Environ Ecol Stat (2014) 21:507–530
DOI 10.1007/s10651-013-0266-z

A hierarchical Bayesian model for forecasting state-level corn yield

Balgobin Nandram · Emily Berg · Wendy Barboza

Received: 9 September 2012 / Revised: 9 July 2013 / Published online: 27 September 2013

Abstract Historically, the National Agricultural Statistics Service crop forecasts and estimates have been determined by a group of commodity experts called the Agricultural Statistics Board (ASB). The corn yield forecasts for the "speculative region," ten states that account for approximately 85 % of corn production, are based on two sets of monthly surveys, a farmer interview survey and a field measurement survey. The members of the ASB subjectively determine a forecast on the basis of a discussion of the survey data and auxiliary information about weather, average planting dates, and crop maturity. The ASB uses an iterative procedure, where initial state estimates are adjusted so that the weighted sum of the final state estimates is equal to a previously-determined estimate for the speculative region. Deficiencies of the highly subjective ASB process are lack of reproducibility and a measure of uncertainty. This paper describes the use of Bayesian methods to model the ASB process in a way that leads to objective forecasts and estimates of the corn yield. First, we use small area estimation techniques to obtain state-level forecasts. Second, we describe a way to

Handling Editor: Ashis SenGupta.

B. Nandram
Department of Mathematical Science, Worcester Polytechnic Institute,
100 Institute Road, Worcester, MA 01609, USA
e-mail: balnan@wpi.edu

E. Berg (✉)
Department of Statistics, Iowa State University, Snedecor Hall,
Ames, IA 50011, USA
e-mail: emilyb@iastate.edu

W. Barboza
National Agricultural Statistics Service, Fairfax, VA 22030-1504, USA
e-mail: wendy.barboza@nass.usda.gov

adjust the state forecasts so that the weighted sum of the state forecasts is equal to a previously-determined regional forecast. We use several diagnostic techniques to assess the goodness of fit of various models and their competitors. We use Markov chain Monte Carlo methods to fit the models to both historic and current data from the two monthly surveys. Our results show that our methodology can provide reasonable and objective forecasts of corn yields for states in the speculative region.

Keywords Constrained model · Covariates · Cross validation · Metropolis-Hastings sampler · Small area estimation

1 Introduction

The National Agricultural Statistics Service (NASS) publishes monthly forecasts of corn yields in August through November. A final yield estimate is published in December after the crop is harvested. The forecasts and estimates are based on multiple surveys as well as auxiliary information about weather, planting dates, and crop condition. We describe a model-based approach to obtaining forecasts and estimates of state-level corn yields for the "speculative region" for corn, the ten states that account for 85 % of corn production.

Traditionally, the NASS corn yield forecast is determined subjectively by a group of commodity experts called the Agricultural Statistics Board (ASB). The ASB process incorporates an iterative benchmarking operation, where state-level corn yield determinations are manually adjusted so that the appropriately weighted sum of the state estimates is equal to a value for the speculative region established by the ASB. Because the ASB process is subjective, the traditional estimation procedures are not reproducible, are difficult to document precisely, and do not furnish a measure of uncertainty.

Wang et al. (2011) specify a model for the speculative region as a whole using the corn yield survey data aggregated to the level of the speculative region. The resulting "regional model" is a hierarchical Bayesian model that combines the multiple survey indications with auxiliary information. The Wang et al. (2011) model is discussed further in Sect. 2.

Modeling state-level yield estimators is more challenging than modeling the regional averages for two main reasons. One is that the survey estimators have higher variances at the state level than at the regional level due to smaller sample sizes. To improve the precision for states with small sample sizes, we specify a model called the "unconstrained state model," which has the form of a small area estimation model (e.g., Battese et al. 1988). The structure of the unconstrained state model allows borrowing information across states. A second challenge associated with state-level forecasting and estimation is that NASS methodologists desire a model-based procedure that mimics the iterative method currently used to produce forecasts and estimates operationally. The current procedure involves determining estimates for the states first, then determining an estimate for the region, and finally adjusting the state-level estimates so that their weighted sum is equal to the estimate for the region. To mimic the iterative process, we use a method similar to the method of Nandram and Sayit (2011) to restrict the probability space for the state-level yields so that the weighted sum of

the state-level yields is equal to the forecast obtained from the regional model. We call the restricted model a "constrained state model." See Datta et al. (2011) for a decision-theoretic approach to benchmarking.

Other studies have used Bayesian hierarchical models to combine multiple survey estimators. See Raghunathan et al. (2007) and Manzi et al. (2011), for example. A unique feature of our approach is that forecasts and estimates for a detailed level of aggregation (in this case, states) are obtained given a suitable model for a larger level of aggregation (in this case, the speculative region).

The topic of corn yield modeling is an active area of research among agronomists and economists. Roper and Wagstaff (2007) use satellite data to predict crop yields in Kansas and California. Tannura et al. (2008) and Kantanantha et al. (2010) develop models forecasting models for corn yields using covariates related to weather.

A feature that distinguishes NASS yield forecasting efforts from other investigations of model-based corn yield forecasting is the availability of the survey indications. The studies discussed in the previous paragraph use the final (end-of-year) published NASS yield estimates as the response variable in their analyses. The published NASS yield estimates are the outcome of the subjective ASB decision procedure and differ from the direct survey estimators. NASS forecasting efforts differ from others because NASS's main goal is to effectively combine the survey estimators with supplemental information about weather and crop condition to predict the true corn yield.

The rest of this paper is organized as follows. In Sect. 2, we discuss the NASS yield forecasting application in more detail. We discuss the structure of the data and previous attempts at model-based yield forecasting at NASS. In Sect. 3, we specify an unconstrained state model that obtains state-level forecasts using small area estimation techniques and compare various versions of the unconstrained state model. In Sect. 4, we specify a constrained state model as a way to restrict the weighted sum of state-level corn yields to equal a previously-determined forecast or estimate for the corn speculative region. We compare the forecasts based on the constrained and unconstrained state models in Sect. 4. We also compare the forecasts and estimates based on the constrained state model to the ASB forecasts and estimates in Sect. 4. Section 5 has concluding remarks.

2 Yield forecasting at NASS

In this section, we first describe the structure of the survey data and the available covariates in more detail. Then, we discuss previous efforts at model-based yield forecasting at NASS.

National Agricultural Statistics Service (NASS) conducts two sets of surveys in August through November to obtain data for forecasting the final corn yield. A large farmer interview survey, conducted in December after the corn is harvested, supports the final yield estimate. We are interested in the states in the "speculative region" for corn, which consists of ten states that account for approximately 85 % of corn production. Between 1993 and 2004, the speculative region consisted of the following seven states: Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio and Wisconsin. NASS added Kansas, Missouri and South Dakota to the corn speculative region in 2004. Due

to changes in data processing and data collection methods, the survey indications are not available for all states and years.

The Objective Yield Survey (OYS) is conducted in August through December. The largest corn fields identified during an area frame based survey conducted in June have the highest selection probabilities in the OYS sample design. For the OYS, enumerators measure characteristics of plants in sampled plots instead of interviewing farmers. Examples of plant characteristics for corn are the number of ears, the kernel width, and the grain weight. The OYS indications are based on models relating the measurements to final corn yields. The states in which the OYS is conducted comprise the "speculative" region. NASS conducted the OYS in Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio and Wisconsin from 1993 to 2010. NASS extended the OYS to Kansas, Missouri and South Dakota in 2004.

The Agricultural Yield Survey (AYS) is a farmer interview survey conducted in August through November. To reduce respondents' burden, the largest operators (who are sampled for many NASS surveys) are excluded from the sampling frame for the AYS. NASS began using a multivariate probability proportional to size sampling design for the AYS in 2001. The AYS indications are available for 2001–2010 for the seven original speculative states (Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio and Wisconsin), and for 2002–2010 for Kansas, Missouri and South Dakota.

A large farmer interview survey called the December Agricultural Survey (DAS) supports the final yield estimate. DAS indications for the original seven speculative states (Illinois, Indiana, Iowa, Minnesota, Nebraska, Ohio and Wisconsin) are available from 1996 to 2010. DAS indications for Kansas, Missouri and South Dakota are available starting in 1999. Given that the DAS is a probability-based survey conducted after the corn is harvested, an assumption that the DAS indications are unbiased is reasonable; see Wang et al. (2011).

The OYS and AYS indications are biased estimators of the true final yields. The potential reasons for the biases differ for the two surveys. One possible reason for the bias of the OYS indications is that the measurement process leads to a systematic overestimation of the plant density in a field. Two potential reasons for the biases in the AYS indications are pessimism on behalf of the farmers and exclusion of farms in the largest size stratum from the sampling frame. Large farms are conjectured to have higher average yields than moderately sized farms because of greater investment in advanced technology.

Crude estimates of the average biases of the OYS (or AYS) indications are the differences between the average of the OYS (or AYS) indications for a particular month and the average of the DAS indications, where the average is across states and years for which data are available. The average differences between the OYS indications and the DAS indications are respectively 13.96, 11.62, 13.13, 14.99 and 14.97 for August, September, October, November and December. The average differences between the AYS indications and the DAS indications are respectively -12.26, -13.42, -9.81, and -4.10 for August, September, October, and November. The OYS biases are approximately constant across the months, while the AYS biases are smaller in magnitude for October and November than for August and September.

