. Implementing BMA with GCMs

a. Using BMA to Quantify Structural Uncertainty

We employ Bayesian Model Averaging (BMA) to derive probabilistic projections of agro-climate indices. BMA was originally developed as a means to account for uncertainty between competing statistical models so as to avoid relying on a single “best” model (Kass and Raftery 1995). Rather than choosing a single model amongst the set of models and using this model for prediction, the prediction, y, is instead conditioned on the entire set of models:

\[ p(y) = \sum_{k=1}^{K} p(y|M_k) p(M_k | y^T) \]  

(1)

Here the prediction is a PDF, \( p(y) \), that is modeled as the summed products of the individual model PDFs, \( p(y|M_k) \), and the probability, \( p(M_k | y^T) \), that the model \( M_k \) is correct given the training data \( y^T \) (Raftery et al. 2005). R05 point out that equation (1) is a restatement of the
law of total probability; that is, if we restrict the sample space of possible models (or GCMs) to the set $K$ (an arguably strong assumption), then our predictive PDF $p(y)$ is conditional on the probability of each given the data, and that this probability across our considered set of models sums to unity. Thus each model has a probability associated with it that can be thought of as a model weight, with the model weights summing to unity. We emphasize however that if $K$ is only a small subset of the possible model representations of the earth’s climate then $p(y)$ can be truncated if the models are not well-dispersed compared to the true predictive PDF.

Equation (1) can be rewritten then in the context of model weights for an ensemble of GCMs:

$$p(y|M_1, ..., M_k) = \sum_{k=1}^{K} w_k g_k(y|M_k),$$  \hspace{1cm} (2)

where the weights $w_k$ are interpreted as the probability of model $k$ being the best ensemble member given the training data $y$ and $g_k(y|M_k)$ is the conditional PDF of $y$ given the hindcast GCM prediction $M_k$. The distribution of $g_k(y|M_k)$ is assumed to follow a normal distribution when appropriate, such as for temperature (and for all three agro-climate indices). When this assumption is not appropriate, such as for precipitation, other distributions (e.g. the gamma distribution) are used to approximate this predictive PDF (Sloughter et al. 2007).

R05 use a simple bias correction formula $a_k + b f_k$ to estimate the mean of the PDF $g_k(y|M_k)$ for model $k$. The bias correction is estimated by linear regression of the training dataset onto the models. R05 note that other estimates for the mean of the PDF are possible. The standard deviation $\sigma$ is estimated by maximizing a likelihood function that is a combination of the expected probability of the model $k$ being the best model, the resulting model weight $w_k$, and the original distance between the model forecasts and observations. The sharpness of the posterior PDF and the variance estimate for each weighted GCM is given by:
\[ \lambda(y|M_1, \ldots, M_K) = \sum_{k=1}^{K} w_k(M_k^B - \sum_{i=1}^{K} w_i M_i^B)^2 + \sigma^2, \]  

(3)

where \( \lambda(y|M_1, \ldots, M_K) \) is the variance of the climate projection given the GCMs \( M_k \) and \( M_k^B \) is the mean of each GCM’s predictive PDF, \( g_k(y|M_k) \), that could be the bias-corrected version of the original model \( M_k \) (given by \( a_k + b_k f_k \) in M05). The combination of bias-corrected means, model weights, and variance estimates yields the predictive distribution in (2), which is a weighted PDF based on model fidelity to observations over time and/or space.

The terms on the right-hand side of equation (3) capture two kinds of uncertainty (cf. the discussion in R05). The first term \([ (M_k^B - \sum_{i=1}^{K} w_i M_i^B)^2 ]\) describes the ensemble spread, i.e. the between-model variance. The second term \((\sigma^2)\), describes the within-model predictive variance based on model fidelity to the training data. Thus BMA provides a way to quantify the structural uncertainty between models as well as the within-model prediction variance. This is a more rigorous treatment of ensemble uncertainty compared to using the ensemble range or standard deviation. Note that this approach still neglects key uncertainties (e.g. parametric uncertainty).

