

Yield and Price Forecasting for Stochastic Crop Decision Planning

Nantachai Kantanantha¹ Nicoleta Serban² and Paul Griffin²

The primary objective of this paper is to develop yield and price forecasting models employed in informed crop decision planning - a key aspect of effective farm management. For *yearly yield prediction*, we introduce a weather-based regression model with time-dependent varying coefficients. In order to allow for within-year climate variations, we predict yearly crop yield using weekly temperature and rainfall summaries resulting in a large number of correlated predictors. To overcome this difficulty, we reduce the space of predictors to a small number of uncorrelated predictors using Functional Principal Component Analysis (FPCA). For *detailed price forecasting*, we develop a futures-based model for long-range cash price prediction. In this model, the cash price is predicted as a sum of the nearby settlement futures price and the predicted commodity basis. We predict the one-year commodity basis as a mixture of historical basis data using a functional model-based approach. In both forecasting models, we estimate approximate prediction confidence intervals that are further integrated in a decision planning model. We applied our methods to corn yield and price forecasting for Hancock County in Illinois. Our forecasting results are more accurate in comparison to predictions based on existing methods. The methods introduced in this paper generally apply to other locations in the US and other crop types.

Key words and phrases: crop decision planning, crop price forecasting, crop yield forecasting, functional model-based clustering, functional principal component analysis, varying-coefficient model.

¹Nantachai Kantanantha is Lecturer in Department of Industrial Engineering, Faculty of Engineering, Kasetsart University.

²Nicoleta Serban is Assistant Professor and Paul Griffin is Professor in the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology.

1 Introduction

Agriculture and related businesses are large industries in the U.S. In 2007, they had a value added of approximately \$161.6B, a 64% increase from 2000 (Bureau of Economic Analysis) and employed about 2.1 million workers (Bureau of Labor Statistics). In addition, the agribusiness has become very complex in recent years, and hence the importance of agricultural planning has increased. Crop producers often base their decisions for crop production and selling on yield and price forecasts. *The objective of this paper is to develop accurate yield and price forecasting models along with prediction confidence bands* further integrated in a decision planning model introduced and analyzed in Kantanantha et al. (2008).

Yield Forecasting: Background and Motivation. Timely and accurate crop yield forecasting is essential for crop production, marketing, storage, and transportation decisions and also helps managing the risk associated with these activities (Lee, 1999; Potgieter et al., 2005). The most well-known and widely used crop information source is the monthly USDA Crop Production reports (Krog, 1988). These reports, prepared by the National Agricultural Statistics Service (NASS), provide statistics and related information on crop production in the U.S. Even though these reports supply broadly utilized yield forecasts, they generate only the mean estimate for each state (Lee, 1999). However, there are significant variations in crop yield by location due to varying environmental conditions. In many previous studies, yield forecasting models incorporate a series of weather predictors (Hoogenboom, 2002; Kandiannan et al., 2002), more specifically, temperature (Peng et al., 2004, Wheeler et al., 2000, Batts et al., 1997) and rainfall (Mkhabela et al., 2005). In line with the existing work, *we develop a crop-weather forecasting model using a regression approach where the weather factors are rainfall and temperature.* Additionally, we account for economic growth by including GDP as a regressor. The common weather-based approach to yield forecast is linear regression with constant coefficients over time. Although linear regression is easy to estimate and interpret, it may be restrictive and with limited prediction power since it does not account for the year-to-year dependence in the yield

variable. To address this limitation, *we propose using a varying-coefficient model, which allows for time-dependent association of weather and economic growth to yield.* We assume that the varying-coefficients are unknown functions of time which are estimated using penalized splines (Ruppert et al., 2003, Chapter 12.4).

For the weather based model, the number of predictors is very large. The planting season consists of about 22 weeks, and for each week, we have temperature and rainfall predictors - a total of 44 predictors. For accurate prediction, we need to reduce the space of predictors. One common approach to reducing the set of predictors is variable selection. However, because the predictors are correlated, variable selection will not provide a valid set of highly significant predictors - the estimated regression coefficients will be unstable due to their joint effect (non-identifiability). To overcome this difficulty, *we reduce the set of correlated predictors to a smaller set of uncorrelated predictors using Functional Principal Component Analysis (FPCA).* Sood et al., (2008) and Cai and Hall (2008) also propose using Functional Principal Component Analysis (FPCA) to transform functional covariates into a finite set of uncorrelated covariates. We applied FPCA to both temperature and rainfall data using a similar estimation approach as discussed in Ramsey and Silverman (1997). Ferraty and Vieu (2006) and Cai and Hall (2008) introduce general estimation frameworks for prediction under functional predictors and scalar response. An important difference between their approach and our model structure is that the response variable varies with time.

The final model is a varying-coefficient model where the temperature and rainfall predictors are replaced by the scores of the functional principal components which explain most of the variability in temperature and rainfall. Our yield forecasting model is novel in that it allows borrowing predictive power from one-year to another and it reduces to a parsimonious model by using a small set of predictors which explain most of the variability in the functional weather predictors. The forecasting results based on this model show improved prediction in comparison to linear regression, the most common method for yield forecasting. Moreover, our model is easy to interpret and implement, and therefore, easy to be incorporated in the existing software for

agribusiness decision.

Price Forecasting: Background and Motivation. Prediction of future crop selling prices is another important aspect in decision planning. Accurate price predictions will help in planning what crops to be planted and when to sell them to optimize the overall profit. The U.S. Department of Agriculture (USDA) publishes reports on crop supply, demand, yields, prices, and other related information. Agricultural prices are provided by the World Agricultural Outlook Board (WAOB) in the World Agricultural Supply and Demand Estimates (WASDE) report. However, the WASDE report is distributed once a month and the prices are national averages. Consequently, a crop price forecasting model for predicting the upcoming prices in any specific location and at a finer aggregation level (e.g. weekly) will help local farmers to optimize their crop selling strategy. A number of models have been developed to forecast the cash prices. Kenyon and Lucas (1998) study the relationship between soybean season average prices and soybean ending stocks - the difference between supply and demand. They propose a simple price forecasting model using price historical data and the ending stocks based on linear regression. Many researchers studied the role of futures contract prices in agricultural price forecasting (Working, 1942; Tomek and Gray, 1970, Kenyon et al., 1993). Futures price is often used as an indicator of the expected cash price (Hoffman, 2005). Eales et al. (1990) examine the difference between futures prices and the average cash prices surveyed from farmers and grain merchandisers in Illinois. In most cases, futures price and cash price are not significantly different.

Because the futures crop price is an indicator of the cash price behavior, in our forecasting model, *we predict the cash price at a specific location by summing the futures price and the predicted local commodity basis*. In the agribusiness literature, *commodity basis* is defined as the difference between the local market cash price and the price of a futures contract for a specific time period. We estimate one-year commodity basis pattern using a functional model-based approach introduced by James and Sugar (2003). Under this approach, we model the underlying commodity basis as a mixture of Gaussian processes. Therefore, using this approach, we predict the commodity basis pattern as a mixture of the historical commodity basis where

the mixture weights are estimated rather than fixed. Our price forecasting model is novel in that it predicts the local crop price using a global component (futures price) and a local predicted component (commodity basis). Moreover, our method allows estimating prediction confidence bands for several months ahead - throughout the planting season and after. This is important in crop decision planning as a farmer will need to decide which crop to grow and when to sell at the beginning of the planting season.

The layout of the paper is as follows. In Section 2, we present the data used for yield and price forecasting. We first introduce the methodology for yield forecast in Sections 3.1 and 3.2 and the prediction results of the proposed model in Section 3.4. In the second part of this paper, we discuss the methodology for commodity basis prediction in Sections 4.1 and 4.2 and the results on price forecasts in Section 4.3. Finally, in Section 5 we discuss how these results could be used in crop planning.