In addition to the survey indications, external information about weather and crop progress are potential covariates. We restrict our attention to covariates that are fully

observed by August 1 to permit use of the same model throughout the forecasting season. The covariates in the selected model are the year (1993–2010), two covariates related to weather, and two covariates related to crop progress and condition. The weather covariates, obtained from NOAA (http://www7.ncdc.noaa.gov/CDO/CDODivisionalSelect.jsp), are the average July temperature and the average July precipitation. The other two covariates are items from the NASS Crop Progress and Condition Survey, a weekly survey of commodity specialists such as extension agents and Farm Service Agency staff. Respondents to the Crop Progress and Condition Survey provide information about planting dates, crop maturity, crop condition, and harvesting progress for crops in their areas. We use two variables from the Crop Progress and Condition Survey for our study. One is an estimate of the percent of corn planted by week 20 of the crop year. The other is the percent of respondents rating the conditions of the corn crop as good or excellent in week 30. Because week 30 of the crop year is near August 1, we refer to the condition rating as the "August rating." The selected covariates have been used in previous studies of corn yield. Two additional covariates are examined in Sect. 3: a drought index called "PMDI" and the number of growing degree days through August 1 (GDD). The GDD is defined as $\sum_{d=1}^{D} GDD_d$, where $GDD_d = \max\{0.5(T_{max,d} + T_{min,d}) - T_{base}, 0\}$, d indexes the date, $d = 1$ is for January 1, $T_{max,d}$ and $T_{min,d}$ are, respectively, the maximum and minimum temperatures on day d measured in degrees Fahrenheit, and T_{base} is 50 degrees Fahrenheit. We obtained the daily maximum and minimum temperatures from NOAA (ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/hcn/).

Previous efforts at model-based yield forecasting using NASS survey data are Keller et al. (2004) and Wang et al. (2011). Keller et al. (2004) and Keller and Olkin (2002) construct an estimated generalized least squares estimator using historical OYS and AYS data to estimate mean and variance parameters. Wang et al. (2011) extend the Keller et al. (2004) method by incorporating data for multiple months simultaneously and utilizing auxiliary information about weather and crop development.

As an initial attempt to obtain state-level forecasts, the regional model of Wang et al. (2011) was applied to the individual states separately. The posterior coefficients of variation of the regression coefficients and variance parameters based on the individual state models are sometimes larger than 100 %, suggesting that sample sizes at the state-level are too small to support a separate model for each state. Another concern with applying the regional model to each state separately is that over-fitting in some states leads to posterior variances that are artificially small. It is more difficult to specify a model for the individual states than for the regional average, primarily because relatively small sample sizes at the state level lead to increased variability relative to the regional averages. As a result, it is difficult to distinguish meaningful differences between state-level yields from the effects of sampling and nonsampling errors. NASS methodologists feel that the effects of sampling and nonsampling errors are greatly reduced when the data are aggregated to the level of the speculative region.

Our approach is to specify a single model for all of the states simultaneously that assumes that certain mean and variance parameters are constant across the states. The model for the states is of the form of the models often used for small area estimation (e.g., Battese et al. 1988). The state model (specified in Sect. 3.1) is called the "unconstrained state model" because the forecast for the region based on the state model

does not necessarily equal the forecast for the region based on the regional model. To impose a constraint on the weighted sum of the state yields, we define a constrained state model in Sect. 4. The distribution of the state yields in the constrained state model is a conditional distribution given that the posterior mean of the weighted sum of the state yields is equal to the posterior mean from the regional model. Nandram and Sayit (2011) develop a similar method for imposing a restriction in the context of a beta-binomial model.

3 A small area model

Let $\boldsymbol{Y}_{kt\ell}$ be the vector of indications for survey k, year t and state ℓ, where $k = O$ for the OYS and $k = A$ for the AYS. More specifically, $\boldsymbol{Y}_{Ot\ell} = (Y_{Ot\ell 1}, \ldots, Y_{Ot\ell 5})'$ is a five-dimensional vector containing the OYS indications from August through December, and $\boldsymbol{Y}_{At\ell} = (Y_{At\ell 1}, \ldots, Y_{At\ell 4})'$ is a four-dimensional vector containing the AYS indications from August through November. The scalar $Y_{Dt\ell}$ denotes the DAS indication for year t and state ℓ. Let $\boldsymbol{S}_{kt\ell}$ denote the vector of estimated sampling variances for $k = O, A$, and let $S^2_{Dt\ell}$ be the estimated sampling variance for the DAS indication.

Let $z_{t\ell}$ be the vector of covariates for year t and state ℓ, where $t = 1$ is for 1993, $t = 18$ is for 2010, and $\ell = 1, \ldots, 10$. For the selected model,

$$z_{t\ell} = (1, t, PCP_{t\ell}, TMP_{t\ell}, PP20_{t\ell}, AGR_{t\ell})', \tag{1}$$

where $PCP_{t\ell}$ is the average July precipitation for year t and state ℓ, $TMP_{t\ell}$ is the average July temperature, $PP20_{t\ell}$ is the reported percent planted by week 20, and $AGR_{t\ell}$ is the August rating. We compare the selected model to two expanded models in Sect. 3.2.

3.1 An unconstrained state model

For the DAS indication, we assume the standard Fay-Herriott model (1979)

$$Y_{Dt\ell} \mid \mu_{t\ell} \sim \mathrm{N}\left(\mu_{t\ell}, S^2_{Dt\ell}\right), \tag{2}$$

where

$$\mu_{t\ell} = \boldsymbol{z}'_{t\ell}\boldsymbol{\beta} + \eta_{t\ell}, \text{ and } \eta_{t\ell} \sim N\left(0, \sigma^2_\eta\right). \tag{3}$$

For $k = O, A$, let

$$\boldsymbol{Y}_{kt\ell} \mid \mu_{t\ell} \sim \mathrm{N}\left\{\boldsymbol{1}_{M(k)}\mu_{t\ell} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k + \mathrm{diag}\left(\boldsymbol{S}_{kt\ell}\right)\right\}, \tag{4}$$

where $\boldsymbol{b}_k = (b_{k1}, \ldots, b_{kM(k)})'$,

$$\boldsymbol{\Sigma}_k = \mathrm{diag}\left(\sigma_{k1}, \ldots, \sigma_{kM(k)}\right) \mathrm{AR1}_{M(k)}(\rho_k) \mathrm{diag}\left(\sigma_{k1}, \ldots, \sigma_{kM(k)}\right),$$

$\mathrm{AR1}_{M(k)}(\rho_k)$ is an AR(1) correlation matrix of dimension $M(k)$ with correlation ρ_k, and $M(k)$ is the number of months for which survey k is conducted ($M(O) = 5$ and $M(A) = 4$). The $\boldsymbol{b}_k$ is a vector of biases for survey k, and $\boldsymbol{\Sigma}_k$ is a covariance matrix of nonsampling errors. The $\boldsymbol{b}_k$ and $\boldsymbol{\Sigma}_k$ are assumed to be constant across states ($\ell = 1, \ldots, L$) and years ($t = 1, \ldots, T$).

Diffuse proper priors are specified for the biases, regression coefficients, variances, and correlations. The priors for the elements of $\boldsymbol{b}_k$ and $\boldsymbol{\beta}$ are independent normal distributions with mean zero and variance 10^6. The priors for σ^2_{km} and σ^2_η are independent inverse gamma distributions with shape and scale parameters equal to 0.001. The correlation parameters have independent uniform priors with support of $(-1, 1)$. These priors are noninformative relative to the sample size and variability in the data. It is worth noting that because all priors are proper, under the mild assumption that the likelihood function is bounded, the joint posterior distribution must be proper. We also note that regression coefficients and the variance components are the same across states and years so that the priors for these parameters are dominated by the likelihood. In fact, we have redone the computations with the 0.001 in the priors for the variance components replaced by unity with virtually no change.

The joint posterior distribution of the parameters is,

$$
\begin{aligned}
f(\boldsymbol{\Omega} \mid \boldsymbol{Y}) \propto & \prod_{\ell=1}^{L} \prod_{k=O,A} \prod_{t=s(\ell,k)}^{T} \Big\{ |\boldsymbol{\Sigma}_k + \boldsymbol{S}_{kt\ell}|^{-0.5} \\
& \times \exp\Big(-0.5\big(\boldsymbol{Y}_{kt\ell} - \boldsymbol{b}_k - \mathbf{1}_{M(k)}\mu_{t\ell}\big)' \left(\boldsymbol{\Sigma}_k + \boldsymbol{S}_{kt\ell}\right)^{-1} \left(\boldsymbol{Y}_{kt\ell} - \boldsymbol{b}_k - \mathbf{1}_{M(k)}\mu_{t\ell}\right)\Big)\Big\} \\
& \times \prod_{\ell=1}^{L} \prod_{t=s(\ell,D)}^{T} \exp(-0.5(Y_{Dt\ell} - \mu_{t\ell})^2 S^{-2}_{Dt\ell}) \\
& \times \prod_{\ell=1}^{L} \prod_{t=\min\{s(l,k):k=O,A,D\}}^{T} (\sigma^2_\eta)^{-0.5} \exp\Big(-0.5(\mu_{t\ell} - \boldsymbol{z}'_{t\ell}\boldsymbol{\beta})^2 \sigma_\eta^{-2}\Big) \\
& \times \exp\Big(-0.5\boldsymbol{\beta}'\boldsymbol{\beta}10^{-6}\Big) \exp\Big(-0.5\boldsymbol{b}'\boldsymbol{b}10^{-6}\Big) \\
& \times \Big(\sigma^2_\eta\Big)^{-(1.001)} \exp(-\sigma_\eta^{-2}0.001) \prod_{k=O,A} \prod_{m=1}^{M(k)} (\sigma^2_{km})^{-(1.001)} \exp(-0.001\sigma_{km}^{-2}), \qquad (5)
\end{aligned}
$$

where $-1 \leq \rho_O, \rho_A \leq 1$, $\boldsymbol{\Omega} = (\boldsymbol{\mu}', \boldsymbol{\beta}', \sigma^2_\eta, \boldsymbol{b}'_O, \boldsymbol{b}'_A, \boldsymbol{\sigma}'_O, \boldsymbol{\sigma}'_A, \rho_O, \rho_A)'$, $M(O) = 5$, $M(A) = 4$, $s(\ell, k)$ is the first year for which we have indications for state ℓ and survey k ($k = O, A, D$), and $\boldsymbol{\sigma}_k = (\sigma^2_{k1}, \ldots, \sigma^2_{kM(k)})'$. We fit (5) using MCMC methods. The full conditional distributions used for the Metropolis-Hastings algorithm are given in the "Appendix". The unconstrained state model and the full conditionals for the Metropolis-Hastings algorithm are described in an internal NASS report (Wang et al. 2010).