b. Using Training Data for GCMs

There is a significant methodological hurdle to extending BMA from numerical weather models producing forecasts on the scale of hours or days to general circulation models that operate on time-scales of decades or even centuries. This difference in temporal scale makes it difficult to develop a set of observable training data that provides a physically meaningful comparison with the GCM output. In the R05 implementation of BMA for use in weather forecasting the evaluation of the training data is straight-forward. This is because numerical weather prediction is accomplished through a constant updating of initial conditions with new
observations. Thus models with predictive skill should have coherence with both the statistical properties of the climate field and the actual observed values (i.e. temporal and spatial coherence). However no such concurrent spatio-temporal coherence is expected for the CMIP3 ensemble; that is, there is no common set of initial values between the observed and simulated climate state. Thus, the GCMs will produce different evolving atmospheric patterns that produce different daily values for the agro-climate indices. Other GCM intercomparison projects such as AMIP (Atmospheric Model Intercomparison Project; Gates et al. 1998) do force some coherence between model and observed internal variability by prescribing observed sea-surface temperatures which are the dominant driver of large-scale weather patterns. This allows for meaningful comparisons of correlation-based skill measures such as RMS error or Taylor skill scores. There is no such expected coherence between short-term observed and simulated short-term patterns (i.e. interannual) in the CMIP3 ensemble. In other words, the actual observed meteorological conditions for a point in time have no relation to a GCM’s meteorological conditions (except by chance). However, the 20C3M experiment performed by the CMIP3 models does specify historic climatic forcings such as greenhouse gas emissions and (for some GCMs) volcanic and solar forcings. Therefore, the scale and timing of longer-term climatic changes due to short-term forcings (e.g. volcanoes and aerosols), long-term forcings (solar and anthropogenic emissions), and the magnitude of unforced climatic variability should be comparable between transitory simulations and the observations.

Ideally one would examine long-term climatic trends and filter out the short-term variability in order to evaluate model performance over time. However, for these three agro-climate indices the required daily data for most CMIP3 models only cover the period 1961-2000 (the downscaled dataset extends the time period to 2010). In addition, it is desirable to have a
large training dataset to increase the degrees of freedom when calibrating the BMA model to avoid overfitting. The agro-climate indices are calculated as annual aggregates of daily data, so there is a maximum of 50 values for each downscaled GCM in the ensemble. Estimating the observed long-term trend is difficult for such a short time-series because the effects of internal variability and uncertain initial conditions can be considerable in this case and can lead to spurious results. These competing factors (i.e. independent inter-annual variability between GCMs and observations and the absence of a long-term dataset) must be balanced when constructing a Bayesian Model Averaging analysis.

4. Methods
a. Observation Data

We use observations of the three agro-climate indices (frost days, thermal time, heat stress index) derived in Terando et al. (2011) based on temperature observations from the National Climatic Data Center’s Global Historical Climatology Network (GHCN; Peterson and Vose 1997). Raw daily observations are converted to 5° by 5° grid-boxes using thin-plate splines (Haylock et al. 2008) before aggregating up to a sub-continental region covering the continental United States, southern Canada and northern Mexico. Details for the interpolation and quality-control measures used to create the observational dataset are given in T11. We calculate agro-climate indices using standard definitions for frost days, thermal time, and heat stress when applied to phenological growth and development of maize. Frost days are defined as the annual number of days where the minimum temperature is below 0 °C. We define thermal time using the remainder method (Tollenaar et al. 1979; Feng and Hu 2004):
\[ TT = \sum_{G_a} \frac{(T_{max} + T_{min})}{2 - T_l}, \]  (4)

where \( TT \) is the thermal time, \( G_b \) and \( G_e \) are the beginning and ending dates of a standard growth period for maize (we use April 1 through October 31), \( T_{max} \) and \( T_{min} \) are the maximum and minimum daily temperature respectively, and \( T_l \) is the limiting temperature where the upper and lower limits define the range of crop growth. We use minimum and maximum thresholds of 10 °C and 30 °C, respectively for thermal time and a threshold of 30 °C for the heat stress index, above which all accumulated degrees count towards a day’s value (with a value of zero if the maximum temperature is below 30 °C).

### b. Empirically Downscaled GCMs

GCMs are a valuable tool for understanding the potential climate response to anthropogenic and natural forcings. However, the coarse resolution of the models limits their applicability to regional impact studies because of the inability to simulate sub-grid scale processes and the localized effects of orography and land use patterns (Fowler et al. 2007). In addition, the coarse scale of the models can result in truncated distributions of temperature and precipitation, especially for shorter time-scales such as the daily time-scales used to calculate agro-climate indices. This is again related to the omission of fine-scale processes that have a large influence on the frequency of extreme events (Diffenbaugh et al. 2005). Downscaling is commonly used to address these issues and refers to the set of techniques that attempt to model or capture the sub-grid scale processes absent in GCMs (cf. Wilby and Wigley 1997 for a review of downscaling techniques). Given the importance of examining the tails of the temperature distribution (e.g. Schlenker and Roberts 2009) and the high temporal resolution required in our
study we employ an ensemble of downscaled GCMs to construct projections of the selected agro-climate indices.