2 Data Background

2.1 Data for Yield Forecasting

Yield Data. Historical corn and soybean yield data are acquired from Quick Stats, an agricultural statistics database, provided by the National Agricultural Statistics Service. Yield data are expressed as a number of bushels harvested per acre. Both corn and soybean yield data are from Hancock County in Illinois from 1927 to 2005. We chose Illinois as our primary state in our study since it is the second largest corn producer in the U.S. (National Agricultural Statistics Service, 2005). Hancock county is chosen as a representative county in Illinois. *Our methodology applies to any crop producer across the country.*

Weather Data. In our weather based model, we use the weather data from National Climatic Data Center (NCDC). These data are collected from La Harpe station in Hancock County, Illinois, from 1927 to 2005. The rainfall variable is the total daily rainfall in inches and the temperature variable is the average daily temperature in

degree Fahrenheit. Even though we have daily data available, we aggregate both the temperature and the rainfall at the weekly level since this aggregation will filter out high frequency variation that would likely be non-predictive. Based on the Planting and Harvesting Dates in the US Field Crops report (National Agricultural Statistics Service, 1997), the planting season for corn in Illinois is approximately beginning of May to late September. We propose predicting yield based on the temperature and rainfall data during this period.

Forecasted Weather Data. In real practice, we do not know the weather condition in advance. Therefore, in predicting the yield, we need to use weather forecast throughout the planting season. In the time series forecasting literature, the problem of obtaining weather forecast for months in advance falls under seasonal or long-range prediction. Existing methods for long-range weather forecasting are based on a broad spectrum of information - satellite data, land surface and atmosphere data, historical and real-time weather data, and historical forecast weather data. To implement these methods, we need to acquire weather-specific data which are not commonly available from non-commercial sources. Consequently, we obtain our weather forecasts using a standard time series technique called autoregressive integrated moving average (ARIMA). We forecast one planting-season ahead using data from the previous years and data in the the same year from January to April.

Economic Growth Factor. Another variable used in yield forecasting is annual GDP from 1926 to 2004. We use the nominal GDP acquired from Economic History Services.

2.2 Data for Price Forecasting

We base our price forecast on both futures and cash price data. The **futures price data** are acquired from an agricultural package provided by the Chicago Board of Trade. A futures contract is classified by its delivery month or contract month. However, there are specific delivery months for each crop. Corn futures contracts are delivered only in March, May, July, September, and December. In this research, we select the *nearby settlement price* to represent the futures price. A *nearby contract* is the

futures contract that is closest to expiration. For example, December corn futures is the nearby futures for corn in October. We acquire the **cash price data** from the USDA Springfield regional office. This office provides the average cash prices of corn traded in central Illinois. Both futures and cash prices are collected every business day from 1991 to the first quarter of 2006. We compute the daily commodity basis from nearby futures and cash prices.

3 Yield Forecasting

3.1 Model Formulation

The underlying yield forecasting model is a varying-coefficient model:

$$\mathbb{E}(Y_i|T, R, GDP) = \mu(t_i) + \alpha_1(t_i)T_{1i} + \dots + \alpha_p(t_i)T_{pi} + \alpha_{p+1}(t_i)R_{1i} + \dots + \alpha_{2p}(t_i)R_{pi} + \alpha_{2p+1}(t_i)GDP_{i-1}, \quad (1)$$

where Y_i is the yield in year t_i for $i = 1, \dots, N$ (N is the number of years). The set of predictors consists of weekly weather data (T_{wi} and R_{wi} are temperature and rainfall observations in week w for $w = 1, \dots, p$ of year i) to account for within-year climate variations and one-year lag nominal GDP to account for the economic growth. In model (1), the regression coefficients $\mu(t)$ and $\alpha_1(t), \dots, \alpha_{2p+1}(t)$ are assumed to vary smoothly over time. This model is an extension of the linear regression model which assumes constant regression coefficients. Allowing for time-dependent regression coefficients will improve the prediction of the response by borrowing predictive power across time.

The weather-based model consists of a large number of predictors (22 for temperature, 22 for rainfall when using weekly data and one for GDP). The number of predictors can be even larger when using other weather predictors (e.g. humidity). One common approach to reducing the set of predictors is variable selection. However, because the predictors are correlated, variable selection will not provide a valid set of highly significant predictors - the estimated regression coefficients will be unstable due to their joint effect (non-identifiability). One way to overcome this limitation is

to first apply a dimensionality reduction method to reduce the set of temperature predictors and the set of rainfall predictors to a smaller set of uncorrelated variables without significant loss of information. In this paper, we apply a functional version of PCA (FPCA) since the temperature predictors as well as the rainfall predictors are functionally dependent. More specifically, to highlight the within-year functionality, we may write

$$T_{wi} = T_i(w) \text{ and } R_{wi} = R_i(w).$$

That is, $T_i(w)$ and $R_i(w)$ vary in time (week-to-week). We apply FPCA to temperature and rainfall data separately. The key references for FPCA are Chapter 8 of Ramsay and Silverman (1997), but recently, other methods for estimating FPC's have been introduced. For example, Yao et al. (2005) developed a method that allows for sparse design.

FPCA Applied to Temperature Data. In FPCA, we assume that $T_i(w)$ is a stochastic process with mean $\mathbb{E}(T_i(w)) = m_T(w)$ and covariance spectral decomposition $\mathbb{C}(T_i(w_1), T_i(w_2)) = \mathbb{C}_T(w_1, w_2) = \sum_{j=1}^{\infty} \tau_j \phi_j(w_1) \phi_j(w_2)$ where $\{\phi_j(w)\}_{j=1, \dots, \infty}$ form an orthogonal basis and are called eigenfunctions. Under these assumptions, according to Karhunen-Loève decomposition, the functional $T_i(w)$ will be decomposed according to

$$T_i(w) = m_T(w) + \sum_{j=1}^{\infty} P_{ji} \phi_j(w) \tag{2}$$

where the coefficients P_{ji} are called scores and they are uncorrelated with $\mathbb{E}(P_{ji}) = 0$ and $\mathbb{E}(P_{ji}^2) = \tau_j$. Based on the Karhunen-Loève decomposition of the covariance $\mathbb{C}_T(w_1, w_2)$, $\{\phi_j(w)\}_{j=1, \dots, \infty}$ are the associated eigenfunctions and $\tau_1 \geq \tau_2 \geq \dots$ are the ordered eigenvalues. However, the decomposition in (2) is impractical because it is based on infinite sum, and therefore we use a truncated version of this decomposition. Moreover, only a small number of eigenvalues are commonly significantly non-zero. For the eigenvalues which are approximately zero, the corresponding scores will also be zero in mean because $\mathbb{E}(P_{ji}^2) = \lambda_j$. We may then approximate the within-year

temperature pattern with

$$\tilde{T}_i(w) = m_T(w) + \sum_{j=1}^I P_{ji} \phi_j(w)$$

where I is the number of significantly non-zero eigenvalues. We choose I such that the first I principal components will explain at least 90% of the variability in the temperature data. For temperature data, $I = 6$. *In conclusion, using this approach we will reduce the set of 22 temperature predictors to only six predictors which will be the scores of the six principal components explaining most of the variability in the temperature data.* In model (1), we replace the predictors $T_{wi}, w = 1, \dots, 22$ with the scores $P_{ji}, j = 1, \dots, 6$.

FPCA Applied to Rainfall Data. We denote $\psi_j(w)$ the principal components and denote S_{ji} the scores for the FPCA decomposition of the rainfall functionals

$$\tilde{R}_i(w) = m_R(w) + \sum_{j=1}^J S_{ji} \psi_j(w).$$

For rainfall data, $J = 4$ since only the first four principal components explain more than 90% of the variability in the rainfall data. In model (1), we will replace the predictors $R_{wi}, w = 1, \dots, 22$ with the rainfall scores $S_{ji}, j = 1, \dots, 4$. After reducing the set of predictors using FPCA, the model in (1) becomes

$$\mathbb{E}(Y_i|T, R, GDP) = \mu(t_i) + \alpha_1(t_i)P_{1i} + \dots + \alpha_I(t_i)P_{Ii} + \alpha_{I+1}(t_i)S_{1i} + \dots + \alpha_{I+J}(t_i)S_{Ji} + \alpha_{I+J+1}(t_i)GDP_{i-1}. \quad (3)$$