We examined trace plots of the MCMC chain to assess the convergence of the Gibbs sampler. We discarded the first 5,000 iterations for burn-in. The jumping probability for the Metropolis-Hastings step is 30.3 %. Because ACF plots showed a high

autocorrelation in the generated samples for σ^2_{km} and ρ_k, we thinned the iterations by an interval of three. We have used 1,000 iterations to make inference. As we have used a single MCMC sequence, rather than multiple sequences, we have performed the Geweke test of stationarity on the 1,000 runs on the variance components and the regression parameters, and it shows that the 1,000 runs for these parameters are stationary.

3.2 Model comparisons

We use three criteria to compare alternative models. The first is the minimum posterior predictive loss developed in Gelfand and Ghosh (1998). For squared error loss and a weight factor of 100, the minimum predictive loss, $D = 100(101)^{-1}G + P$, where

$$G = \sum_{kt\ell m} \left\{E[y^{rep}_{kt\ell m} \mid \boldsymbol{Y}] - y_{kt\ell m}\right\}^2, \text{ and } P = \sum_{kt\ell m} V\left\{y^{rep}_{kt\ell m} \mid \boldsymbol{Y}\right\}.$$

We compute $E[y^{rep}_{kt\ell m} \mid \boldsymbol{Y}]$ and $V\{y^{rep}_{kt\ell m} \mid \boldsymbol{Y}\}$ using the properties,

$$E\left[y^{rep}_{kt\ell m} \mid \boldsymbol{Y}\right] = E\left[E\left[y^{rep}_{kt\ell m} \mid \boldsymbol{\Omega}, \boldsymbol{Y}\right] \mid \boldsymbol{Y}\right]$$

and

$$V\left\{y^{rep}_{kt\ell m} \mid \boldsymbol{Y}\right\} = E[V\left\{y^{rep}_{kt\ell m} \mid \boldsymbol{\Omega}, \boldsymbol{Y}\right\} \mid \boldsymbol{Y}] + V\left\{E\left[y^{rep}_{kt\ell m} \mid \boldsymbol{\Omega}, \boldsymbol{Y}\right] \mid \boldsymbol{Y}\right\},$$

where $\boldsymbol{\Omega}$ is defined after (5). The second criterion, DIC (Spiegelhalter et al. 2002), is defined,

$$\text{DIC} = 2E[D(\boldsymbol{\Omega}) \mid \boldsymbol{Y}] - D\left(E[\boldsymbol{\Omega} \mid \boldsymbol{Y}]\right),$$

where

$$\begin{aligned} D(\boldsymbol{\Omega}) = &\sum_{\ell=1}^{L} \sum_{t=s(\ell,O)}^{T} (\boldsymbol{y}_{Ot\ell} - \boldsymbol{b}_O - \mathbf{1}_5\mu_{t\ell})'(\boldsymbol{\Sigma}_O + \boldsymbol{S}_{Ot\ell})^{-1}(\boldsymbol{y}_{Ot\ell} - \boldsymbol{b}_O - \mathbf{1}_5\mu_{t\ell}) \\ &+ \sum_{\ell=1}^{L} \sum_{t=s(\ell,A)}^{T} (\boldsymbol{y}_{At\ell} - \boldsymbol{b}_A - \mathbf{1}_4\mu_{t\ell})'(\boldsymbol{\Sigma}_A + \boldsymbol{S}_{At\ell})^{-1}(\boldsymbol{y}_{At\ell} - \boldsymbol{b}_A - \mathbf{1}_4\mu_{t\ell}) \\ &+ \sum_{\ell=1}^{L} \sum_{t=s(\ell,D)}^{T} (y_{Dt\ell} - \mu_{t\ell})^2/S^2_{Dt\ell}. \end{aligned}$$

The third criterion, CV, defined in Wang et al. (2011), is

$$\mathrm{CV} = \sum_{k\in\{O,A\}} \sum_{m=1}^{M(k)} \sum_{\ell=1}^{L} \sum_{t=s(\ell,k)}^{T} \left| y_{kt\ell m} - E\left[y_{kt\ell m} \mid \boldsymbol{y}_{(ktlm)}\right] \right| + \sum_{\ell=1}^{L} \sum_{t=s(\ell,D)}^{T} \left| y_{Dtl} - E[y_{Dtl} \mid \boldsymbol{y}_{(Dtl)}] \right|$$

where $\boldsymbol{y}_{(ktlm)}$ is the vector of observations not including $y_{kt\ell m}$. The CV is computed using the importance sampling method described in Wang et al. (2011). The D, DIC, and CV are measures of global goodness of fit.

To evaluate the quality of the model in more detail, we examine deleted residuals. The deleted residual for observation $y_{kt\ell m}$ is defined,

$$r_{mkt\ell} = \left(y_{kt\ell m} - E\left[y_{kt\ell m} \mid \boldsymbol{y}_{(kt\ell m)}\right]\right) \left[V\left\{y_{kt\ell m} \mid \boldsymbol{y}_{(kt\ell m)}\right\}\right]^{-0.5},$$

where we compute $V\{y_{kt\ell m} \mid \boldsymbol{y}_{(kt\ell m)}\}$ using the definition, $V\{y_{kt\ell m} \mid \boldsymbol{y}_{(kt\ell m)}\} = E[y^2_{kt\ell m} \mid \boldsymbol{y}_{(kt\ell m)}] - E[y_{kt\ell m} \mid \boldsymbol{y}_{(kt\ell m)}]^2$, and the importance sampling method described in Wang et al. (2011) to compute the expected values.

We compare the selected model to two expanded models. One of the expanded models includes Palmer's Modified Drought Index (PMDI) and cumulative growing degree days (GDD). Previous studies have used squares of precipitation or temperature to approximate nonlinear relationships between these weather-related variables and final yields (e.g., Vado and Goodwin 2010; Irwin et al. 2008; Rosenzweig et al. 2002). We consider a model with squares of average July precipitation and temperature and an interaction between average July precipitation and average July temperature.

We compute D, DIC, and CV for the selected model and the two expanded models described above. As stated in (1), the covariates in the selected model are the year, the average July precipitation, the average July temperature, the percent planted by week 20, and the August rating. The values of D, DIC, and CV for the selected model are 238911.9, 7843.721, and 35.03, respectively. The values of D, DIC, and CV for an expanded model containing GDD and PMDI in addition to the covariates in the selected model are 241187.0, 7855.068, and 34.98, respectively. The values of D, DIC, and CV for an expanded model containing the squares and interactions of precipitation and temperature in addition to the covariates in the selected model are 240645.6, 7844.177, and 34.93, respectively. The values of D and DIC are smaller for the selected model than for the expanded models. The value of CV for the selected model is larger than the values of CV for both of the expanded models. Because the values of D, DIC, and CV are similar for all of the models, we conclude that expanding the model does not substantially improve the quality of the fit.

Figure 1 contains the deleted residuals for the DAS indications based on the selected model plotted against PMDI. The different symbols in Fig. 1 distinguish the different states. The residuals have approximate mean zero and do not show a clear trend as a function of PMDI. Plots of deleted residuals against the other potential additional

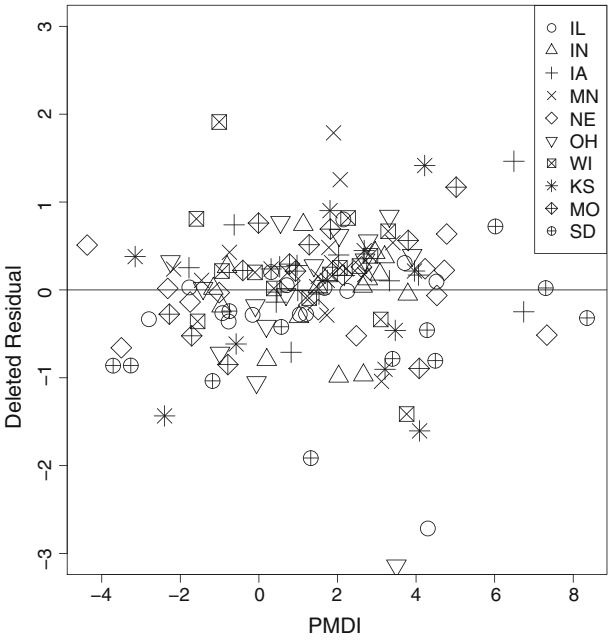

Fig. 1 Deleted residuals based on the unconstrained state model plotted against PMDI

covariates (GDD and squares and interactions of average July precipitation and temperature) also support the conclusion that expanding the model does not substantially improve the fit.