We use output from six downscaled GCMs that were originally part of the World Climate Resource Project (WCRP) Coupled Model Intercomparison Project (CMIP3) multi-model dataset (Meehl et al. 2007c). The output from these GCMs is part of a new dataset of downscaled climate models for North America (Stoner et al. In Prep). The GCMs are statistically downscaled, meaning an empirical relationship is developed between local observations and large-scale atmospheric patterns that are then related to the coarse GCM output. There are several advantages to using this dataset of downscaled models: (1) this particular downscaling method known as Statistical Asynchronous Regression (SAR; O’Brien et al. 2001) is designed to better reflect the distribution of daily temperature extremes, (2) the GCM output are a continuous time series from 1961-2099 which allows for a longer calibration period (i.e. 1961-2010) between models and observations for the BMA, and (3) the downscaled dataset contains four GCMs with output for the A1Fi GHG emissions scenario (the other two GCMs are part of the A2 emissions scenario, which also is consistent with a future of high GHG emissions). This scenario (A1Fi) most closely tracks the current trajectory of anthropogenic GHG emissions (Raupach et al. 2009). We use the full six-member ensemble to estimate the BMA weights and then restrict our projections to the four-member A1Fi ensemble.

d. Bootstrapped Trend Analysis

We construct our training dataset using a 50-year trend estimate (1961-2010) for each agro-climate index. We use a boot-strapping procedure, known as a sieve bootstrap (Bühlmann 1998) to estimate the associated uncertainties. Accordingly, we fit a series of polynomials...
ranging from a linear fit to a sixth-order polynomial, to the 50-year observation period and choose the best fit by selecting the model with the lowest Akaike Information Criterion (AIC) value. We then use bootstrap resampling of the residuals from this model to estimate the variance around the best-fit trend. We test for potential autocorrelation in the residuals using standard diagnostics. The autocorrelation function and partial autocorrelation function indicate that the residuals are white (i.e. no evidence of autocorrelation), the standard errors around the fitted AR terms (we tested both AR1 and AR2 models) show evidence for uncorrelated residuals, and the AIC test suggests a pure trend model with no AR term is the best model. Thus we proceed under the assumption of IID (independent and identically distributed) residuals. Taking these residuals we create 1,000 bootstrap residual samples and superimpose these model draws on the best fit polynomial and then re-estimate the trend once again using AIC to select the model. The results are 1,000 bootstrap estimates of the low-frequency change for the 50-year period 1961-2010. Finally, we estimate the trends in the GCM ensemble using the same procedure. There is evidence of serial correlation in several of the simulated agro-climate trends, and we use the best-fit AR model (determined by AIC) accordingly to generate bootstrap replicates that are added back onto the best-fit polynomial. We repeat this procedure for the projections using two 50-year periods, 2011-2060 and 2050-2099.

c. Bayesian Model Averaging

We approximate the predictive PDF $g_k(y|M_k)$ with a weighted PDF of normal distributions for each downscaled GCM trend. Since we are examining changes in agro-climate indices over a time interval, the normal distribution satisfactorily represents the distribution of bootstrapped trends. We use the 50-year bootstrapped trends from the GCMs for the model mean.
(M^R_k). In this study the SAR downscaling technique applied to the GCMs serves as our bias correction in that the model variance and mean are transformed and more closely match the observations over the period of model calibration. Therefore, we do not apply any additional bias-correction measures.

We use the R package ensembleBMA (Fraley et al. 2010) to estimate the BMA weights by comparing the cumulative distribution function (CDF) of the 1,000 bootstrap trend estimates to the corresponding GCM trend estimates. As in R05, we use the EM Algorithm (Dempster et al. 1977) to calculate the posterior weighted PDF centered on individual GCM hindcasts. These equate to the model means (M^R_k) in g_k(y|M_k) and are given by the sorted bootstrap trends in the CDF (Figure 2). The posterior predictive variance (σ^2) and the model weights are jointly estimated by maximum likelihood as in R05 for the model in Equation (2). Each model σ is a function of the goodness-of-fit between the GCM CDF and observed CDF and the set of weights that maximize the log-likelihood function. We measure goodness-of-fit using the squared deviations between the sorted trends for GCMs and observations. For the case where we force equal model weights, the likelihood function converges almost immediately to values very close to the initial guess for g_k(y|M_k) based on the goodness-of-fit of the model and the inter-model variance.

5. Results
   a. Historic and Projected Changes in Agro-Climate Indices

The GCM hindcasts for frost days are similar to the observed trends and variability. There appears to be a bias towards greater warming in the mean and trends for thermal time and the heat stress index (Figure 1). The increasing effects of anthropogenic GHG emissions are
reflected in the GCM output by corresponding warming trends in the agro-climate indices. In addition there is some separation between the two emission scenarios by the end of the 21st century, especially for the heat stress index. One key feature of this comparison is that the considerable future trends projected by the models have, thus far, not been seen to a large extent in the model hindcasts or in the observations. Thus, the current observations may be poor constraints on system behavior if the sensitivity of the signal to anthropogenic perturbations is large.