3.2 Model Estimation

The regression coefficients μ and α_j for $j = 1, \dots, (I + J + 1)$ are assumed to be unknown smooth functions. Using a nonparametric approach, we decompose the regression coefficients using an orthonormal basis of functions. Common basis of functions are Fourier, wavelets and splines. In our approach, we use penalized splines but other basis may be applied also. The spline basis is a (low-rank) radial basis with

degree $p = 2$. Therefore, we assume

$$\mu(t) = \mu_0, \alpha_j(t) = \alpha_{0j} + \alpha_{1j}t + \sum_{r=1}^R u_{jr} |t - \kappa_r|^3, j = 1, \dots, (I + J + 1)$$

where κ_r for $r = 1, \dots, R$ are equally distributed knots over the time domain (1927-2005). The number of knots is not important since we control the smoothness of the regression functions by penalizing the goodness-of-fit. We estimate the coefficients $\mu_0, \alpha_{0j}, \alpha_{1j}$ and u_{jr} for $j = 1, \dots, (I + J + 1)$ by minimizing the penalized least sum of squares:

$$\sum_{i=1}^N \left(Y_i - \mu(t_i) - \sum_{j=1}^I \alpha_j(t_i) P_{ji} - \sum_{j=I+1}^{I+J} \alpha_j(t_i) S_{ji} - \alpha_{I+J+1}(t_i) GDP_{i-1} \right) + \sum_{j=1}^{I+J+1} \lambda_j J(\alpha_j) \quad (4)$$

where $J()$ is a wiggleness penalty (Wahba, 1990) and λ 's are penalty tuning parameters which control the trade-off between goodness-of-fit and smoothness. Similar to the semi-parametric models discussed in Ruppert et al. (2003), the solution to the model in (4) is equivalent to the solution of the mixed-effects model described below. Parameter estimation under the mixed effects model formulation allows automatic evaluation of the smoothing penalty parameters, which is an important aspect in our model estimation since the number of penalty parameters ($I + J + 1$) is large.

Under the equivalent mixed effects model, u_{jr} 's are normally distributed random effects with

$$\Omega^{1/2} u_{jr} \sim N(0, \sigma_{u_j}^2 I_R) \text{ where } \Omega = [|\kappa_r - \kappa_{r'}|^3].$$

Because we use a low-rank penalized spline basis, we need to rescale the random effects by $\Omega^{-1/2}$. Other assumptions in the model are that the predictors have mean zero and constant variance across all years. These assumptions hold since P_{ji} and S_{ji} are *the scores of functional principal components for temperature and rainfall*. GDP used in this model is standardized.

Define the X and Z matrices as

$$\begin{aligned} X &= [1 \ t_i P_{1i} \ \dots \ t_i P_{Ii} \ t_i S_{1i} \ \dots \ t_i S_{Ji} \ t_i GDP_{i-1}]_{i=1, \dots, N}, \\ Z_{P_j} &= \{[|t_i - \kappa_1|^3 \ \dots \ |t_i - \kappa_R|^3] P_{ji}\}_{i=1, \dots, N}, \quad j = 1, \dots, I, \\ Z_{S_j} &= \{[|t_i - \kappa_1|^3 \ \dots \ |t_i - \kappa_R|^3] S_{ji}\}_{i=1, \dots, N}, \quad j = 1, \dots, J, \\ Z_{GDP} &= \{[|t_i - \kappa_1|^3 \ \dots \ |t_i - \kappa_R|^3] GDP_{i-1}\}_{i=1, \dots, N}. \end{aligned}$$

Because the random effects are rescaled by $\Omega^{-1/2}$, it is common practice to multiply the Z matrices with $\Omega^{-1/2}$ to predict independent random effects. Denote

$$\{\tilde{Z}_{P_j} = \Omega^{-1/2} Z_{P_j}, \quad j = 1, \dots, I\}, \{\tilde{Z}_{S_j} = \Omega^{-1/2} Z_{S_j}, \quad j = 1, \dots, J\}, \tilde{Z}_{GDP} = \Omega^{-1/2} Z_{GDP}.$$

Finally, define \tilde{Z} matrix as $\tilde{Z} = [\tilde{Z}_{P_1} \ \dots \ \tilde{Z}_{P_I} \ \tilde{Z}_{S_1} \ \dots \ \tilde{Z}_{S_J} \ \tilde{Z}_{GDP}]$. Under these notations, the model in (3) is equivalent to a linear mixed model

$$Y_i = \alpha X + u \tilde{Z} + \varepsilon_i, \quad i = 1, \dots, N, \quad (5)$$

where α are fixed effects and u are random effects. Assuming that the variance of the errors $\mathbb{V}(\varepsilon_i) = \sigma^2$, the penalty smoothing parameters defined in (4) are $\lambda_j = \frac{\hat{\sigma}_{u_j}^2}{\hat{\sigma}^2}$ for $j = 1, \dots, I + J + 1$.

3.3 Prediction

Under the mixed effects model in (5), we estimate the fixed effects α and predict the random effects u using best linear unbiased prediction (BLUP) derived in Ruppert et al. (2003), subsection 4.3. Denote the BLUP estimators $\hat{\alpha}$ and \hat{u} . Let $t^* = t_{N+1}$ be the prediction year, and $P_{1t^*}, \dots, P_{It^*}$ the temperature scores and $S_{1t^*}, \dots, S_{Jt^*}$ the rainfall scores for the prediction year. We obtain these scores by applying FPCA to observed historical weather data for years t_1, \dots, t_N and predicted weather data for the prediction year T_{N+1} . With this notations, define the matrix $C_{t^*} = [X_{t^*} \ Z_{t^*}]$

where

$$\begin{aligned} X_{t^*} &= [1 \ t^* P_{1t^*} \ \dots \ t^* P_{It^*} \ t^* S_{1t^*} \ \dots \ t^* S_{Jt^*}] \\ \tilde{Z}_{t^*} &= \Omega^{-1/2} [|t_i - \kappa_1|^3 P_{1t^*} \ \dots \ |t_i - \kappa_R|^3 P_{It^*} \ |t_i - \kappa_1|^3 S_{1t^*} \ \dots \ |t_i - \kappa_R|^3 S_{Jt^*}] \end{aligned}$$

The predicted yield for year t^* is derived from the mixed model (5) using the notations above

$$Y_{t^*} = \hat{\alpha} X_{t^*} + \hat{u} \tilde{Z}_{t^*}.$$

We compute an approximate prediction confidence interval for Y_{t^*} with bias correction extending on the derivation by Ruppert et al. (2003), pp.137-140. The $(1 - \alpha)$ prediction interval is

$$\hat{Y}(t^*) \pm z_{(1-\frac{\alpha}{2})} \hat{\sigma}_\varepsilon \sqrt{C_{t^*} \left(C^T C + \frac{\sigma_\varepsilon^2}{\sigma_u^2} D \right)^{-1} C_{t^*}^T + 1}, \quad (6)$$

where $C = [X \ \tilde{Z}]$ and $D = \text{diag}(0, \dots, 0, 1, \dots, 1)$. The number of zeros in D is equal to the number of columns in X and the number of ones in D is equal to the number of columns in \tilde{Z} . This confidence interval is an approximation because it does not take into account the variability in the smoothing parameters λ_j , $j = 1, \dots, I + J + 1$ and the variability in the predicted weather scores $P_{t^*} = (P_{1t^*}, \dots, P_{It^*})$ and $S_{t^*} = (S_{1t^*}, \dots, S_{Jt^*})$ for the prediction year. The variability in the smoothing parameters is commonly negligible under the assumption of a large sample but the variability in the predicted scores depends on the accuracy of the predicted weather data. Assuming $\hat{\alpha}$ and \hat{u} independent and homoscedastic with respect to each other, we can extend the confidence interval in (6) using the variance and covariance derivations in Bohrnstedt and Goldberger (1969). Specifically, we replace $C_{t^*} \left(C^T C + \frac{\sigma_\varepsilon^2}{\sigma_u^2} D \right)^{-1} C_{t^*}^T$ with a variance estimator which takes into account the variability in the predicted temperature and rainfall scores. However, for this, we need to input the expectations of the prediction scores, $\mathbb{E}(P_{t^*})$ and $\mathbb{E}(S_{t^*})$, and the variances of the prediction scores $\mathbb{V}(P_{t^*})$ and $\mathbb{V}(S_{t^*})$, which in turn depend on the expectation and the variance of the predicted weather data. Because of this difficulty, we may use the confidence interval

in (6) as an approximation for prediction upper and lower bounds for the yield.

To evaluate the precision or coverage of the prediction confidence interval described in (6), we performed a simulation study described in the Appendix. We computed the precision as the number of times a 95% prediction interval covers the simulated true value. We simulate 100 datasets and predict ten years for each dataset. For this simulation study, 52 prediction intervals out of 1000 (100 simulations \times 10 years) do not cover the true simulated value when predicting based on observed weather data. Therefore, the coverage is close to the confidence level of the prediction confidence intervals. The precision decreases slightly when predicting based on weather data forecast - 60 prediction intervals out of 1000 do not cover the true simulated value. In the Appendix, we evaluate the mean absolute prediction errors and compare the accuracy of the prediction confidence intervals under both scenarios. Overall, the prediction method performs well when using the forecast instead of observed weather data although the prediction precision decreases.