The extreme residual with a value below -3 is for Ohio for 2002. Keller et al. (2004) examine the 2002 data for Ohio. They explain that 2002 was "a year with much variability in the weather throughout the nation," and that "many areas of the Corn Belt, especially Ohio and Indiana, had lingering rain showers which delayed some corn and soybean planting." The percent planted by week 20 and the percent of respondents rating the conditions for growing corn as good or excellent for Ohio in 2002 are, respectively, 3.0 and 2.8 standard deviations below the corresponding averages.

One method for outliers, which has been investigated for estimation of crop yields at the county level, is to assume that the observed yields arise from a mixture of two normal distributions. The normal distributions have different variances, and the outliers are assumed to arise from the distribution with the larger variance. If the number of "outliers" is small, then the estimator of the larger variance may be unstable. For this reason, we did not pursue the mixture approach for handling outliers.

Because the survey indications are associated with spatial locations (states), it is natural to consider a model with spatial dependence. We fit a model to the data for the ten speculative states with spatially correlated random slopes and intercepts in the model for $\mu_{t\ell}$. Specifically, we fit a model where

$$\mu_{t\ell} = \boldsymbol{z}'_{t\ell}\boldsymbol{\beta} + \gamma_{o,\ell} + t\gamma_{1,\ell} + \eta_{t\ell}, \quad \eta_{t\ell} \sim N\left(0, \sigma_\eta^2\right),$$

where

$$\boldsymbol{\gamma}_o \sim \mathrm{N}\left(\mathbf{0}, (\boldsymbol{D}_w - \rho_o \boldsymbol{W})^{-1}\right), \boldsymbol{\gamma}_1 \sim \mathrm{N}\left(\mathbf{0}, (\boldsymbol{D}_w - \rho_1 \boldsymbol{W})^{-1}\right), \tag{6}$$

$\boldsymbol{\gamma}_o = (\gamma_{o,1}, \ldots, \gamma_{o,L})'$, $\boldsymbol{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{1L})'$, $\boldsymbol{W}$ is a matrix with element (i, j) equal to one if state i is adjacent to state j and zero otherwise, and $\boldsymbol{D}_w$ is a diagonal matrix with the column sums of $\boldsymbol{W}$ on the diagonal. (See Banerjee et al. 2004, Chapter 7 for a discussion of the spatial model specified by (6).) The prior for (ρ_o, ρ_1) is $(\rho_o, \rho_1) \sim \mathrm{Unif}[(-1, 1) \times (-1, 1)]$.

The 95 % credible intervals for the parameters of the spatial covariance matrix (ρ_o and ρ_1) do not differ significantly from zero. This is true because ten states are insufficient for estimating the parameters of the spatial covariance matrix. It is also possible that the covariates explain variability across space so that remaining spatial structure in the model residuals is negligible.

For the spatial model, the values of D, DIC and CV are, respectively, 240718, 7816 and 34.22, as compared with 238911.9, 7843.721 and 35.03 for the selected model. There are two main differences between the spatial model and the selected model. First, the spatial model has random intercepts and random slopes, while the selected model does not. Second, the parameters ρ_o and ρ_1 in the spatial model allow us to estimate spatial structure in the random effects, η_{tl}, while ρ_o and ρ_1 are assumed to equal zero in the non-spatial model. In this case, we conjecture that the relatively small decrease in the DIC and CV is due to the addition of the random intercepts and random slopes and not the addition of the spatial correlation parameter. Because the difference between the spatial model and the selected model is minor, we discard the spatial model from further analysis.

4 Constraining the small area model

Let $\boldsymbol{Y}_{kt}$ $(k = O, A)$, Y_{Dt}, and $\boldsymbol{z}_t$ denote the survey indications and covariates aggregated to the regional level for year t. Letting $h_{t\ell}$ be the acres harvested in corn for year t and state ℓ,

$$\left(\boldsymbol{Y}'_{Ot}, \boldsymbol{Y}'_{At}, Y_{Dt}, \boldsymbol{z}'_t\right)' = \sum_{\ell=1}^{L(t)} w_{t\ell} \left(\boldsymbol{Y}'_{Ot\ell}, \boldsymbol{Y}'_{At\ell}, Y_{Dt\ell}, \boldsymbol{z}'_{t\ell}\right)',$$

where $w_{t\ell} = h_{t\ell}(\sum_{\ell=1}^{L(t)} h_{t\ell})^{-1}$, $L(t)$ is the number of states for which data are collected in year t, $L(t) = 7$ for $t = 1, \ldots, 6$, and $L(t) = 10$ for $t = 7, \ldots, 18$. We compute an estimated sampling variance for a regional indication under an assumption that sampling is independent across states; $\boldsymbol{S}_{kt} = \sum_{\ell=1}^{L(t)} w_{t\ell}^2 \boldsymbol{S}_{kt\ell}$. The model of Wang et al. (2011) is a model for $(\boldsymbol{Y}'_{Ot}, \boldsymbol{Y}'_{At}, Y_{Dt})$, treating $\boldsymbol{S}_{kt}$ as a fixed value.

For the selected unconstrained state model discussed in Sect. 3.1, inference for the regional yield is conducted with respect to the posterior distribution of the state yields, $f_S(\boldsymbol{\mu}_t \mid \boldsymbol{Y})$, where $\boldsymbol{\mu}_t = (\mu_{t1}, \ldots, \mu_{t,L(t)})'$, $\boldsymbol{Y}$ is the vector of state-level survey indications, and S denotes the unconstrained state model. By definition, the true

regional yield for year t is $\mu_t = \sum_{\ell=1}^{L(t)} w_{t\ell}\mu_{t\ell}$, and one can conduct inference for the regional yield using the posterior distributions of the state yields. The posterior mean and variance of the regional yield for year t based on the unconstrained state model are, respectively,

$$\hat{\mu}_t^S = E_S\left[\mu_t \mid \boldsymbol{Y}\right] = E_S\left[\sum_{\ell=1}^{L(t)} w_{t\ell}\mu_{t\ell} \mid \boldsymbol{Y}\right] = \sum_{\ell=1}^{L(t)} w_{t\ell}\hat{\mu}_{t\ell},$$

and

$$\hat{V}_t^S = V\left\{\sum_{\ell=1}^{L(t)} w_{t\ell}\mu_{t\ell} \mid \boldsymbol{Y}\right\},$$

where $\hat{\mu}_{tl} = E_S[\mu_{tl} \mid \boldsymbol{Y}]$. It is not necessarily the case that $\hat{\mu}_t^S$, the posterior mean of the regional yield based on the unconstrained state model, is equal to the posterior mean of the regional yield based on the regional model. To mimic the iterative process currently used by NASS to produce corn yield forecasts and estimates, we desire the posterior mean for the regional yield based on the state model to equal the posterior mean for the regional model. To impose this restriction, we specify a constrained state model.

4.1 Constrained state model

For the constrained state model, the conditional distribution of $(\boldsymbol{Y}'_{Ot\ell}, \boldsymbol{Y}'_{At\ell}, Y_{Dt\ell})'$ given $\mu_{t\ell}$ is the model specified by (2) and (4), and (3) is replaced by a conditional distribution for $\mu_{t\ell}$ given the regional yield for year t. To specify the conditional distribution, let

$$\phi_t = \sum_{\ell=1}^{L(t)} w_{t\ell}\mu_{t\ell} - \theta_t^R, \tag{7}$$

where θ_t^R represents the regional yield for year t. Let

$$f\left(\mu_{t1}, \ldots, \mu_{t,L(t)-1}, \phi_t \mid \boldsymbol{\beta}, \sigma_\eta^2, \theta_t^R\right)$$

be the joint distribution of $(\mu_{t1}, \ldots, \mu_{t,L(t)-1}, \phi_t)$ implied by (3) if θ_t^R is known. From (3),

$$\begin{pmatrix} \mu_{t1} \\ \vdots \\ \mu_{t,L(t)-1} \\ \phi_t \end{pmatrix} \sim \mathrm{N}\left(\tilde{\boldsymbol{\mu}}_t, \begin{pmatrix} \sigma_\eta^2 \boldsymbol{I}_{L(t)-1} & (w_{t1}, \ldots, w_{t,L(t)-1})'\sigma_\eta^2 \\ (w_{t1}, \ldots, w_{t,L(t)-1})\sigma_\eta^2 & \sum_{\ell=1}^{L(t)} w_{t\ell}^2\sigma_\eta^2 \end{pmatrix}\right), \tag{8}$$

where $\tilde{\boldsymbol{\mu}}_t = (\boldsymbol{z}'_{t1}\boldsymbol{\beta}, \ldots, \boldsymbol{z}'_{t,L(t)-1}\boldsymbol{\beta}, \sum_{\ell=1}^{L(t)} w_{t\ell}\boldsymbol{z}'_{t\ell}\boldsymbol{\beta} - \theta_t^R)'$. By the definition of ϕ_t in (7), the yield for state $L(t)$ is $\mu_{t,L(t)} = w_{t,L(t)}^{-1}(\phi_t + \theta_t^R - \sum_{\ell=1}^{L(t)-1} w_{t\ell}\mu_{t\ell})$. The conditional distribution that replaces (3) for the constrained model is $f(\mu_{t1}, \ldots, \mu_{t,L(t)-1} \mid \phi_t = 0, \boldsymbol{\beta}, \sigma_\eta^2, \theta_t^R)$, which is obtained from (8) using properties of the multivariate normal distribution.

The constrained state model requires a prior distribution for θ_t^R. To impose the restriction that the posterior mean of the weighted sum of state yields is equal to the posterior mean from the regional model, we specify the prior distribution for θ_t^R to be a point mass at the posterior mean from the regional model. An unpublished NASS technical report (Berg et al. 2011) contains a discussion of a class of normal priors for θ_t^R. The priors for $\boldsymbol{b}_k$, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_k$, and σ_η^2 are the same as the priors used for the unconstrained state model. It is worth noting that because all priors are proper, under the mild assumption that the likelihood function is bounded, the joint posterior distribution must be proper. We also note that regression coefficients and the variance components are the same across states and years so that the priors for these parameters are dominated by the likelihood. In fact, we have redone the computations with the 0.001 in the priors for the variance components replaced by unity with virtually no change.