The positive bias in the GCM hindcasts is apparent in the bootstrap trend estimator as well (Figure 2). The bootstrap estimator of the observed trends indicates a warming trend for frost days and thermal time but less evidence for such warming in the heat stress index. There is a general GCM bias towards warming trends greater than the observed trends which is most pronounced for the heat stress index (Figure 2c) and less pronounced for frost days (Figure 2a). While there is a consistent warm bias across most models, there is more variability in the shape of the CDF curves with some models closely matching the shape of the observed trend CDF.

b. BMA Weights and Out-of-Sample Model

We use the empirical cumulative distribution functions to estimate BMA weights and predictive variances for the six-member model ensemble. The resulting model weights and variances are used to derive probabilistic projections of the three agro-climate indices for the four-member A1Fi ensemble across North America. Specifically, we calculate six PDFs for the following groups:

- Three agro-climate indices (frost days, thermal time, heat stress index) for the A1Fi emissions scenario
Two projections based on 50-year trend estimates for the periods 2011-2060 and 2050-2099 based on the polynomial-fitting procedure.

The common result across the three agro-climate indices’ posterior model weights is for a large majority of the weight to be placed on one or two GCMs (Table 1). Three models (ccsm3.0, cgcm3.1 (T63), and gfdl-cm2.1) receive weights less than 0.01 for all agro-climate indices. The cgcm3.1 (T47) model received almost the entire ensemble weight for heat stress index (0.99) but there are no projection results for the A1Fi emissions scenario, therefore we redistribute the remaining model weights across the other GCMs. We return to this issue of very high weights being placed on one model in the discussion section.

We now show the results for out-of-sample predictions based on calibration and prediction periods that overlap with the total observational record. For this test we use the same polynomial fitting procedure over the 25-year period 1961-1985 for calibration and 1986-2010 for prediction. We compare this observed change to the GCM predicted change over the same 25-year period. Ideally the out-of-sample tests would have at least as long a period for calibration as was available for the full BMA fitting exercise. Unfortunately this is not possible since we are limited to a 50-year period of overlapping observations and GCM hindcasts. Thus the trend estimates are more likely to be influenced by random internal variability and the anthropogenic perturbation of the climate system will have less of an impact. Nevertheless, the two consecutive 25-year periods provide one useful test to gauge the risk of overfitting when BMA is applied to GCMs at this spatial scale and for these agro-climate indices.

For frost days and the heat stress index the out-of-sample prediction tests indicate overconfidence when using the BMA predictive PDF (Figures 3a, 3b, 5a, 5b). Overconfidence in this case occurs when the 95% prediction interval for the agro-climate change period 1986-2010
does not contain the verifying observation over the same time period. This holds whether using the four-member A1Fi ensemble or the full six-member ensemble. Only the thermal time observed trend is within the 95% interval of the BMA PDF (Figure 4a and 4b). For all cases however, the verifying observation is outside the range of the four-member ensemble but is within the six-member ensemble.

There are several potential explanations for the apparent overconfidence of the BMA projections. First, there is positive bias in the model trends compared to the observed bootstrap trends (Figure 2). The observed trends are generally outside the range of the simulated trends. It follows then that the models fail to reproduce the continuation of modest observed warming trends from 1986-2010. Second, the ensemble is small, which reduces the ability to sample multiple instances of internal climatic variability as well as the full range of climate sensitivity to natural and anthropogenic forcings. Third, there is uncertainty about how well the downscaling method captures the local climatic variability and relates it to large-scale dynamics.

In contrast to the PDFs derived from the BMA calibrated weights, the predictive PDFs based on equal model weights all contain the verifying trends within the 95% interval. Notably, this is the case for both the four-member and six-member ensembles. However the reduction in overconfidence also comes at the expense of a possibly underdispersed predictive PDF. This is due to the fact that the 95th percentile of bootstrapped trends for the calibration period (1961-1985) is narrower than the equally weighted 95% BMA PDF and yet also covers the verifying observations (Figure 6). Therefore, for near-term forecasts (e.g. ten to thirty years), the naïve forecast given by climatology may be a more skillful predictor of future trends (defined as whether the verifying observation is within the 95th percentile of the observations), than the very conservative equally weighted BMA PDF or the potentially overconfident calibrated BMA PDF.
c. Probabilistic Projections

Figure 7 shows the probabilistic projections for the A1Fi emissions scenario for two 50-year periods: 2011-2060 and 2040-2099. The projected anomalies are in relation to each GCM’s mean agro-climate values over the period 1981-2010. The 95% projection intervals are shown as rectangles above the shaded PDFs along with the GCM point projections and the four-member ensemble mean. The high warming rates associated with the high GHG emissions specified in the A1Fi scenario are reflected across the agro-climate indices as well. However, there are some differences between the three indices. Both the frost day and thermal time PDFs show bimodality in the projection, owing to large projected differences between two models that both received non-negligible posterior weights. The heat stress index PDFs are unimodal since nearly all available ensemble weight is placed on one model. In both time periods almost the entire area of the PDFs is greater than zero (for thermal time and heat stress index) and less than zero for frost days. This indicates that even with a wide uncertainty estimate, the expected probability of experiencing a warming trend as opposed to a cooling trend (or no trend) for these indices is high. On the other hand, the poor out-of-sample performance of the calibrated BMA approach suggests that considerable caution is warranted in interpreting these projections (Figure 6).

