Improved Index Insurance Design and Yield Estimation using a Dynamic Factor Forecasting Approach

Hong Li*, Lysa Porth†, Ken Seng Tan‡, and Wenjun Zhu§

January 21, 2020

Abstract

Accurate crop yield forecasting is central to effective risk management for many stakeholders, including farmers, insurers, and governments, in various practices, such as crop management, sales and marketing, insurance policy design, premium rate setting, and reserving. This paper first investigates an innovative approach of yield forecasting using a dynamic factor model. Based on the proposed approach, we then design an enhanced weather index-based insurance (IBI) policy. The dynamic factor approach is motivated by its ability to effectively summarize the information in a large set of explanatory variables with common factors of a much lower dimension. This makes it possible to use an extensive set of variables in crop yield prediction without worrying about identification issues. Using both county-level and state-level crop production data from the state of Illinois, U.S., the empirical results show that the dynamic factor approach produces more accurate in- and out-of-sample forecasting results compared to the classical statistical models. The empirical results also support that the proposed IBI policy based on the dynamic forecasting model has small basis risk. This, in turn, greatly improves the IBI's hedge effectiveness against agricultural production as well as increases its practicality as an insurance policy for agriculture.

Keywords: Crop Yield Forecasting; Factor Model; Index-based Insurance.

"As we have seen, unexpected events can easily take markets away from these central trends. In the past I have talked about 'the hikes in the spikes'; now we have to beware of 'the bumps in the slumps'!"

Launch of the OECD-FAO Agricultural Outlook 2017-2026
Remarks by Angel Gurría, OECD Secretary-General
Paris, France, July 10, 2017

1 Introduction

Agriculture is often recognized as the foundation of the economy. While significant technological advancement has occurred in the agricultural sector over the past decades, great need for further improvement remains to meet the estimated 70 percent increase in food productivity needed to feed the world's growing population by 2050 (FAO, 2009). Further, with mounting concerns over possible changes in climate, which can lead to

*li.hong@umanitoba.ca. Warren Centre for Actuarial Studies and Research, Asper School of Business, University of Manitoba, Canada.
†lysa.porth@umanitoba.ca. Warren Centre for Actuarial Studies and Research, University of Manitoba, Canada.
‡kstan@uwaterloo.ca. Department of Statistics and Actuarial Science, University of Waterloo, Canada.
§wjzhu@ntu.edu.sg. Division of Banking & Finance, Nanyang Business School, Nanyang Technological University, Singapore.

significant widespread crop losses, finding more efficient and sustainable risk management approaches is of great importance.

Accurate yield forecasting is central to effective risk management for many stakeholders, including farmers, insurers, reinsurers and governments. From the farmers' perspective, the ability to accurately forecast crop yield before the harvest season allows them to make decisions regarding crop sales and marketing, and risk management practices such as buying insurance policies to mitigate losses due to reduced crop yield. For insurers and governments, improved approaches to predict yield can contribute to more effective and sustainable ratemaking (Ozaki et al., 2008), as well as the formation of more informed policies and innovative risk management approaches that can protect producers from potentially disastrous outcomes due to extreme adverse weather, such as droughts or floods. Therefore, the first goal of this paper is to investigate an innovative approach of crop yield projection which could potential improve forecast accuracy.

There have been two popular types of crop yield forecasting approach. The first type is the process-based models, which are based on intensive simulation algorithms with a large number of covariates on crop management, such as planting dates, phenology, and seed variety, nutrient conditions, water and energy availability (Hoogenboom et al., 2004; Bondeau et al., 2007; Basso et al., 2010; Asseng et al., 2013). However, while the process-based models are more popular in developed countries, broader applications have been limited. In particular, it is not always possible to obtain such detailed information, especially in less developed countries or regions.

In the latter cases, the second type of approach, statistical models, are more useful. Statistical models use more easily measurable meteorological variables, such as temperature and precipitation over a broader region, to forecast crop yields on particular farms. Moreover, statistics-based models have simpler structures, and are more straightforward and transparent in assessing goodness-of-fit using statistics such R^2 and information-based criteria. In addition, technological advancements, such as remote sensing, could provide statistics models with more valuable inputs, and thus improve their ability to better forecast yields.

Despite the advantages of statistical models, limitations exist as well. In particular, it is often difficult to determine the appropriate variables to include in the regressions. Existing literature typically considers a large number of variables, including monthly temperatures and precipitations during the growing season, as well as their higher order terms and interactions (Thompson, 1969, 1988). Candidate explanatory variables of such a large size can easily lead to variable selection problems. On the one hand, crop yield data are scarce. Even for countries with fairly good historical data, e.g., Canada and the U.S., there are typically only annual observations for around 40 years. With limited historical data, estimation errors will increase rapidly with the number of regressors. Furthermore, estimation would even become infeasible if dimension of regressors exceeded the number of observations. On the other hand, a model with too few regressors will often result in lower explanatory and prediction power (Lobell and Burke, 2010). To make things worse, the

lack of an effective guideline for variable selection makes the estimation problem more difficult.