3.4 Yield Forecasting: Results and Discussion

Comparative Models. The first model is described in equation (3) which is a varying-coefficient regression model with 11 predictors. We refer to this model as *Model 1*. However, some of the estimated regression coefficients in Model 1 have high penalty smoothing parameters ($\lambda_j = \frac{\hat{\sigma}_{u_j}^2}{\hat{\sigma}^2}$ large). This suggests using linear or constant functions to estimate these regression coefficients implying a more parsimonious model, which will be easier to predict and interpret. We therefore use the smoothing parameter as an indicator for identifying predictors with a linear or constant relationship to the yield response. The selected model consists of three nonlinear regression components - standardized lag nominal GDP, first temperature principal component score and third rainfall component score and four constant components - second, third, fourth, and fifth temperature principal component scores. This model is referred as *Model 2*.

The common approach to yield prediction using weather data is linear regression (Sheehy et al., 2006 and Kandiannan et al., 2002). Because of the large number of

temperature and rainfall predictors, we investigate the reduced model in (3) assuming constant regression coefficients ($\alpha_j(t) = a_j$ for $j = 1, \dots, I + J + 1$). However, not all $I + J + 1$ predictors are significant; consequently, we further reduce the set of predictors using the minimum R^2 -adjusted criterion. The resulting model is *Model 3* in this paper.

One other alternative to variable reduction is to use monthly rather than weekly aggregated data, and therefore, reduce from 45 to only 11 predictors - five for temperature and five for rainfall since we consider only the growing season period. The linear regression model using selected aggregated variables is referred as *Model 4*.

Evaluation Results. We evaluate the predictions over a period of 10 years (1996-2005). We investigate the performance of the four models according to two evaluation measures, mean square prediction error (MSPE) and mean absolute percentage prediction error (MAPE) reported in Tables 1 and 2. The first table reports the prediction results when using observed weather data for the prediction year. Although the weather data are not available at the time of yield prediction (beginning of the planting season), this model allows us to evaluate prediction errors under perfect knowledge of weather data and to perform variable selection. The second table presents the mean prediction errors for the four models based on predicted weather data for the prediction year. The least parsimonious model (Model 1) has a high MSPE and MAPE although has the highest R^2 -adjusted. On the other hand, Model 2 which is a parsimonious version of Model 1 has the lowest MSPE and MAPE for both observed and forecasted weather data (highlighted in Tables 1 and 2). We compare the predicted and the observed corn yields for years 1996 to 2005 in Figure 1. Models 3 and 4 consistently over-forecast the yield. The largest differences between the four models is for the period from 1996 to 1999. The prediction accuracy improves for Model 2 throughout these years because this model allows borrowing information across time. The observed corn yield in year 2005 is lower than the predicted yield from all four models. This is because of extreme drought conditions during the 2005 growing season (Zhang et al., 2006).

Prediction Confidence Band. We use Model 2 to determine the prediction confidence

intervals defined in Section 3.2. The 95% pointwise prediction interval for year 2005 in bushels per acre is **(150.08, 210.46)** for forecasted weather data and **(137.48, 204.25)** for observed weather data. The observed corn yield in year 2005 is 142 bushels per acre. It is close to the lower bound of the confidence interval due to the severe drought conditions during the 2005 growing season. The prediction confidence intervals for years 1996 to 2005 derived using Model 2 are illustrated in Figure 2. In crop decision planning, it is important to obtain the prediction confidence interval for the crop yield to cope with different scenarios or yield outcomes that may occur over the planning horizon for crop decisions (Kantanantha et al, 2008).

4 Price Forecasting

4.1 Model Formulation

In this section, we develop a cash price forecasting model under a futures-based framework where cash price is forecasted as the sum of the nearby futures price and predicted commodity basis. Under this formulation, we only need to predict the commodity basis, which typically does not vary as much as the cash price and can generally be predicted from historical commodity basis patterns (Chicago Board of Trade, 2000). For decision planning, we need to obtain weekly price forecasts, and therefore, weekly commodity basis predictions throughout the planting season and after (May to December). To plan the crop allocation and selling time, a farmer needs to obtain these predictions at the beginning of the planting season (late April).

We first introduce our model for long-range commodity basis prediction. Given the historical commodity basis data

$$Y_{il} = Y_i(t_l), \quad l = 1, \dots, nw, i = 1, \dots, ny$$

where t_l is the l th week for nw number of weeks ($nw = 35$ in our application) and $ny = 14$ is the total number of years, we predict the commodity basis pattern for the upcoming year as a mixture of historical commodity basis patterns. Denote $Y^*(t_l)$

for $l = 1, \dots, nw$ the weekly predictions of the commodity basis for the upcoming year. Under our model formulation, we assume that $Y^*(t)$ is distributed as a mixture of Gaussian processes

$$Y^*(t) \sim \sum_{k=1}^K \pi_k f_k(\cdot|t), \quad (7)$$

where K is the number of mixtures, f_k is the density function of the k th gaussian mixture component with mean $\mu_k(t)$ and covariance $\Sigma_k(t, t')$. Both the mean functions and the covariance surfaces are unknown for $k = 1, \dots, K$. The mixture weights π_k for $k = 1, \dots, K$ are also unknown. We predict the commodity basis using this mixture approach because the basis displays similar underlying patterns across years with a mixture of cyclical patterns. To illustrate this, we present the corn basis from years 1991 to 2006 in Figure 3. Overall, the basis fluctuates within a year with five common local maxima and one local minimum. In these plots, the local maxima are indicated by dashed lines and the local minimum is marked by a straight line. The 1996 corn basis behaves differently from other years because of the passage of the 1996 Farm Act, which increased the planting flexibility and resulted in a large amount of corn released to the market.

4.2 Model Estimation

Since the number of mixtures in (7) is unknown, the estimation problem is an unsupervised clustering problem. We follow the approach for clustering functional data proposed by James and Sugar (2003). Under this model-based approach, the cluster memberships Z_i 's are treated as latent variables. Assuming Z_i for $i = 1, \dots, ny$ are multinomial with parameters (π_1, \dots, π_K) and π_k is the probability that a commodity basis curve belongs to the k^{th} cluster. We estimate the mixture weights π_k , the mean functions $\mu_k(t)$ and covariance surfaces Σ_k as described in James and Sugar (2003) and briefly presented below.

For each year i , the commodity basis is modeled as

$$Y_{il} = \beta_i(t_l) + \varepsilon_{il}, l = 1, \dots, nw,$$

where we decompose the regression function $\beta_i(t)$ using a spline basis $\beta_i(t) = s(t)^T \varphi_i$, where $s(t)$ is a vector of spline basis functions of length q and φ is a spline coefficient vector of length q . We assume that the spline coefficients follow a Gaussian distribution:

$$\varphi_i = \mu_{z_i} + \gamma_i, \quad \gamma_i \sim N(0, \Gamma), \text{ for } i = 1, \dots, ny$$

where μ_{z_i} is the cluster coefficient vector, z_i is the unknown cluster membership of year i taking values from 1 to K , and γ_i is a random effect vector for i th year.

Define $S = (s(t_1)^T, \dots, s(t_{nw})^T)^T$ to be the spline basis matrix, $b_i = (Y_{i1}, \dots, Y_{inw})$, and $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{inw})$ to be the vector of measurement errors. Using these notations, the functional clustering model (FCM) becomes

$$b_i = S(\mu_{z_i} + \gamma_i) + \varepsilon_i, \quad i = 1, \dots, ny, \quad (8)$$

$$\varepsilon_i \sim N(0, \sigma^2 I), \quad \gamma_i \sim N(0, \Gamma).$$

Under this formulation, the distribution of b_i is

$$b_i \sim N(S\mu_{z_i}, \Sigma), \text{ where } \Sigma = \sigma^2 I + S\Gamma S^T. \quad (9)$$

As in FCM (James and Sugar, 2003), we estimate the model parameters by maximizing the mixture likelihood function. Denote the estimated coefficients $\hat{\mu}_k$ for the k th cluster. It follows that for cluster k with $k = 1, \dots, K$, the estimated cluster mean is $\hat{\mu}_k(t) = s(t)^T \hat{\mu}_k$ and the estimated cluster covariance is $\hat{\Sigma}_k(t, t') = \hat{\Sigma}(t, t') = s(t)^T \hat{\Gamma} s(t')^T + \hat{\sigma}^2 I \{t \neq t'\}$.