The vector of unknown parameters for the unconstrained state model is

$$\boldsymbol{\Lambda} = \left((\boldsymbol{\mu}_1^{[L(1)]})', \ldots, (\boldsymbol{\mu}_T^{[L(T)]})', (\boldsymbol{\theta}^R)', \boldsymbol{\beta}', \boldsymbol{b}'_O, \boldsymbol{b}'_A, \boldsymbol{\sigma}'_O, \boldsymbol{\sigma}'_A, \rho_O, \rho_A\right)',$$

where $\boldsymbol{\mu}_t^{[L(t)]} = (\mu_{t1}, \ldots, \mu_{T,L(t)-1})'$, and $\boldsymbol{\theta}^R = (\theta_1^R, \ldots, \theta_T^R)'$. Because the Jacobian of the transformation of $(\mu_{t1}, \ldots, \mu_{tL(t)})$ to $(\mu_{t1}, \ldots, \mu_{t,L(t)-1}, \phi_t)$ is a function of $\boldsymbol{w}_t$, the full joint posterior of the parameters is,

$$\begin{aligned}
f(\boldsymbol{\Lambda} \mid \boldsymbol{Y}, \boldsymbol{\phi} = \boldsymbol{0}) \propto & \prod_{t=1}^{T}\prod_{\ell=1}^{L(t)-1}\prod_{k\in S(\ell,t)} \{|\boldsymbol{\Sigma}_k + \boldsymbol{S}_{kt\ell}|^{-0.5} \\
& \times \exp(-0.5(\boldsymbol{Y}_{kt\ell} - \boldsymbol{b}_k - \boldsymbol{1}_{M(k)}\mu_{t\ell})'(\boldsymbol{\Sigma}_k + \boldsymbol{S}_{kt\ell})^{-1}(\boldsymbol{Y}_{kt\ell} - \boldsymbol{b}_k - \boldsymbol{1}_{M(k)}\mu_{t\ell}))\} \\
& \times \prod_{t=1}^{T}\prod_{k\in S(L(t),t)} \{|\boldsymbol{\Sigma}_k + \boldsymbol{S}_{ktL(t)}|^{-0.5}\exp\left(-0.5\boldsymbol{r}'_{ktL(t)}(\boldsymbol{\Sigma}_k + \boldsymbol{S}_{ktL(t)})^{-1}\boldsymbol{r}_{ktL(t)}\right) \\
& \times \prod_{t=1}^{T}\prod_{\ell=1}^{L(t)-1} \exp\{-0.5(Y_{Dt\ell} - \mu_{t\ell})^2 S_{Dt\ell}^{-2}\} I[\text{DAS observed year } t \text{ and state } \ell] \\
& \times \prod_{t=1}^{T} \exp\{-0.5(Y_{DtL(t)} - a_t)^2 S_{DtL(t)}^{-2}\} I[\text{DAS observed year } t \text{ and state } L(t)] \\
& \times \prod_{t=1}^{T}\prod_{\ell=1}^{L(t)-1} \sigma_\eta^{-2}\exp\{-0.5(\mu_{tl} - \boldsymbol{z}'_{t\ell}\boldsymbol{\beta})^2\sigma_\eta^{-2}\}
\end{aligned} \tag{9}$$

$$\times \prod_{t=1}^{T} \sigma_\eta^{-2} \exp\{-0.5(a_t - \boldsymbol{z}'_{tL(t)}\boldsymbol{\beta})^2 \sigma_\eta^{-2}\}$$
$$\times \exp(-0.5\boldsymbol{\beta}'\boldsymbol{\beta}10^{-6}) \exp(-0.5\boldsymbol{b}'\boldsymbol{b}10^{-6}) (\sigma_\eta^2)^{-(0.001+1)} \exp(-\sigma_\eta^{-2} 0.001)$$
$$\times \prod_{k=O,A} \prod_{m=1}^{M(k)} (\sigma_{km}^2)^{-(0.001+1)} \exp(-0.001\sigma_{km}^{-2}) \times \prod_{t=1}^{T} g(\theta_t^R),$$

where

$$\boldsymbol{r}_{ktL(t)} = \boldsymbol{Y}_{ktL(t)} - \boldsymbol{b}_k - \boldsymbol{1}_{M(k)} a_t$$

$a_t = w_{tL(t)}^{-1} \left(- \sum_{\ell=1}^{L(t)-1} w_{t\ell} \mu_{t\ell} + \theta_t^R \right)$, $S(\ell, t)$ is the set of k in $\{O, A\}$ such that survey k is conducted for state ℓ in year t, and $g(\theta_t^R)$ is the prior for θ_t^R. We sample from (9) using the full conditional distributions given in the "Appendix". We monitor convergence using trace plots of the generated samples. We discard the first 5,000 iterations and thin the generated samples by an interval of three to reduce the autocorrelation in the simulated parameters. The jumping probability for the Metropolis-Hastings step in the constrained state model is 30.5 %. A Geweke test of stationarity on the 1,000 iterations used for inference on the variance components and the regression parameters shows that the generated sequences of parameters is stationary.

4.2 Comparison of constrained and unconstrained state models

As a way to compare the constrained and unconstrained state models, we construct forecasts and estimates for final yields using the indications that would be available to construct a forecast or estimate in practice. For concreteness, let t_f denote the year of interest, and let m_f be the month in which the forecast or prediction will be obtained. Let $\boldsymbol{Y}_{t_f,m_f}$ be the set of survey indications obtained through the month m_f for year t_f. For example, if t_f is 2006 and m_f is September, then $\boldsymbol{Y}_{t_f,m_f}$ consists of the survey indications obtained through September of 2006. Inference for $\mu_{t_f\ell}$ is based on the conditional distribution of $\mu_{t_f\ell}$ given $\boldsymbol{Y}_{t_f,m_f}$. We obtained forecasts and estimates for t_f = 2004–2010 and m_f = August–December.

Figure 2 shows the differences between the state forecasts based on the constrained and unconstrained models for the years 2004–2010. The different symbols distinguish the different years. The states are listed on the horizontal axes of the plots in decreasing order by the average harvested acres. For each month, the differences between the state forecasts based on the two models decrease as the harvested acres decrease. The largest August differences are for Iowa and Illinois in 2007, a year when the forecast for the region based on the unconstrained state model is closer to the December estimate than the forecast for the region based on the constrained state model. The largest November differences are for Iowa and Illinois for 2010. Given that a standard error for a November forecast is approximately 3.0 bu/acre, a difference of −6 bu/acre is substantial. The forecasts for the regional yield based on the constrained state model

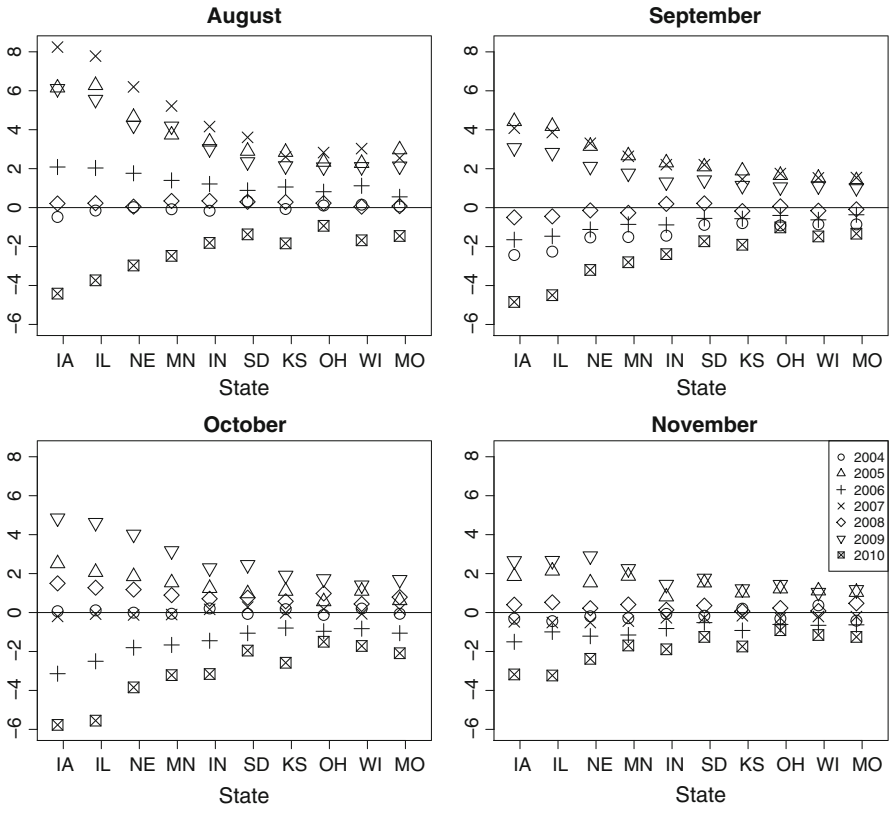

Fig. 2 Differences between state forecasts based on constrained and unconstrained state models. States are listed in decreasing order by average acres harvested for corn. *Shapes* correspond to the years (2004–2010)

for 2010 are closer to the final December estimate than the predictions for the region based on the unconstrained state model.