Therefore, the usefulness of a statistical model for crop yield prediction depends crucially on its ability to cope with a large number of regressors. To this end, different statistical approaches were proposed in the literature. For example, the LASSO (Tibshirani, 1996) approach uses penalized regression with an $\mathcal{L}_1-$penalty imposed, and can effectively achieve continuous shrinkage and automatic variable selection. However, when the number of regressors (J) is larger than the number of observations (T), LASSO can only select at most T regressors (Efron et al., 2004). To address this limitation, elastic net has been proposed (Zou and Hastie, 2005). An alternative dimension reduction method is the factor model, which summarizes the information in a high-dimensional set of explanatory variables by lower-dimensional latent (unobservable) common factors. In contrast to the LASSO approach, factor models are able to deal with situations where $J \geqq T$. While factor models have a long tradition in the modeling of large dimension data, such as the principal component analysis (Jolliffe, 2002), it has been applied extensively in modern researches of various fields. Examples include theoretical econometrics (Bai and Ng, 2002; Pesaran et al., 2013), high-dimension and high-frequency data modeling (Lam et al., 2012; Aït-Sahalia and Xiu, 2019), big data analysis (Fan et al., 2014; Zhu et al., 2017), macroeconomic analysis (Ludvigson and Ng, 2009; Jurado et al., 2015), corporate finance (Eberhardt et al., 2013), asset pricing (Ludvigson and Ng, 2007; Weron, 2014), and mortality modeling (French and O'Hare, 2013; Hyndman et al., 2013), among many others.

In this paper, we follow the factor modeling literature, and propose a dynamic factor model for crop yield prediction. The proposed method first summarizes a large set of explanatory variables with common latent factors of a lower dimension, based on the information criteria proposed by Bai and Ng (2002). Then, an optimal subset of the common factors are selected to predict crop yield. Bai and Ng (2002) propose a factor selection method to consistently estimate the number of factors included in a regression when both the cross-section dimension and the time dimension grow to infinity. Moreover, this method is shown to be satisfying finite sample properties, and can thus lead to a reasonable selection of factors when the sample size is limited. Their method, which will be discussed in Section 2, has been used in recent research in many areas, for example, see Ludvigson and Ng (2007, 2009); McCracken and Ng (2016). After the factor model is estimated, the prediction model of crop yield using original variables can be approximated by linear transformation. The factor model makes it possible to start the investigation from a rich database, and evaluate the impact of each single variable on crop yield. Finally, the proposed method is able to deal with situations where the number of explanatory variables is larger than the number of observations. In the empirical analysis, the dynamic factor model is applied to a design matrix consisting of 116 variables, including 61 weather variables and 55 soil condition variables. A data-driven approach is used to determine the optimal factors included in the regression model. Our empirical study show that the dynamic factor approach is able to produce more accurate forecasting results than the traditional statistical models.

This paper also contributes to the weather index insurance literature by developing a new agricultural index-based insurance (IBI) approach using the proposed dynamic factor approach. Formal insurance contracts are viewed as important instruments to help ensure long-term stability and growth of the agriculture sector and provide income stabilization to farmers in times of loss (Iturrioz, 2009). In many cases, agricultural insurance policies are designed based on traditional, indemnity approaches that require on-farm loss assessments. For this type of policies, assessments can be quite costly, and there can also be significant moral hazard, i.e., the tendency to take on greater risk once insured; and adverse selection, i.e., when higher risk farmers are more inclined to seek insurance (Miranda, 1991; Miranda and Glauber, 1997; Rowell and Connelly, 2012; Goodwin, 1993; Hao et al., 2018). These obstacles are often regarded as major contributors to the failure of private agricultural insurance market (Goodwin and Smith, 1995; Woodard and Garcia, 2008).

As a potential alternative to traditional indemnity-based insurance policies, IBI has received increasing attention in recent years. IBI policies deliver payments related to the realization of an underlying index, which is based on public information relevant to the actual loss experiences on the farm. In case of IBI, administrative costs are reduced given that there is no need of field investigations needed to measure the actual loss on the farm. Additionally, there is significantly less adverse selection and moral hazard since farmers are unlikely to be more informed about the (public) index than the insurer, and they have no control over the outcome of the index (Barnett and Mahul, 2007; Chantarat et al., 2013; Conradt et al., 2015; Maestro et al., 2016; Mahul and Skees, 2007; Masshoff et al., 2011; Jensen et al., 2016; Shen and Odening, 2013). Moreover, since IBI is derived from an independently verifiable index, insurers could efficiently transfer risk to international reinsurance markets (Hazell et al., 2010), or in some cases through securitization to capital markets (Cummins et al., 2004; Lee and Yu, 2002; Phélippé-Guinvarc'h and Cordier, 2006; Hainaut, 2019).