Confidence Band Estimation. We assume that the prediction curve $Y^*(t)$ follows a mixture of gaussian processes where the k^{th} component has an estimated mean $\hat{\mu}_k(t)$ and estimated covariance $\hat{\Sigma}_k(t, t')$. Also assuming that the clustering variable Z is multinomial with proportion parameters π_k for $k = 1, \dots, K$, we compute an $(1 - \alpha)$

prediction confidence band for $Y^*(t)$ using the formula

$$(\widehat{L}(\alpha, t), \widehat{U}(\alpha, t)) = \left(\sum_{k=1}^K \pi_k \widehat{\mu}_k(t) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \widehat{v}(t) \right), \quad (10)$$

where \widehat{v} is the diagonal of square root of the estimated covariance matrix of $Y^*(t)$ computed from

$$\begin{aligned} \mathbb{V}(Y^*(t)) &= \mathbb{V}(\mathbb{E}(Y^*(t)|Z)) + \mathbb{E}(\mathbb{V}(Y^*(t)|Z)) \\ &= \mathbb{V}\left(\sum_{k=1}^K Y_k(t) \mathbb{E}(1_{Z=k})\right) + \mathbb{E}\left(\sum_{k=1}^K Y_k^2(t) \mathbb{V}(1_{Z=k})\right) \\ &= \sum_{k=1}^K \pi_k^2 \mathbb{V}(Y_k(t)) + \sum_{k=1}^K \pi_k (1 - \pi_k) \mathbb{E}(Y_k^2(t)) \\ &= \text{diag}(s(t) \Sigma s(t)^T) + \sum_{k=1}^K \pi_k (1 - \pi_k) \mu_k^2(t). \end{aligned}$$

A more conservative confidence interval has been suggested by James and Sugar (2003). They suggest estimating the confidence bands for each cluster $k = 1, \dots, K$, $(\widehat{L}_k(\alpha, t), \widehat{U}_k(\alpha, t))$ and take the minimum over all lower bounds and the maximum over all upper bounds

$$\left(\min_{k=1}^K \widehat{L}_k(\alpha, t), \max_{k=1}^K \widehat{U}_k(\alpha, t) \right).$$

The confidence band for $Z = k$ is

$$(\widehat{L}_k(\alpha, t), \widehat{U}_k(\alpha, t)) = \left(\widehat{\mu}_k(t) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \widehat{v}_k(t) \right), \quad (11)$$

where Φ^{-1} is the inverse cumulative density function and v_k is the diagonal of the estimated covariance matrix of the k^{th} cluster, $\widehat{\Sigma}_k$. James and Sugar (2003) also suggest a more accurate (less conservative) interval using only the confidence bands of the clusters with the largest estimated weights $\widehat{\pi}_k$ such that the sum of their weights is equal to a fixed τ_1 . An $\alpha = \tau_1 \tau_2$ confidence band is then

$$\left(\min_{\sum_{k=1}^{K-c} \pi_k \leq \tau_1} \widehat{L}_k(\tau_2, t), \max_{\sum_{k=1}^{K-c} \pi_k \leq \tau_1} \widehat{U}_k(\tau_2, t) \right)$$

where the clusters $K - c, \dots, K$ are those with the lowest weights. In the next section, we compare the two methods for years 2004 and 2005.

4.3 Price Forecasting: Results and Discussion

Functional Clustering Analysis. The functional model-based clustering technique outlined in Section 4.1 is applied to the commodity basis data. We first predict the commodity basis pattern based on the historical commodity basis data scaled to mean zero. For the method briefly discussed in Section 4.2, we need to specify the dimension of the spline basis, q , and the number of clusters, K . The predicted commodity basis cluster means $\hat{\mu}_k(t)$ are similar because we estimate the mixture in (7) using a small number of similar patterns ($ny = 14$). This implies that the estimated mixture is not sensitive to the number of clusters. However, for a larger number of years, the number of clusters may play a significant role. There are techniques that can be used to determine the number of clusters, for example, gap statistic (Tibshirani et al., 2001), and jump method (Sugar and James, 2003) but here we choose a small value for K . We choose q by investigating the smoothness level under a series of values for this tuning parameter and select q to optimize the trade-off between local and global variations.

Prediction Confidence Band. The yearly commodity basis data is scaled to mean zero before estimating the mixture based on the historical commodity basis data. Therefore, both the predicted commodity basis and its confidence bands are not on the same scale as the corresponding observed basis. We therefore re-scale the predicted basis and the prediction confidence bands. Because we are interested in prediction starting with May - beginning of the planting season, we use as scaling constant the average over the commodity basis within the first four months of the prediction year. The proposed re-scaling technique works well in the years with no severe conditions. Lastly, we add the nearby settlement futures price to the predicted commodity basis to obtain the predicted cash price and the prediction confidence bands. We show the prediction results for cash price in Figure 4 for years 2004 and 2005.

We proposed two different methods for estimating the prediction confidence bands in Section 4.2. In Figure 4, we compare the two methods - Method 1 in red is the method introduced in James and Sugar (2003) and Method 2 in blue follows from our assumption that the predicted pattern comes from a mixture of gaussian processes. The first method provides wider confidence bands and most of the observed cash price values are within this confidence band. Using the proposed prediction method, the cash price is over-predicted around the end of the planting season - end of September for both years (2004 and 2005). Moreover, for months September to December, the prediction confidence bands are also wider since the historical commodity basis varies during this time period significantly. Year 2005 shows greater variability and larger prediction errors than 2004. This may come from the severe drought condition in 2005, which affected the corn production and hence the price.

Comparisons. We compare our forecasting model with a simple times series method - ARIMA. We obtain weekly predictions five-week ahead using an ARIMA model with the ARMA and seasonality orders selected using a modified AIC criterion (Brockwell and Davis, 2002). One has to bear in mind that the predictions based on ARIMA are medium-range (lag of five weeks) whereas the predictions using our method are long-range (lag of 35 weeks). The decision planning model makes use of long-range predictions, but for comparison reasons, we obtain the medium-range ARIMA predictions since long-range ARIMA with 35 lags performs poorly. The predictions for both comparison years, 2004 and 2005 are presented in Figure 5. For both years, our method follows the price pattern closely with larger prediction errors in September to November. The mean squared prediction errors for both years are also lower when using our long-range forecasting method than the medium-range forecasting ARIMA. The price forecasting model in this paper allows for long-range prediction because it accounts for yearly seasonality and cycles of the commodity basis. However, for outlying patterns due to severe weather conditions - for example, 2005 with a draught condition, our model consistently over-predicts the price. As more information about the cash price level is available, we can simply adjust the price predictions by rescaling the predicted price pattern. Updating the price prediction may require modifying

the crop decision planning, for example, the dates when to sell the crops. However, not all decisions made at the beginning of the planting season are reversible. One example is the crop allocation. For these decisions, we need long-range predictions as provided in this paper.

5 Discussion

Decision planning plays an important role in agriculture just as it does in other industries. It is a key factor that determines in part the success and failure of business. Farmers' decisions include: i) which crops to produce, ii) how much land to allocate to each crop, and iii) when to grow, harvest, and sell. Uncertain factors such as weather and demand, along with the limited resources used to cultivate, store, and supply crops make crop decisions difficult for farmers and can therefore significantly affect returns.

In order to make good decisions under these uncertainties, forecasting of important factors is a crucial step. The models described in previous sections facilitate computation of not only point prediction but also prediction confidence bands for yield and price. These results enable the analysis of a range of scenarios in decision planning. For example, before each cropping period begins, farmers have to consider which crops they will grow. This can be a difficult decision since farmers do not have information for the outcome of their crops before the cropping season. The expectations for yield and price would help producers to estimate the expected return for each crop. The prediction confidence bands would provide other possible outcomes beside the expected values. A second example is the decision of when to harvest and when to sell the crop. If crops are storable, farmers may keep them for sale at higher prices after the harvest time. By doing so, producers may obtain higher return. However, this decision will depend on whether the increased price is high enough to compensate for the storage costs. The price prediction confidence band would help the growers to estimate the maximum loss or maximum gain. With these estimations along with the realized harvest cash price, farmers can make better decisions about

whether they should keep their crops for later sale or not. Therefore, our forecasting models are usable tools for rigorous decision planning.