The equally weighted probabilistic projections as expected show more uncertainty than the BMA PDFs (Figure 7d, 7e, and 7f). However, even with this greater uncertainty, the 95% projection intervals are completely outside of the zero-trend or cooling-trend projection regions. Thus the analyzed models, data, and our analysis suggest a high degree of confidence that there will be a warming trend for these three agro-climate indices. The bimodality seen in the BMA frost days prediction (Figure 7a) remains in the equally-weighted PDF (Figure 7d), largely because of small predictive variance estimated for two GCMs. Since these two models had
nearly equal weights in the BMA PDF (weights of 0.39 and 0.61), the bimodality is maintained in the equally weighted PDF. The largest change between the BMA and equally weighted PDFs is seen in the heat stress index (Figure 7f). The PDF changes from unimodal to bimodal or multimodal due to the relatively equal predictive variances for the four GCMs.

Table 2 summarizes the uncertainty estimates for the probabilistic projections and point projections. For frost days, the 50% projection interval is similar for all three projections (calibrated BMA, equally weighted BMA, and ensemble range). Interestingly, the 50% projection interval is almost twice as large for the BMA thermal time projection compared to the equally weighted projection. This occurs because one of the four models projects smaller thermal time anomalies but receives the majority of the weight in the calibrated BMA projection. The 50% projection interval is shifted towards this model while also including the other models in the ensemble. In contrast the equally weighted probabilistic projection is centered towards the ensemble mean since multiple GCMs are clustered together. The ensemble range for heat stress index is greater than all three BMA prediction intervals and also greater than the equally weighted 50% projection interval. As in the cross-validation test, the equally weighted PDF has wider uncertainty bounds than the ensemble range.

6. Discussion

Our analysis indicates consensus among the climate model projections of substantial warming over the next century for all three agro-climate indices. This result holds regardless of whether the raw ensemble is used to characterize structural uncertainty or if Bayesian Model Averaging is used to create a probabilistic projection. Using equal weights did not change the
overall picture of future warming, although it did produce the widest uncertainty bounds for the
95% and 99% prediction intervals.

We have shown that using BMA to derive probabilistic climate change projections has an
impact on the structural uncertainty estimate. However, this impact does not always manifest in a
consistent way across all cases. As such, whether the forecast probability density function is
sharp, diffuse, multi-modal, or unimodal depends on how well the model hindcasts match the
observations over the training period. In the original use of Bayesian Model Averaging by
Raftery et al. (2005) for short-term numerical weather forecasting, this is not a problem and in
fact is the desired requirement for the model weighting process so that the posterior PDFs can be
both calibrated and sharp (Gneiting et al. 2007). But as stated earlier, this poses a problem when
attempting to more rigorously quantify projection uncertainty with a small model ensemble that
is not constantly updated with observations to create comparable temporal comparisons.

We address this issue by filtering and dampening internal variability to enhance the
signal from forcings that are temporally congruent in models and observations. This approach
still faces several challenges, as evidenced by the poor cross validation results as well as the
extremely high weights placed on a few models that contribute to overly calibrated projections.
These issues are related to the choice of the training dataset. The deterministic short-term
forecasts used in Raftery et al. (2005) simulate processes that can be predicted with high
accuracy and precision given enough observations and information to solve the equation of state
of the atmosphere. That is, getting the forecast 'right' is possible in terms of a forecast for day $x$
matching the observed value for day $x$. For surface temperature and precipitation, the forecast
errors are often equally distributed among the ensemble members and that the amount of error is
often correlated with the spread of the ensemble (Raftery et al. 2005). In such cases more
equality is expected amongst the forecast models if the biases and errors are randomly distributed amongst the models. For other variables however (such as atmospheric pressure), the error variance is small compared to forecast values (note small error variance does not necessarily equate to small forecast errors). In this case the BMA approach will assign a majority of weight to one model since little additional information is gained from the other models in the ensemble (Hamill 2007; Wilson et al. 2007). Similarly in our analysis, we limit the impact of internal variability in our comparison to reduce the possibility that a GCM will undeservedly receive too high or low of a weight. Consequently we also reduce the degrees of freedom available for the BMA as we are now comparing sorted CDF values.