While promising, IBI also faces its own challenges, and in practice the demand has been unfortunately low (Cole et al., 2013). Among a number of possible explanations, the most important one is basis risk (Chantarat et al., 2013; Cummins et al., 2004; Deng et al., 2007; Jensen et al., 2016). Basis risk refers to the risk that the underlying index and actual losses are mismatched, and hence could lead to situations in which farmers are not indemnified for actual losses, or are paid indemnities despite having no actual losses (Skees et al., 1997; Zhang et al., 2018). Therefore, minimizing basis risk is crucial because farmers will not be fully insured if basis risk is present even when rates are actuarially sound (Mobarak and Rosenzweig, 2012). Although some basis risk reduction methods have been suggested for IBI products, the lack of demand for such contracts indicates that more effective approaches are needed. Using a dynamic factor crop yield forecasting model, this paper proposes a new IBI policy with a focus on reducing basis risk.

The rest of the paper proceeds as follows. Section 2 describes the econometric framework of the dynamic factor approach. Section 3 discusses the empirical results based on three major crops in the State of Illinois

in the U.S.: corn, soybeans, and wheat. In Section 4, a new IBI policy is proposed based on the dynamic factor approach, and the resulting basis risk of the product is evaluated. Section 5 concludes. Proofs and the explanatory variables considered in this paper are collected in the Appendix.

2 Econometric Framework

In this section, we describe our proposed dynamic factor based econometric framework. In particular, Subsection 2.1 focuses on two popular model specifications, namely the time-series and the cross-section specifications. Then Subsection 2.2 presents an algorithm of selecting an optimal combination of common factors. Finally Subsection 2.3 describes a procedure to approximate the original relationship between the crop yield and the explanatory variables implied by the optimal factor model.

2.1 Model Specification

Various model specifications for forecasting crop yields have been proposed in the literature.[1] In this paper, we consider two types of model specifications: (a) the time-series specification, which focuses on the dynamic relationship between crop yields and the explanatory variables for a given location (e.g. farm, county, or state); and (b) the cross-section specification, which studies the average relationship between crop yields and the explanatory variables across multiple locations (Lobell and Burke, 2010). A further advantage of the cross-section model is that spatially varying, but time-invariant, variables (such as many soil quality measurements) can be incorporated, and this may provide additional explanatory power for crop yields. In this paper, we focus on data at the county and the state level. For a particular crop, suppose that we have N counties and T years of data, the above two types of specifications are given below.

(a) Time-series specification:

$$\log(y_{i,t}) = a_i + b_i t + \boldsymbol{\gamma}_i' \boldsymbol{X}_{i,t} + \varepsilon_{i,t}, \tag{1}$$

where $y_{i,t}$ is the yield in county i at year t, $i = 1, \ldots, N, t = 1, \ldots, T$; $\boldsymbol{X}_{i,t}$ is a $J \times 1$ column vector containing the explanatory variables; a_i is the intercept; $\boldsymbol{\gamma}_i$ is a $J \times 1$ column vector of coefficients; $b_i t$ is the time trend that accounts for the impact of technology improvements; and $\varepsilon_{i,t}$ represents the independently and identically distributed (iid) normal residuals.

(b) Cross-section specification:

$$\log(y_{i,avg}) = a_i + \boldsymbol{\gamma}_i' \boldsymbol{X}_{i,avg} + \varepsilon_i, \tag{2}$$

where $y_{i,avg}$ is the county-level time average of the crop yield for county i and $\boldsymbol{X}_{i,avg}$ is a $K \times 1$ column vector of explanatory variables. The subscript "avg" emphasizes that explanatory variables

[1] See Shi et al. (2013) for a comprehensive review.

are averaged over time, and ε_i represents iid normal residuals.

The time-series model specification (1) is based on the county level. An analogous specification for a given state is also possible, as shown below:

$$\log(y_t) = a + bt + \boldsymbol{\gamma}' \boldsymbol{X}_t + \varepsilon_t, \tag{3}$$

where y_t is the state level yield, $\boldsymbol{X}_t$ is a $J \times 1$ vector of state level explanatory variables in year t,[2] bt is the time trend, and ε_t are iid normal residuals. It's worth noticing that, in Models (1) and (3), the validity of including the time trend depends on the stationarity of the data. Hence, in the empirical application, the stationarity of each crop yield series should be tested before one can decide whether to include the time trend. To conclude this section, we point out that there should not be any confusion in terms of the interpretation of a_i and γ_i even though we have used the same notation in the above model specifications.