The forecasting models in this paper allow prediction of the yield and price at the beginning of the planting season when most of the decisions are made. For yield forecasting, we need to use long-range weather forecasts since the observed weather data throughout the planting season are not available. Because of this limitation, the yield predictions are less precise. However, because our model accounts for within-year dependence using time-varying regression coefficients, the yield prediction borrows information across years, and therefore, smooth changes may be accurately predicted. Years 1996 to 1999 are example of smooth variations in the yield; our model provides accurate predictions for these years. On the other hand, year 2005 is an example of a sharp change in the yield due to a severe drought. For years with severe weather conditions, the model relies on good weather predictions. We use simple weather forecasts based on the ARIMA model but more accurate predictions may be used here as described in Section 2.1. Another important aspect of our model is dimensionality reduction. Because weather data may be represented by multiple between-year dependent predictors, we propose using Functional Principal Component Analysis to reduce the space of predictors. This step is important since a large number of predictors leads to over-fitting, and therefore, inaccurate predictions. In conclusion, varying-coefficients and variable reduction by FPCA enhance the yield prediction as provided in our empirical case study. We compared yield forecasts over 10 years (1996-2005). The model proposed in this paper out-performs the most common approach to yield forecasting - linear regression in average across the prediction years.

Under the futures-based model, we obtain commodity basis predictions and sum these predictions to the futures price to obtain a cash price forecast. A multiple-year average technique is often used as a tool to compute the expected commodity basis. It is simple and provides relatively insightful results. However, it assumes equal weights for all years without accounting for yearly patterns which are predominant across multiple years. For example, if there is a year which has an outlying pattern from all

the other years, then this year will be equally weighted as all the other years. In our model, we estimate the mixture weights. The model based formulation also allows us to estimate (pointwise) confidence band, which are further used in decision planning. We evaluate the precisions of the prediction bands for years 2004 and 2005. For year 2004, the prediction bands cover most of the observed price values throughout the eight-month prediction period. For year 2005, the confidence bands are not precise during the harvesting period due to the severe weather condition. If we were to have available data for a longer period of time, we might be able to cluster years with severe conditions like flood and drought. This would allow us to accurately identify 2005 as a drought year and correctly account for this severe condition.

In a subsequent paper, we introduce an optimization-based decision planning model for these decision problems. (Kantanantha et al, 2008) Since yield and price are stochastic and affect both revenues and costs, these stochastic variables will be incorporated in the model to derive probabilities for different planning scenarios. Our model takes into account the resource limitations as its major constraints as well as the limitation on the planting and harvesting periods of each crop.

Appendix

In this Appendix, we discuss a simulation study for evaluating the precision and the accuracy of the predictions derived from the yield forecasting model. Specifically, we simulate from the full model in (1) where the regression coefficients take one of the four forms

$$f_1(t) = a_0$$

$$f_2(t) = a_0 + a_1t$$

$$f_3(t) = a_0 + a_1t + a_2t^2$$

$$f_4(t) = a_0 + a_1t + a_2t^2 + a_3t^3$$

with the function coefficients being randomly sampled from a uniform distribution on $(0, 1)$ ($a_0, a_1, a_2, a_3 \sim \mathcal{U}(0, 1)$). The error term ϵ_i is normally distributed with variance $\sigma^2 = 1$. We consider the same number of years as for the observed yield in our empirical example and the same set of predictors (GDP, 22 temperature and 22 rainfall predictors). For 100 simulated data sets, we apply the yield forecasting model as described in Sections 3.1-3.3. We first reduce the set of predictors using FPCA and then estimate the model in (3). We obtain predictions for years 1996 to 2005 using a parsimonious version of the model in (3). For each simulation, we count the number of years the prediction interval covers the observed value.

We apply this simulation study under two scenarios discussed in this paper - prediction under the observed weather data (scenario I) and prediction under forecasted weather data (scenario II) for the prediction year. The second scenario allows us to evaluate the precision and accuracy of the confidence interval in the presence of the prediction error due to using the forecast rather than observed weather data. Overall, 52 prediction intervals out of 1000 (100 simulations times ten prediction years) and 60 out of 1000 do not cover the true simulated value when predicting based on observed weather data, and respectively, when predicting based on forecast weather data. Another precision measure is to evaluate for how many datasets out of 100, there are two or more years with confidence intervals not covering the simulated value. For example, for one simulation, the prediction confidence intervals may not cover the observed values for years 1996, 1997 and 1998. For the first scenario, 5 simulations out of 100 provide prediction confidence intervals not covering the simulated values for two prediction years. In contrast, there are 10 simulations out of 100 under the second scenario. Using this precision measure, the prediction confidence intervals under the second scenario are much less precise than for the first scenario.

In Table 3, we report mean absolute prediction error (MAPE) and the length of the prediction confidence interval divided by the simulated true value. We show these accuracy measures for each of the prediction years. The reported prediction errors and standardized confidence interval lengths are averages over 100 simulations. The MAPE's do not vary much for years 1996 to 2005. The standardized confidence

interval length is a measure of accuracy of the confidence interval. The lower this value is, the higher the accuracy is at a given precision ($1 - \alpha = .95$ in this simulation). The accuracy of the prediction intervals do not differ for the two scenarios. In conclusion, using forecast weather data does not affect the accuracy of the confidence intervals but their precision.

References

- [1] Batts, G.R., Morison, J.I.L., Ellis, R.H., Hadley, P., Wheeler, T.R. (1997), “Effects of CO₂ and temperature on growth and yield of crops of winter wheat over four seasons,” *European Journal of Agronomy*, 7, 43-52.
- [2] Bohrnstedt, G.W., Goldberger, A.S. (1969), “On the exact covariance of products of random variables,” *Journal of the American Statistical Association*, December, 1439-1442.
- [3] Bureau of Economic Analysis, United State Department of Commerce.
- [4] Brockwell P.J., Davis, R.A. (2002), *Introduction to Time Series and Forecasting*, Springer.
- [5] Bureau of Labor Statistics, United State Department of Labor (2005), “Career guide to industries 2006-2007 editon.”
- [6] Cai, T., Hall, P. (2006), “Prediction in functional linear regression,” *Annals of Statistics*, 34,5, 2159.
- [7] Eales, J.S., Engel, B.K., Hauser, R.J., Thompson, S.R. (1990), “Grain price expectations of Illinois farmers and grain merchandisers,” *American Journal of Agricultural Economics*, 72(3), 701-708.
- [8] Ferraty, F., Vieu, P. (2006), *Nonparametric Functional Data Analysis*, Springer.

- [9] Hoogenboom, G. (2000), "Contribution of agrometeorology to the simulation of crop production and its applications," *Agricultural and Forest Meteorology*, 103, 137-157.
- [10] Hoffman, L.A. (2005), "Forecasting the counter-cyclical payment rate for U.S. corn: An application of the futures price forecasting model," Outlook Report No. FDS-05a-01, Economic Research Service.
- [11] James, G.M., Sugar, C.A. (2003), "Clustering for sparsely sampled functional data," *Journal of the American Statistical Association*, 98, 397-408.
- [12] Kandiannan, K., Chandaragiri, K.K., Sankaran, N., Balasubramanian, T.N., Kailasam, C. (2002), "Crop-weather model for turmeric yield forecasting for Coimbatore District, Tamil Nadu, India," *Agricultural and Forest Meteorology*, 112, 133-137.
- [13] Kantanantha, N., Serban, N., Griffin, P.M., Assavapokee, T. (2008), "Crop Decision Planning under Yield and Price Uncertainties", submitted.
- [14] Kenyon, D., Jones, E., McGuirk, A. (1993), "Forecasting performance of corn and soybean harvest futures contracts," *American Journal of Agricultural Economics*, 75, 399-407.
- [15] Kenyon, D., Lucas, K. (1998), "Soybean pricing guide," Department of Agricultural and Applied Economics, Virginia Tech, Virginia, USA, REAP Report No. 37.
- [16] Krog, D.R. (1998), Plant-process model corn yield forecasts for Iowa, Ph.D. Dissertation, Iowa State University, Iowa.
- [17] Lee, R. (1999), Modeling corn yields in Iowa using time series analysis of AVHRR data and vegetation phenological metrics, Ph.D. Dissertation, University of Kansas, Kansas.