The state-level differences between the constrained and unconstrained state models for December are not shown because they are small relative to the differences for the other months. The medians of the absolute values of the differences between the state-level December estimates based on the unconstrained state model and the December estimates based on the constrained state model are 0.031, 0.125, 0.118, 0.149, 0.267, 0.285 and 0.043 for 2004, 2005, 2006, 2007, 2008, 2009 and 2010. The differences between the two sets of model-based estimates for December are relatively small because the posterior mode of σ_η^2 is large relative to the $S_{Dt\ell}^2$; the posterior mode of σ_η^2 in the unconstrained model is 190, and a 95 % credible interval for σ_η^2 is [160, 250]. The ratios of the posterior mode of σ_η^2 to $S_{Dt\ell}^2$ range from 7.02 (Kansas, 2003) to 422 (Minnesota, 2000). The median ratio is 113 (Iowa, 2005), and the 25 and 75 percentiles of the ratios are 55 (Wisconsin, 2001) and 164 (Minnesota, 2010), respectively. For Iowa, the state with the largest average harvested acres, the smallest ratio is 39 and the second-smallest ratio is 112. For a ratio of 10, the weight assigned to the DAS indication in the estimated yield is approximately 0.90, and for a ratio of 40, the weight

Table 1 Averages of posterior standard deviations for the unconstrained state model (USM) and the constrained state model (CSM)

State	August		September		October		November		December	
	CSM	USM	CSM	USM	CSM	USM	CSM	USM	CSM	USM
IA	5.656	6.785	4.455	5.440	3.503	4.287	2.107	2.503	0.867	0.999
IL	5.801	6.841	4.594	5.408	3.627	4.295	2.112	2.495	0.822	0.925
NE	6.329	6.806	5.058	5.484	4.007	4.317	2.534	2.785	1.246	1.435
MN	6.516	6.887	5.199	5.515	4.071	4.294	2.557	2.732	1.146	1.226
IN	6.559	6.781	5.239	5.424	4.098	4.237	2.503	2.567	0.847	0.868
SD	6.942	7.028	5.542	5.651	4.291	4.407	2.855	2.911	1.334	1.365
KS	6.910	7.043	5.561	5.686	4.504	4.560	3.086	3.097	2.077	2.170
OH	6.798	6.806	5.426	5.442	4.253	4.253	2.622	2.694	1.158	1.180
WI	6.947	7.068	5.556	5.576	4.427	4.442	3.028	3.040	1.626	1.670
MO	6.866	6.868	5.540	5.560	4.443	4.517	3.002	3.070	1.477	1.498

Averages are across the years 2004–2010 for each combination of state and month

is approximately 0.978. Because most of the ratios of the posterior mode of σ_η^2 to $S_{Dt\ell}^2$ are large, the December estimates based on both models are very close to the DAS estimates. (The approximation ignores the OYS and AYS variances. See Kass and Steffey (1989) for a discussion of such approximations.)

The averages of the posterior standard deviations based on the two models are given in Table 1, where the averages are across years (2004–2010) for a given state and month. The columns are labeled "CSM" and "USM" for the constrained and unconstrained state models, respectively. As we expect, the average posterior standard deviations based on the constrained state model are smaller than the average posterior standard deviations based on the unconstrained state model because the prior for the regional yield used in the constrained state model treats an estimate of the regional yield as a fixed parameter. For each month, the ratios of the standard deviations based on the constrained state model to the averages of the posterior standard deviations based on the unconstrained state model approach one as the average harvested acres decrease because imposing the constraint leads to greater adjustments to yields in states with larger harvested acres. It is not surprising that the averages of the standard deviations decrease as the months approach December because the survey indications become more reliable as the corn grows and is eventually harvested.

Because the quality of the regional forecast is a priority for NASS, we compare the forecasts for the region based on the constrained and unconstrained state models. Table 2 contains empirical root mean squared errors and maximum absolute differences at the regional level for the constrained and unconstrained state models. The empirical mean squared error is defined as the average squared difference between the forecasts and the final estimates published at the end of the year. The maximum absolute difference is the maximum of the absolute differences between the forecasts and the final estimates. The results in Table 2 for the constrained state model differ from Wang et al. (2011) because we incorporate the year 2010, we include the three

Table 2 Empirical root means squared errors and average absolute differences for the unconstrained (USM) and constrained (CSM) state models

Month	CSM		USM		MSE(USM)/MSE(CSM)
	RMSE	Max |Diff|	RMSE	Max |Diff|	
August	7.04	12.17	8.09	15.00	1.32
September	6.88	11.27	7.38	14.47	1.15
October	2.95	4.66	4.43	7.99	2.26
November	1.49	3.32	2.46	5.51	2.73

The empirical root mean squared error is the square root of the average of the squared differences between the model-based forecasts of the yields and final published yields at the regional level. The maximum absolute difference is the largest absolute difference between the model-based forecasts of the yields and the final published yields. Averages and maxima are across the years 2004–2010. Each entry of the last column is a ratio of the MSE for the unconstrained state model to the MSE for the constrained state model

states that were added to the speculative region in 2004, and we include the August rating as a covariate. The empirical root mean squared errors and maximum absolute differences for the region are smaller for the constrained state model than for the unconstrained state model. The ratios of the empirical mean squared errors for the unconstrained model to the empirical mean squared errors for the constrained model are larger for October and November than for August and September. (See Table 2.)

4.3 Comparisons of model-based and published forecasts

To compare the forecasts based on the constrained state model to the published forecasts, we repeat the experiment described at the beginning of Sect. 4.2, where we mimic the realistic situation by deleting data for certain years and months and constructing the forecast as if the data were not yet observed. We plot the differences between the two sets of state forecasts (model-based and published) and final published yields in Fig. 3. The differences between the model-based forecasts and the final published yields are in blue. The differences between the published forecasts and the final published yields are in red. The states are listed in decreasing order by average harvested acres, and shapes correspond to the different years (2004–2010). We call the difference between a forecast and a final published yield a "forecast error." The magnitudes of the forecast errors tend to decrease as the months progress, which is expected because the survey indications become more reliable as more of the corn is harvested.

The variation in the forecast errors across years for a given state is greater than the variation across states for a given year. For example, the forecast errors for August of 2010 based on both procedures (constrained state model and ASB) are positive for all states except for Wisconsin, where the error in the ASB forecast -3 bu/acre. Also, of the 40 forecast errors for 2009, 29 of the model-based forecasts are negative, and 30 of the ASB forecasts are negative.

The differences between the model-based and published forecasts for the same year, month and state are usually smaller in absolute value than the difference between either forecast and the final published yield. For example, $|\hat{\mu}_{tlAug,Model} - \hat{\mu}_{tlAug,ASB}| <$

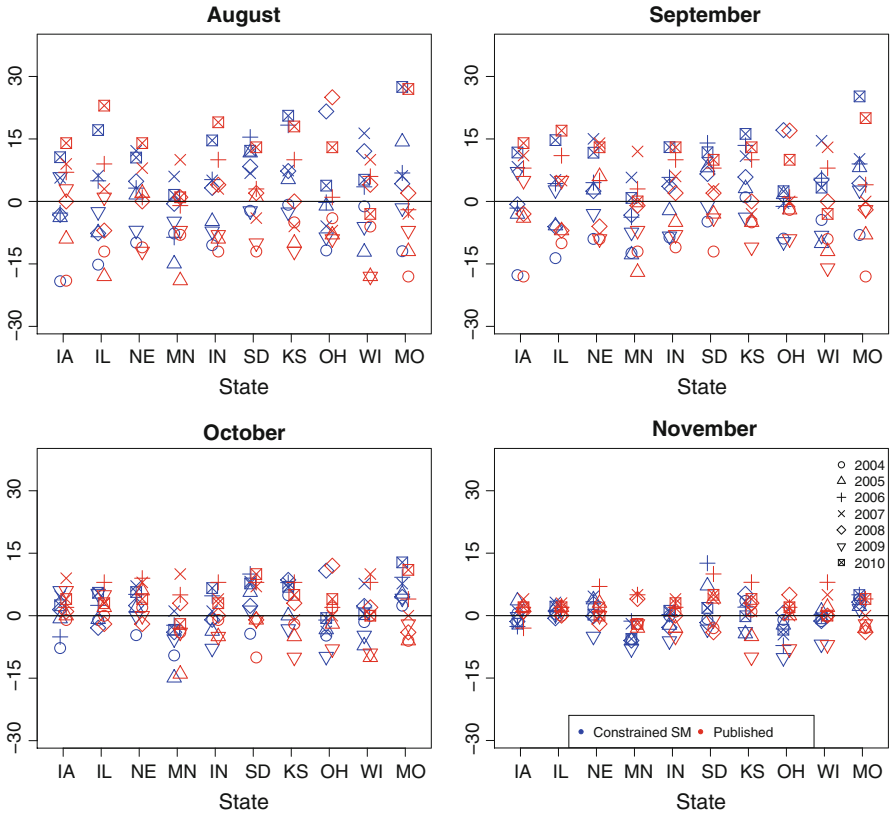

Fig. 3 Forecast errors: differences between forecasts and final published estimates of state-level yields. States are listed in decreasing order by average acres harvested for corn. *Shapes* correspond to the years (2004–2010). For each state, the forecast errors based on the constrained state model are in *blue*, and the errors in the published forecasts are in *red* (Color figure online)

$\min\{|\hat{\mu}_{tlAug,Model} - \hat{\mu}_{tlDec,ASB}|, |\hat{\mu}_{tlAug,ASB} - \hat{\mu}_{tlDec,ASB}|\}$ for 59 out of the 70 state×year combinations in Fig. 3, where $\hat{\mu}_{tlm,Model}$ and $\hat{\mu}_{tlm,ASB}$ denote, respectively, the model-based and published forecasts of the yield for state ℓ, month m, and year t.