What then, are the alternatives to assessing predictive uncertainty for long-term GCM projections and using the information present in observations? Other methods approach the problem by removing consideration of the evolving anthropogenic signal and instead focus on model precision and accuracy (i.e. model variance and model mean, respectively; cf. Tebaldi et al. 2005, Smith et al. 2009). However, removing temporal comparisons in the skill assessment precludes consideration of how effective the models are able to simulate the timing and rate of climate change. It is, thus far, an open and hotly debated question whether model weighting increases GCM predictive skill at all given the constraints of the observations and the time-scales of interest for climate projections (e.g. Brekke et al. 2008; Santer et al. 2009, Weigel et al. 2008; Knutti et al. 2010, Knutti 2010). As a result, some studies have eschewed model weighting and assign equal weights to all ensemble members (e.g. Meehl et al. 2007b). As our study demonstrates, this approach can outperform BMA in terms of reducing overconfidence (Figure 6). Given the small sample size available, this would seem to be a prudent way forward for characterizing projection uncertainty in a probabilistic manner while reducing the risk of
overconfident predictions. But even this method involves implicit model weighting choices because the ensemble is finite and older generation climate models are typically excluded because it is assumed they are less accurate than the most current generation (Knutti 2010). This assumption may or may not be accurate depending on the climate variable, scale, and region of interest. Thus, there are complex tradeoffs between sharpness, calibration, and overconfidence when using past observations and model hindcasts to inform projection uncertainty.

7. Caveats

Our approach is subject to several caveats. First, our results are sensitive to how we choose to derive the model weights. This is shown in previous studies to be a constraint on any type of study that uses model weighting or model culling (Santer et al. 2009; Knutti et al. 2010). Of primary concern is how to measure model skill in the context of temporal trends that do not share the same sequences of internal variability. Our particular skill score used to estimate optimal model weights (root square error between bootstrap estimates of observed and simulated trends) is very simple and we realize it is possible to construct many more comparisons based on different statistical or physical attributes. Constructing the “best” set of optimal weights is an area of ongoing and fruitful research (Wilks 2001; Knutti 2010; Räisänen and Ylhäisi 2011).

The posterior model weights are generally biased towards strong weighting of a few models rather than diffuse weighting of all models in the ensemble. There is a risk that the posterior PDFs will be underdispersed and overconfident due to our decision to reduce the influence of internal variability on the training dataset trends. As more observations accumulate through time, the weighting should become more robust. However, this also points to the need for GCMs with longer hindcast time series for these impact-relevant climate variables.
Additionally, efforts that use prescribed sea-surface temperatures as initial conditions such as the AMIP (Gates et al. 1998) or CLIVAR (Scaife et al. 2009) projects may also serve as useful surrogates for temporally-consistent sets of training data; especially as it relates to model skill in simulating the internal variability of the climate system.

8. Conclusions

We perform a Bayesian Model Averaging (BMA) analysis to derive probabilistic projections for three agro-climate indices (frost days, thermal time, and heat stress index). The projections are developed for North America for two time periods in the 21st century using a very high ‘business as usual’ emissions scenario. We use a bootstrap trend analysis to develop the training dataset as part of the BMA procedure which shows differing model performance depending on the agro-climate index. A cross-validation test indicates that using calibrated weights from the BMA analysis results in overconfident predictions, similar to the GCM point projections. However, a much wider uncertainty bounds with no evidence of overconfidence is found when using BMA to derive an equally weighted probabilistic prediction. The results of our analysis suggest that at this time it appears that using a combination of equally-weighted GCMs along with the BMA estimate of the predictive variance is the best strategy to reduce the risk of overconfident projections of impact-relevant climate indices.

There is agreement and high confidence among all the analyzed models and the probabilistic projections that there will be large warming trends for these agro-climate indices. These results suggest large potential changes in the conditions that farmers will experience over the remainder of the 21st century. Of particular concern is the projected change in the heat stress index which has the potential to result in large damages to rain-fed agriculture, even with
increases in beneficial thermal time and reductions in frost days that increase the growing season length (Schlenker and Roberts 2009). Our results suggest that given our current emissions trajectory, adaptive measures will be required to adjust to changes in the climate system that are most relevant to agriculture.

Acknowledgements

This study was supported by the National Oceanic and Atmospheric Administration under U.S. Department of Commerce Agreement EL133E07SE4607. This work was partially supported by the National Science Foundation (SES-0345925) as well as the Penn State Center for Climate Risk Management. We thank M. Haran, S. Bhat and N. Urban for helpful feedback and discussion about BMA. Any opinions and errors are those of the authors.
References


Hamill, T., 2007: Comments on “Calibrated surface temperature forecasts from the Canadian
Ensemble Prediction System using Bayesian Model Averaging”. Monthly Weather
Review, 135, 4226-4230.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate

Hay, L.E. and G.J. McCabe, 2010: Hydrologic effects of climate change in the Yukon River

Hayhoe, K., C. Wake, B. Anderson, X-Z Liang, E. Maurer, J. Zhu, J. Bradbury, A. DeGaetano,
A.M. Stoner, and D. Wuebbles, 2008: Regional climate change projections for the
Northeast USA. Mitigation and Adaptation Strategies for Global Change, 13, 425-436.