- [18] Mkhabela, M.S., Mkhabela, M.S., Mashinini, N.N. (2005), “Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA’s-AVHRR,” *Agricultural and Forest Meteorology*, 129, 1-9.
- [19] National Agricultural Statistics Service, United State Department of Agriculture (2005), “Agricultural statistics 2005.”
- [20] National Agricultural Statistics Service, United State Department of Agriculture (1997), “Usual planting and harvesting dates for U.S. field crops.”
- [21] National Climatic Data Center, United State Department of Commerce.
- [22] Peng, S., Huang, J., Sheehy, J.E., Laza, R.C., Visperas, R.M., Zhong, X., Centeno, G.S., Khush, G.S., Cassman, K.G. (2004), “Rice yields decline with higher night temperature from global warming,” *Proceedings of the National Academy of Sciences of the United States of America*, 101, 9971-9975.
- [23] Potgieter, A.B., Hammer, G.L., Doherty, A., de Voil, P. (2005), “A simple regional-scale model for forecasting sorghum yield across North-Eastern Australia,” *Agricultural and Forest Meteorology*, 132, 143-153.
- [24] Ramsay, J.O., Silverman, B.W. (1997), *Functional Data Analysis*, 1st edition, Springer, New York.
- [25] Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge, New York.
- [26] Sheehy, J.E., Mitchell, P.L., Ferrer, A.B. (2006), “Decline in rice grain yields with temperature: Models and correlations can give different estimates,” *Field Crops Research*, 98, 151-156.
- [27] Sood, A., James, G. and Tellis, G. (2008), “Functional regression: A new model for predicting market penetration of new products,” *Marketing Science*, to appear.

- [28] Sugar, C.A., James, G.M. (2003), "Finding the number of clusters in a data set: An information theoretic approach," *Journal of the American Statistical Association*, 98, 750-763.
- [29] Tibshirani, R., Walther, G., Hastie, T. (2001), "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society, B*, 63, 411-423.
- [30] Tomek, W.G., Gray, R.W. (1970), "Temporal relationships among prices on commodity futures markets: Their allocative and stabilizing roles," *American Journal of Agricultural Economics*, 52, 372-380.
- [31] Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics.
- [32] Wheeler, T.R., Craufurd, P.Q., Ellis, R.H., Porter, J.R., Prasad, P.V.V. (2000), "Temperature variability and the yield of annual crops," *Agriculture, Ecosystems & Environment*, 82, 159-167.
- [33] World Agricultural Outlook Board, United State Department of Agriculture, "World agricultural supply and demand estimates."
- [34] Working, H. (1942), "Quotations on commodity futures as price forecasts," *Econometrica*, 10, 39-52.
- [35] Yao, F., Müller, H.G., Wang, J.L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577-590.
- [36] Zhang, P., Anderson, B.T., Myneni, R. (2006), "Monitoring 2005 corn belt yields from space," *EOS, Transactions American Geophysical Union*, 87, 150.

	Model 1	Model 2	Model 3	Model 4
MSE*	566.61	142.81	190.35	185.41
MAPE**	0.156	0.053	0.078	0.074
R ² -adj ***	90.10 %	88.45 %	78.07 %	77.80 %

Table 1: Corn yield prediction results from observed weather data for Model 1 to Model 4. (*Mean prediction square error; ** Mean absolute percentage error; *** Using Data 1927-2005)

	Model 1	Model 2	Model 3	Model 4
MSE*	308.10	234.94	281.07	278.88
MAPE**	0.102	0.072	0.080	0.094
R ² -adj ***	89.97 %	88.30 %	78.07 %	77.80 %

Table 2: Corn yield prediction results using forecasted weather data for Model 1 to Model 4. (*Mean prediction square error; ** Mean absolute percentage error;*** Using Data 1927-2004)

Year	MAPE (I)	MAPE(II)	Stand CI length (I)	Stand CI length (II)
1996	0.127	0.119	0.62	0.61
1997	0.130	0.126	0.58	0.58
1998	0.110	0.108	0.60	0.55
1999	0.109	0.109	0.52	0.52
2000	0.116	0.117	0.55	0.50
2001	0.112	0.092	0.50	0.46
2002	0.093	0.113	0.49	0.41
2003	0.095	0.095	0.49	0.44
2004	0.097	0.091	0.46	0.41
2005	0.093	0.091	0.44	0.39

Table 3: Simulation study: Accuracy of point prediction and confidence interval prediction under two scenarios: using observed weather data (I) and using forecast weather data (II).

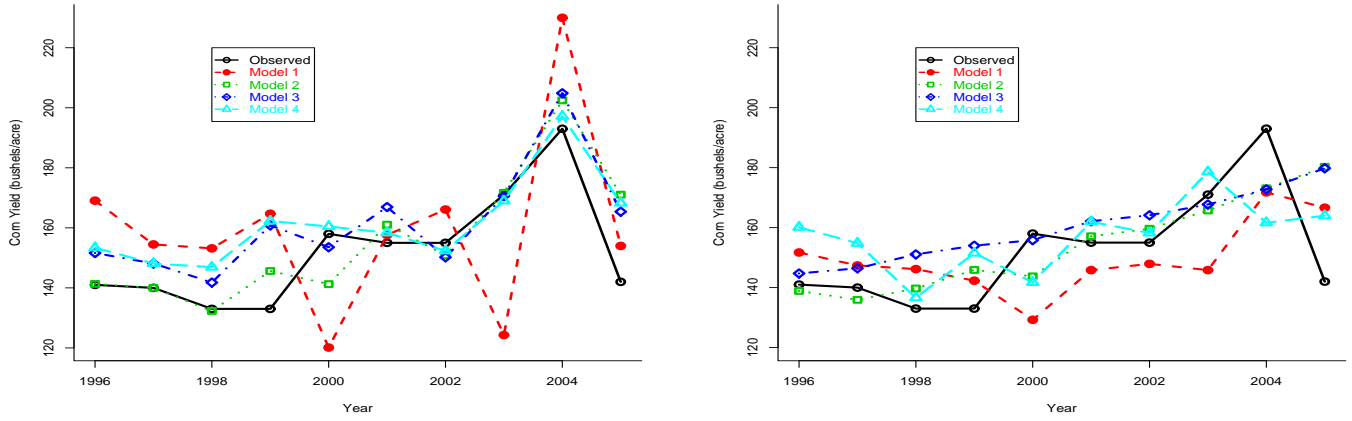


Figure 1: Observed corn yield and predicted corn yield: (left) Model 1-4 predictions for years 1996-2005 based on observed weather data; (right) Model 1-4 predictions for years 1996-2005 based on forecasted weather data.

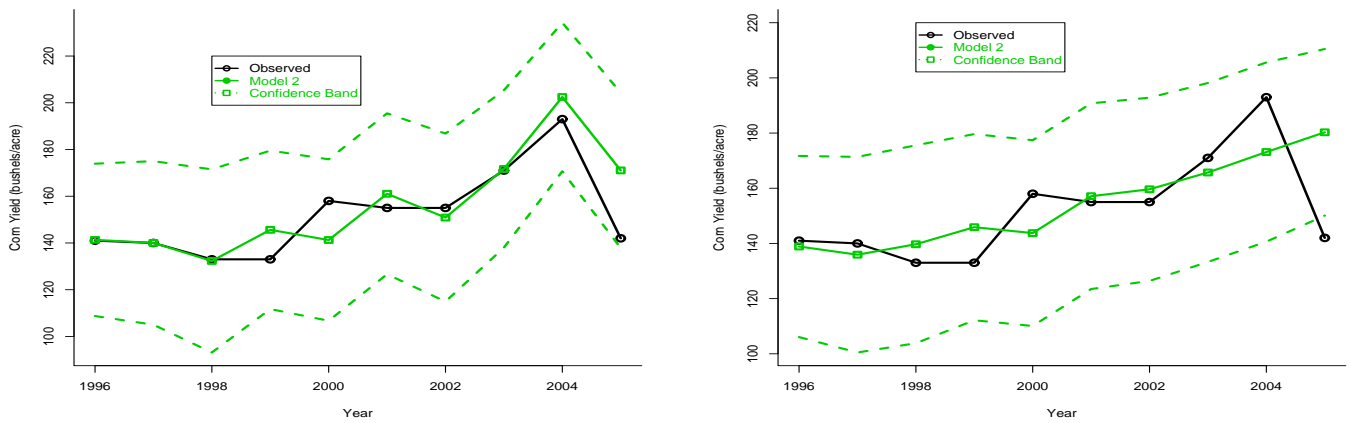


Figure 2: 95% pointwise prediction confidence bands estimated using Model 2 for the prediction years 1996-2005: (left) for the observed weather data and (right) for the forecasted weather data.