4.4 Comparisons of model-based and published final estimates

Out of the 162 combinations of years and states, the published estimate is contained in the 95 % credible interval for the final yield for 142 year×state combinations. For the cases where the published estimate is contained in the 95 % credible interval, the model-based estimates and the published estimates are both heavily influenced by the DAS indications. As discussed in Sect. 4.2, an approximation for the weight assigned to the DAS indication is a function of the ratio of the posterior mode of σ_η^2 to $S_{Dt\ell}^2$. Because all but two of the ratios are larger than 10 and most of the ratios

are larger than 100, the approximate weight assigned to the DAS indication in the final estimate is usually larger than 0.9. Of the 20 instances where the published estimate is outside of the 95 % credible interval, 7 are for the years 1993, 1994, or 1995. The relatively large differences between the model-based and published final estimates for 1993, 1994, and 1995 are expected because the model-based estimates and the published estimates are based on different data. The only data from 1993 to 1995 used to construct the model-based estimates are the OYS indications, and the estimated biases of the OYS indications are largely determined by the average differences between OYS and DAS indications for 1996 through 2010. The published estimates, in contrast, are based partly on DAS indications from before 1996 that we can not access because the NASS data processing environment changed in 1996. Of the remaining 13 state×year combinations where the final published estimates are not contained in the 95 % credible intervals, 8 are from Ohio. For these 8 cases, the absolute difference between the published December estimate and the published November estimate is always smaller than the absolute difference between the model-based December estimate and the published November estimate. The model-based December estimate for Ohio is always closer to the DAS indication than the published estimate is to the DAS indication. As discussed above, the posterior means are close to the DAS indications because the posterior mode of σ_η^2 is 110 times the median DAS sampling standard error.

The only instance where the DAS indication is not contained in the 95 % credible interval is the year 2003 for Kansas. The reason is that the sampling variance for the DAS indication for Kansas for 2003 is the largest DAS sampling variance in the data set. Because of the relatively large sampling standard deviation for the DAS indication, the posterior mean of $\mu_{t\ell}$ for Kansas in 2003 is closer to the forecasts based on the October and November survey estimators than to the DAS indication.

5 Concluding remarks

This paper develops a model-based approach for combining multiple sources of survey information and auxiliary data to forecast and estimate state-level corn yields. The survey information is from two farmer interview surveys (AYS and DAS) and a field measurement survey (OYS) that are conducted monthly in August–December. We focus on the states in the speculative region for corn, which consists of ten states that account for approximately 85 % of corn production. The auxiliary variables are average July precipitation, average July temperature, the percent of corn planted by week 20 of the crop year, and the percent of surveyed individuals who rate the conditions of the corn crop as good or excellent in week 30 of the crop year. The selected variables have been used previously in studies of corn yields. Adding complexity to the model through additional covariates, random coefficients, and spatial dependence does not significantly improve the fit of the model according to the goodness of fit measures of D, DIC, and CV.

By incorporating a restriction on a weighted sum of state-level yields, the forecasts for the speculative region based on the constrained state model proposed in this paper are equal to the forecasts for the speculative region based on the model of Wang et

al. (2011). As mentioned in Sect. 3, the covariates used for our implementation of the "regional model" differ from the covariates used in Wang et al. (2011) in that we include the percent rated good or excellent by week 30. The quality of the forecasts based on the constrained state model is judged superior to the quality of the forecasts based on the unconstrained model in the sense that average squared deviations from the final yields based on the constrained state model are smaller than those based on the unconstrained state model. This is expected because the constraint incorporates useful information. It is also *necessary* to incorporate this constraint to *mimic* the ASB process. Thus, it is not an ad hoc constraint, but it is necessary. The effect of the constraint is greatest for states with largest harvested acres in corn. (See Table 2; Fig. 3).

The differences between the model-based and published forecasts are usually smaller in magnitude than the difference between either forecast and a final published yield (see Fig. 3). The final published estimates are contained in 95 % credible intervals for 88 % of the states and years used for this study. An important advantage of the model-based approach relative to the ASB procedure is the objectivity associated with the use of a formal statistical model. Unlike the subjective ASB method, the statistical procedure can be documented and provides a measure of uncertainty through the posterior mean square error.

The unconstrained state model proposed in this paper implies a distribution for the regional yield. The assumptions for the regional yield and the regional survey estimators implied by the unconstrained state model differ from the assumptions of the regional model. More specifically, under the state model, the regional yield is $\mu_t = \bar{z}'_{w,t}\boldsymbol{\beta} + \bar{\eta}_{w,t}$, where $(\bar{z}'_{w,t}\boldsymbol{\beta}, \bar{\eta}_{w,t}) = \sum_{\ell=1}^{L(t)} w_{t\ell}(z_{t\ell}, \eta_{t\ell})$. Under the unconstrained state model, $V\{\bar{\eta}_{w,t} \mid \sigma_\eta^2\} = \sum_{\ell=1}^{L(t)} w_{t\ell}^2 \sigma_\eta^2$, which is not a constant across years. In contrast, one of the assumptions of the regional model is that the variance of the parameter representing the true regional yield is constant across years. One way to reconcile the regional and state models is to let $V\{\eta_{t\ell} \mid \sigma_\eta^2\} = h_{t\ell}^{-1}\sigma_\eta^2$, where $h_{t\ell}$ is the acres harvested for year t and state ℓ. Similarly, by letting $V\{\delta_{tk\ell m} \mid \sigma_{km}^2\} = \sigma_m^2 h_{t\ell}^{-1}$ one can avoid a contradiction between the nonsampling error variances for the state and regional models.

In the models proposed here, the biases and the variances of the survey indications are assumed constant across years and across states. So, one may ask, do the biases and variances vary systematically across states, years, or weather conditions? Keller et al. (2004) suggest "using the subset of historical data corresponding to years similar to the current year to estimate biases and the covariance structure." One specific question is whether or not the biases, variances, or regression coefficients change when the three new states (Kansas, Missouri, and South Dakota) are added to the speculative region. It is conjectured in a NASS manual that the bias of the AYS indication is larger in magnitude during droughts. Further examination of variation in the biases as a function of covariates or across states is an area for future work.

Acknowledgments The authors thank Jay Wang for allowing us to use the R code which was developed in the collaborative project of the National Institute for Statistical Sciences (NISS) and the National Agricultural Statistics Service (NASS) for predicting corn yields.

Appendix: full conditionals for Metropolis-Hastings

Notation used in this section: $\mathbf{1}_d$ is a d—dimensional column vector of ones, $\boldsymbol{I}_d$ is a d—dimensional identity matrix, $\boldsymbol{\Omega}$ denotes the vector of parameters excluding the parameter being generated, and $\boldsymbol{Z}$ is the matrix with rows $z'_{t\ell}$ listed in the order with states grouped together.

Unconstrained state model

Looking at the joint posterior density in (5), we can obtain the conditional posterior densities (CPD) to run the Metropolis-Hastings sampler.

– For state×year combinations where data from all surveys are available, the conditional posterior distribution (CPD) of $\mu_{t\ell} \mid \boldsymbol{\Omega}$ is

$$\mu_{t\ell} \mid \boldsymbol{\Omega} \sim \mathrm{N}(\Delta_{1,t\ell}^{-1}\Delta_{2,t\ell},\, \Delta_{1,t\ell}^{-1}),$$

where

$$\begin{aligned}\Delta_{1,t\ell} &= [\mathbf{1}_5'(\boldsymbol{\Sigma}_O + \boldsymbol{S}_{t\ell O})^{-1}\mathbf{1}_5 \\ &\quad + \mathbf{1}_4'(\boldsymbol{\Sigma}_A + \boldsymbol{S}_{t\ell A})^{-1}\mathbf{1}_4 + 1/S_{Dt\ell}^2 + 1/\sigma_\eta^2], \qquad (10)\\ \Delta_{2,t\ell} &= \mathbf{1}_5'(\boldsymbol{\Sigma}_O + \boldsymbol{S}_{t\ell O})^{-1}(\boldsymbol{Y}_{t\ell O} - \boldsymbol{b}_O) \\ &\quad + \mathbf{1}_4'(\boldsymbol{\Sigma}_A + \boldsymbol{S}_{t\ell A})^{-1}(\boldsymbol{Y}_{t\ell A} - \boldsymbol{b}_A) + S_{Dt\ell}^{-2}Y_{Dt\ell} + z'_{t\ell}\boldsymbol{\beta}/\sigma_\eta^2. \qquad (11)\end{aligned}$$

For combinations of states and years where at least one survey indication is not available, the terms in $\Delta_{1,t\ell}$ and $\Delta_{2,t\ell}$ associated with the missing survey indications are omitted.