Meteorology, 43, 170-181.

Kiktev, D., D.M.H. Sexton, L. Alexander, and C.K. Folland, 2003: Comparison of modeled and

Keller, K., A. Robinson, D.F. Bradford, and M. Oppenheimer, 2007: The regrets of

Knutti, R. 2010: The end of model democracy? An editorial comment. Climatic Change, 102,
395-404.


Liang, X-Z., J. Zhu, and K.E. Kunkel, 2008: Do CGCMs simulate the North America monsoon
precipitation season-interannual variability? Journal of Climate, 21, 4424-4448.

Lobell, D. B., and M.B. Burke, 2008: Why are agricultural impacts of climate change so
uncertain? The importance of temperature relative to precipitation. Environmental
Research Letters, 3, 034007.


Table Headers

**Table 1:** BMA model weights for six empirically downscaled members of the CMIP3 GCM ensemble based on observations and simulations from the 20C3M experiment. Weights are given three agro-climate indices (FR, TT, and HSI) corresponding to frost days, thermal time, and heat stress index. Asterisks indicate model weights less than 0.01.

**Table 2:** Point projections and uncertainty estimates for the BMA PDF, equally weighted PDF, and the four-member GCM ensemble for the A1Fi emissions scenario. Uncertainty estimates are shown for the probabilistic projections for the 50%, 95%, and 99% prediction intervals while the ensemble range is given for the raw climate model projections.
**Figure Legends**

**Figure 1:** Observed (heavy black line) and simulated (gray dashed and solid lines) agro-climate indices (frost days, thermal time, and heat stress index) for the period 1960-2099. Dashed gray lines show output from the two downscaled GCMs using the A2 emissions scenario and solid gray lines show the output for the four downscaled GCMs using the A1Fi scenario. These are the “raw” observations and model hindcasts and projections without the Bayesian Model Averaging applied.

**Figure 2:** Empirical CDF comparison between bootstrapped trends of GCM simulated agro-climate change (gray lines) and bootstrapped observed trend estimate (heavy black line) after applying best-fit polynomials to the period 1961-2010.

**Figure 3:** Out of sample predictive PDFs of frost days (and individual model PDFs) for the period 1986-2010 based on the 25-year calibration period from 1961-1985. Top row plots are the predictive PDFs based on the BMA model weights for the 4-member A1Fi ensemble (a) and the full ensemble (b). Bottom row plots (c and d) are the predictive PDFs using equal weights for the ensemble members. Abscissae refer to the agro-climate index change observed or predicted in 2010 relative to beginning of the trend period (1986). GCM point predictions are shown above PDFs as open circles with ensemble mean given as solid black diamond. The verifying observation and 95% prediction interval are given by the solid vertical line and thin dashed lines, respectively.

**Figure 4:** Same as Figure 3 except for thermal time trend predictions.

**Figure 5:** Same as Figure 3 except for heat stress index trend predictions.

**Figure 6:** Uncertainty estimates from the out-of-sample prediction test for frost days (a), thermal time (b), and heat stress index (c) based on the period 1961-1985 for calibration and 1986-2010 for prediction. Estimates are shown for the four GCMs in the A1Fi emissions scenario and include the raw prediction range (GCM Range), the 95% BMA PDF prediction interval (BMA 95%), the same 95% PDF interval using equal weights (Equal BMA 95%), and 95% climatology interval (Climatology 95%). The verifying observation is denoted in the plots by the vertical line.
Figure 7: BMA calibrated and equally weighted probabilistic projections of agro-climate index change for two 50-year periods relative to 2010 values based on the four GCMs in the A1Fi emissions scenario. Light gray PDFs and rectangles (corresponding to 95% projection interval) show projections for 2011-2060 and dark gray PDFs and rectangles are projections for 2040-2099. Also shown are the median projection (vertical dashed lines (2011-2060) and solid lines (2040-2099)), GCM point projections (open circles), and the ensemble mean (solid large diamond).
Table 1: BMA model weights for six empirically downscaled members of the CMIP3 GCM ensemble based on observations and simulations from the 20C3M experiment. Weights are given three agro-climate indices (FR, TT, and HSI) corresponding to frost days, thermal time, and heat stress index. Asterisks indicate model weights less than 0.01.