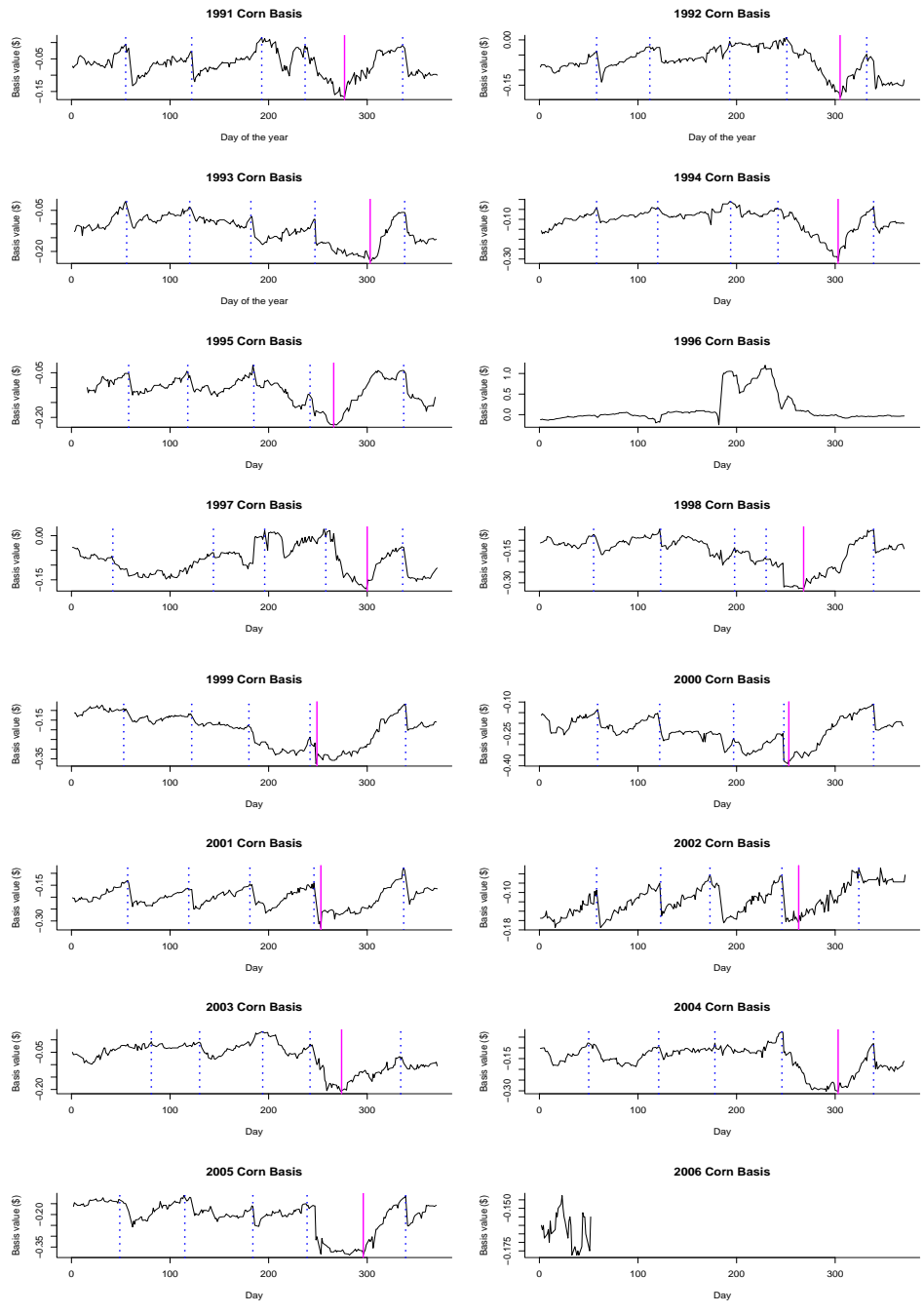


Figure 3: Corn basis plots from 1991 to 2006.

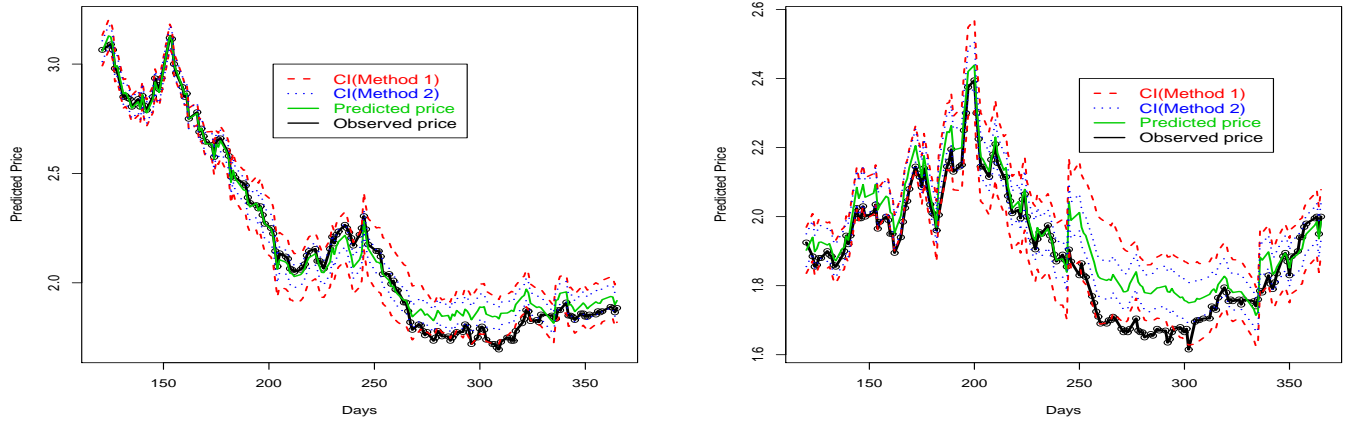


Figure 4: 95% confidence prediction bands for cash price throughout the planting season and after (May-December) for 2004 (left plot) and 2005 (right plot).

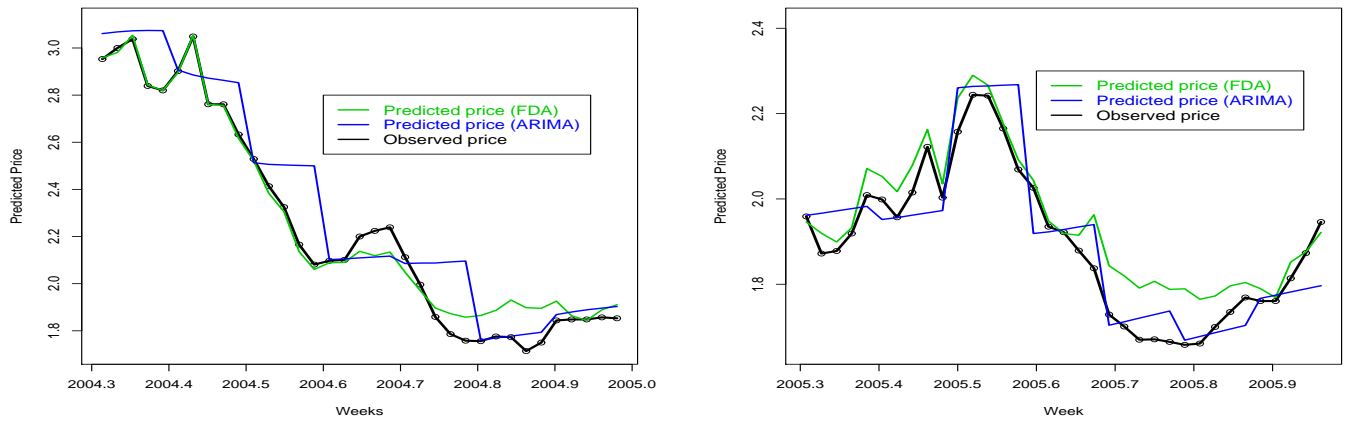


Figure 5: Price prediction comparison for year 2004 (left plot) and year 2005 (right plot). The FDA prediction method is the method introduced in this paper.