– The CPD of $\boldsymbol{b}_k \mid \boldsymbol{\Omega}$ is $\boldsymbol{b}_k \mid \boldsymbol{\Omega} \sim \mathrm{N}(\boldsymbol{\Delta}_{1,k}^{-1}\boldsymbol{\Delta}_{2,k},\, \boldsymbol{\Delta}_{1,k}^{-1})$, where

$$\begin{aligned}\boldsymbol{\Delta}_{1,k} &= \sum_{\ell=1}^{L}\sum_{t=s(\ell,k)}^{T}(\boldsymbol{\Sigma}_k + \boldsymbol{S}_{t\ell k})^{-1} + 10^{-6}\boldsymbol{I}_{M(k)},\\ \boldsymbol{\Delta}_{2,k} &= \sum_{\ell=1}^{L}\sum_{t=s(\ell,k)}^{T}(\boldsymbol{\Sigma}_k + \boldsymbol{S}_{t\ell k})^{-1}(\boldsymbol{Y}_{t\ell k} - \mathbf{1}_{M(k)}\mu_{t\ell}).\end{aligned}$$

– Letting p denote the number of covariates including the intercept, the CPD of $\boldsymbol{\beta} \mid \boldsymbol{\Omega}$ is $\boldsymbol{\beta} \mid \boldsymbol{\Omega} \sim \mathrm{N}(\boldsymbol{\Delta}_1^{-1}\boldsymbol{\Delta}_2,\, \boldsymbol{\Delta}_1^{-1})$, where

$$\boldsymbol{\Delta}_1 = \boldsymbol{Z}'\boldsymbol{Z}/\sigma_\eta^2 + 10^{-6}\boldsymbol{I}_p,\ \boldsymbol{\Delta}_2 = \boldsymbol{Z}'\boldsymbol{\mu}/\sigma_\eta^2.$$

– The CPD of $\sigma_\eta^2 \mid \boldsymbol{\Omega}$ is $\sigma_\eta^2 \mid \boldsymbol{\Omega} \sim$ Inverse-gamma(a, b), where

$$a = 0.001 + 0.5 \sum_{t=1}^{T} L(t), b = 0.001 + \sum_{t=1}^{T} \sum_{\ell=1}^{L(t)} 0.5(\boldsymbol{z}'_{t\ell}\boldsymbol{\beta} - \mu_{t\ell})^2.$$

– The CPD of σ_{km}^2 and ρ_k are not proportional to known distributions, so we use Metropolis-Hastings with the vector of transformed variables,

$$\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2)',$$

where

$$\boldsymbol{\gamma}_1 = \left(\log(\sigma^2_{O,1}), \ldots, \log(\sigma^2_{A,4})\right)',$$

and

$$\boldsymbol{\gamma}_2 = \left(\log\left[(1-\rho_O)^{-1}(1+\rho_O)\right], \log\left[(1-\rho_A)^{-1}(1+\rho_A)\right]\right)'.$$

The proposal distribution is a multivariate normal distribution with a fixed covariance matrix and mean equal to the current value.

Constrained state model

For the constrained state model, $\mu_{tL(t)} = w_{tL(t)}^{-1}(\theta_t^R - \sum_{\ell=1}^{L(t)-1} w_{t\ell}\mu_{t\ell})$. We sample from the joint posterior density (9) using Metropolis-Hastings. The full conditional distributions for the biases, regression coefficients, and variance parameters are the same in the unconstrained and constrained state models. To define the distribution used to generate $\mu_{t\ell}$ for $\ell = 1, \ldots L(t) - 1$, we introduce some notation. Let $\boldsymbol{\mu}_t^{(1)} = (\mu_{t1}, \ldots, \mu_{t,L(t)-1})'$,

$$\boldsymbol{V}_{t,1} = \begin{pmatrix} \text{diag}(\Delta_{1,t1}^{-1}, \ldots, \Delta_{1,t(L(t)-1)}^{-1}) & \boldsymbol{\omega}'_{t,1} \\ \boldsymbol{\omega}_{t,1} & \sum_{\ell=1}^{L(t)} w_{t\ell}^2 \Delta_{1,t\ell}^{-1} \end{pmatrix},$$
$$\boldsymbol{\omega}_{t,1} = (w_{t1}\Delta_{1,t1}^{-1}, \ldots, w_{t(L(t)-1)}\Delta_{1,t(L(t)-1)}^{-1})',$$

and

$$\boldsymbol{m}_{t,1} = \left(\Delta_{1,t1}^{-1}\Delta_{2,t1}, \ldots, \Delta_{1,t(L(t)-1)}^{-1}\Delta_{2,t(L(t)-1)}, \sum_{\ell=1}^{L(t)} w_{t\ell}\Delta_{1,t\ell}^{-1}\Delta_{2,t\ell} - \theta_t\right)',$$

where $\Delta_{1,t\ell}$ and $\Delta_{2,t\ell}$ are defined in the specification of the full conditional distribution for $\mu_{t\ell}$ in the unconstrained state model in Eqs. (10) and (11). Partition $\boldsymbol{V}_{t,1}$ and $\boldsymbol{m}_{t,1}$ above as,

$$\boldsymbol{V}_{t,1} = \begin{pmatrix} \boldsymbol{\Sigma}_t^{(11)} & (\boldsymbol{\sigma}_t^{(12)})' \\ \boldsymbol{\sigma}_t^{(12)} & \sigma_t^{(22)} \end{pmatrix},$$

and $\boldsymbol{m}_{t,1} = (\boldsymbol{m}_{t,1}^{(1)}, m_t^{(2)})'$, respectively, where $\boldsymbol{m}_t^{(1)}$ is of dimension $L(t) - 1$, $\boldsymbol{\Sigma}_t^{(11)}$ is of dimension $(L(t) - 1) \times (L(t) - 1)$, and $\boldsymbol{\sigma}_t^{(12)}$ is of dimension $(L(t) - 1) \times 1$.

– The CPD of $\boldsymbol{\mu}_t^{(1)} \mid \boldsymbol{\Omega}$ is $\boldsymbol{\mu}_t^{(1)} \mid \boldsymbol{\Omega} \sim \mathrm{N}(\boldsymbol{\delta}_t, \boldsymbol{\Delta}_t)$, where

$$\boldsymbol{\delta}_t = \boldsymbol{m}_t^{(1)} - \boldsymbol{\sigma}_t^{(12)}/\sigma_t^{(22)} m_t^{(2)}, \quad \boldsymbol{\Delta}_t = \boldsymbol{\Sigma}_t^{(11)} - \boldsymbol{\sigma}_t^{(12)}/\sigma_t^{(22)} (\boldsymbol{\sigma}_t^{(12)})'.$$

References

Banerjee S, Carlin B, Gelfand A (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall CRC, London

Battese GE, Harter RM, Fuller WA (1988) An error-components model for prediction of county crop areas using survey and satellite data. J Am Stat Assoc 83:28–36

Berg E, Barboza W, Nandram B (2011) A constrained Bayesian hierarchical model for forecasting state-level corn yield. Unpublished NASS technical report

Datta G, Ghosh M, Steorts R, Maples J (2011) Bayesian benchmarking with applications to small area estimation. TEST 20:574–588

Fay R, Herriot R (1979) Estimates of income for small places: an application of James–Stein procedures to census data. J Am Stat Assoc 74:341–353

Gelfand AE, Ghosh SK (1998) Model choice: a minimum posterior predictive loss approach. Biometrika 85:1–11

Irwin S, Good D, Tannura M (2008) Weather, technology, and corn and soybean yields in the U.S. corn belt. Forming expectations about 2008 U.S. corn and soybean yields application of crop weather models that incorporate planting progress. Marketing and outlook briefs. Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign

Kass RE, Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J Am Stat Assoc 84(407):717–726

Kantanantha N, Serban N, Griffin P (2010) Yield and price forecasting for stochastic crop decision planning. J Agric Biol Environ Stat 15(3):362–380

Keller T, Olkin I (2002) Combining correlated unbiased estimators of the mean of a normal distribution. Technical Report No. 2002–5. National Agricultural Statistics Service

Keller T, Wigton W, Garber S, McEwen B, Rumberg D, Schleusener M, DeWalt D, Ellison H, Onig L, Jantzi D, Thessen G, Guss P, Parks B (2004) Research on composite indications of crop yield. Technical Report. National Agricultural Statistics Service

Manzi G, Spiegelhalter DJ, Turner RM, Flowers F, Thompson SG (2011) Modelling bias in combining small area prevalence from multiple surveys. J R Stat Soc A 174:31–50

Nandram B, Sayit H (2011) A Bayesian analysis of small area probabilities under a constraint. Surv Methodol 37:137–152

Raghunathan TE, Xie D, Schenker N, Parsons VL, Davis WW, Dodd KW, Feuer EJ (2007) Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. J Am Stat Assoc 102:474–486

Roper A, Wagstaff KL (2007) A Support-vector based machine Learning Approach to prediction of crop yield from multispectral satellite images. J Mach Learn Res 1:1–48

Rosenzweig C, Tubiello FN, Goldberg R, Mills E, Bloomfield J (2002) Increased crop damage in the US from excess precipitation uncer climate change. Glob Environ Change 12:197–202

Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc B 64:583–630

Tannura MA, Irwin SH, Good DL (2008) Weather, technology, and corn and soybean yields in the U.S. corn belt. Marketing and outlook research report 2008–01. Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign

Vado L, Goodwin B (2010) Analyzing the effects of weather and biotechnology adoption on corn yields and crop insurance performance in the U.S. corn belt. Selected Paper prepared for presentation at the Agricultural and Applied Economics Association's 2010 AAEA, CAES and WAEA Joint Annual Meeting, Denver, CO., July 25–27, 2010

Wang JC, Holan SH, Nandram B, Barboza W, Toto C, Anderson EA (2010) Internal NASS Report

Wang JC, Holan SH, Nandram B, Barboza W, Toto C, Anderson EA (2011) Bayesian approach to estimating agricultural yield based on multiple repeated surveys. J Agric Biol Environ Stat 1085–7117:1–23. doi:10.1007/s13253-011-0067-5

Author Biographies

Balgobin Nandram is a Professor of statistics in the Department of Mathematical Science at Worcester Polytechnic Institute.

Emily Berg is a Research Assistant Professor in the Department of Statistics and the Center for Survey Statistics and Methodology at Iowa State University.

Wendy Barboza is a Senior Mathematical Statistician and Chief of the Statistical Methodology Research Branch at the National Agricultural Statistics Service.