<table>
<thead>
<tr>
<th>n</th>
<th>Model</th>
<th>Model Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FR</td>
</tr>
<tr>
<td>1</td>
<td>cccsm3.0</td>
<td>**</td>
</tr>
<tr>
<td>2</td>
<td>cgcm3.1(T47)</td>
<td>**</td>
</tr>
<tr>
<td>3</td>
<td>cgcm3.1(T63)</td>
<td>**</td>
</tr>
<tr>
<td>4</td>
<td>gfdl-cm2.1</td>
<td>**</td>
</tr>
<tr>
<td>5</td>
<td>hadcm3</td>
<td>0.39</td>
</tr>
<tr>
<td>6</td>
<td>pcm</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Table 2: Point projections and uncertainty estimates for the BMA PDF, equally weighted PDF, and the four-member GCM ensemble for the A1Fi emissions scenario. Uncertainty estimates are shown for the probabilistic projections for the 50%, 95%, and 99% prediction intervals while the ensemble range is given for the raw climate model projections.

<table>
<thead>
<tr>
<th>Year</th>
<th>Index</th>
<th>BMA Projection Median</th>
<th>50%</th>
<th>95%</th>
<th>99%</th>
<th>Equally Weighted Projection Median</th>
<th>50%</th>
<th>95%</th>
<th>99%</th>
<th>Ensemble Projection Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011-2050</td>
<td>FR</td>
<td>-16</td>
<td>18</td>
<td>20</td>
<td>21</td>
<td>-30</td>
<td>17</td>
<td>25</td>
<td>30</td>
<td>-27</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>TT</td>
<td>316</td>
<td>289</td>
<td>493</td>
<td>576</td>
<td>427</td>
<td>168</td>
<td>485</td>
<td>582</td>
<td>442</td>
<td>294</td>
</tr>
<tr>
<td></td>
<td>HSI</td>
<td>267</td>
<td>46</td>
<td>135</td>
<td>177</td>
<td>224</td>
<td>137</td>
<td>396</td>
<td>506</td>
<td>224</td>
<td>204</td>
</tr>
<tr>
<td>2050-2099</td>
<td>FR</td>
<td>-33</td>
<td>29</td>
<td>31</td>
<td>32</td>
<td>-49</td>
<td>28</td>
<td>35</td>
<td>43</td>
<td>-47</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>TT</td>
<td>576</td>
<td>350</td>
<td>554</td>
<td>637</td>
<td>728</td>
<td>202</td>
<td>526</td>
<td>619</td>
<td>665</td>
<td>355</td>
</tr>
<tr>
<td></td>
<td>HSI</td>
<td>489</td>
<td>46</td>
<td>135</td>
<td>177</td>
<td>356</td>
<td>157</td>
<td>335</td>
<td>411</td>
<td>369</td>
<td>193</td>
</tr>
</tbody>
</table>
Figure 1: Observed (heavy black line) and simulated (gray dashed and solid lines) agro-climate indices (frost days, thermal time, and heat stress index) for the period 1960-2099. Dashed gray lines show output from the two downscaled GCMs using the A2 emissions scenario and solid gray lines show the output for the four downscaled GCMs using the A1Fi scenario. These are the “raw” observations and model hindcasts and projections without the Bayesian Model Averaging applied.
Figure 2: Empirical CDF comparison between bootstrapped trends of GCM simulated agro-climate change (gray lines) and bootstrapped observed trend estimate (heavy black line) after applying best-fit polynomials to the period 1961-2010.
Figure 3: Out of sample predictive PDFs of frost days (and individual model PDFs) for the period 1986-2010 based on the 25-year calibration period from 1961-1985. Top row plots are the predictive PDFs based on the BMA model weights for the 4-member A1Fi ensemble (a) and the full ensemble (b). Bottom row plots (c and d) are the predictive PDFs using equal weights for the ensemble members. Abscissae refer to the agro-climate index change observed or predicted in 2010 relative to beginning of the trend period (1986). GCM point predictions are shown above PDFs as open circles with ensemble mean given as solid black diamond. The verifying observation and 95% prediction interval are given by the solid vertical line and thin dashed lines, respectively.
Figure 4: Same as Figure 3 except for thermal time trend predictions.
Figure 5: Same as Figure 3 except for heat stress index trend predictions.
Figure 6: Uncertainty estimates from the out-of-sample prediction test for frost days (a), thermal time (b), and heat stress index (c) based on the period 1961-1985 for calibration and 1986-2010 for prediction. Estimates are shown for the four GCMs in the A1Fi emissions scenario and include the raw prediction range (GCM Range), the 95% BMA PDF prediction interval (BMA...
95%), the same 95% PDF interval using equal weights (Equal BMA 95%), and 95% climatology interval (Climatology 95%). The verifying observation is denoted in the plots by the vertical line.
**Figure 7:** BMA calibrated and equally weighted probabilistic projections of agro-climate index change for two 50-year periods relative to 2010 values based on the four GCMs in the A1Fi emissions scenario. Light gray PDFs and rectangles (corresponding to 95% projection interval) show projections for 2011-2060 and dark gray PDFs and rectangles are projections for 2040-2099. Also shown are the median projection (vertical dashed lines (2011-2060) and solid lines
GCM point projections (open circles), and the ensemble mean (solid large diamond).