2.2 Dynamic Factor Approach

In this subsection, the dynamic factor approach for selecting common latent factors is discussed. The underlying assumption of the dynamic factor model is that sufficiently large proportion of variations in the agricultural production processes can be summarized by a relatively small number of unobserved factors. The dynamic modeling consists of two steps. First, a set of latent factors is estimated through principal component analysis (PCA).[3] Second, a dynamic factor procedure is used to select key factors that are important for yield forecasting. As a result, yield forecasting becomes a model selection problem, in which likelihood-based information criterion is used to achieve a good balance between accuracy and parsimony.

The procedure begins with constructing a high-dimensional design matrix (J dimensions for time-series specification and K dimensions for cross-section specification[4]). The issue of including too many variables in the regression models is the significant loss of degree-of-freedom, which results in the regression models quickly losing reliability and credibility. The extreme case is when $J+1 > T$ or $K > N$, where estimation is not even feasible with the classical regression approach. As a result, previous research typically constructs yield forecasting models based a limited number of explanatory variables, which may cause the problem of underfitting (Lobell et al., 2011; Lobell and Burke, 2010). It is not until recently, however, that advanced statistical learning techniques were applied to crop yield forecasting. For example, Bornn and Zidek (2012) incorporate the principal component regression (PCR) into a Bayesian framework to deal with the problem

[2] This paper will mostly focus on the relationship between crop yields and meteorological and soil condition variables. However, lagged terms of $\log(y_t)$ can also be included in the regressors.

[3] For the theoretical background of PCA, we refer to Jolliffe (2002).

[4] In practical application, both J and K can be very large. The database used in our empirical analysis has $J = 61$ and $K = 116$, whereas the sample size is at most 55 (resp. 76) for the time-series (resp. cross-section) specification.

of spatial dependence; Sharif et al. (2017) compare different dimension reduction techniques in crop yield forecasting, including regularization methods such as LASSO, ridge regression, and elastic net.

In this paper, we propose to use a dynamic factor analysis to achieve efficient dimension reduction. In particular, the proposed factor-based technique starts with a large set of dynamic factors estimated by PCA, and then the most useful information is filtered out following a dimension reduction algorithm.

For notation brevity, the county subscript "i" is suppressed in the remainder of this subsection. Denote by $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1', ..., \widetilde{\boldsymbol{X}}_T')'$ the demeaned $J \times T$ data matrix, i.e., the mean of each row in $\widetilde{\boldsymbol{X}}$ is normalized to be 0. We posit that each conditioning variable in the demeaned matrix, denoted by $\tilde{x}_{j,t}$, has the following dynamics

$$\tilde{x}_{j,t} = \boldsymbol{\lambda}_j' \boldsymbol{f}_t + \omega_{j,t}, \tag{4}$$

where $\boldsymbol{f}_t$ is a $r \times 1$ vector of latent factors common to each j, $\boldsymbol{\lambda}_j$ is a $r \times 1$ vector of parameters to be estimated, and $\omega_{j,t}$'s are iid error terms. The number of common factors, r, is constant for all t. Of particular importance is the condition that $r \ll \min(J, T)$, so that substantial dimension reduction can be achieved. Therefore,

$$\log(y_t) = a + bt + \boldsymbol{\theta}' \boldsymbol{F}_t + \epsilon_t, \tag{5}$$

where $\boldsymbol{F}_t$ is a vector constructed from $\boldsymbol{f}_t$ and with a dimension smaller than r. With slight abuse of notation, we denote this by $\boldsymbol{F}_t \subseteq \boldsymbol{f}_t$. To achieve an effective forecasting model, it is crucial to determine $\boldsymbol{F}_t$ in an appropriate manner, as factors that have large explanatory power on $\widetilde{\boldsymbol{X}}$ may not be important for predicting crop yields. We propose to determine the optimal $\boldsymbol{F}_t$ based on the following two steps. First, the latent factor $\boldsymbol{f}_t$ is estimated through PCA with the optimal number of PC's determined by the panel information criteria proposed by Bai and Ng (2002). Second, the optimal factor is determined via a model selection procedure. We explain the procedure in more details in the remaining of this subsection.

Since neither the common factors $\boldsymbol{f}_t$, nor its dimension r, are observable, they are to be estimated and replaced by the $\hat{r}$-dimensional estimator $\hat{\boldsymbol{f}}_t$. Let $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_J)'$ be the $J \times \hat{r}$ matrix, and $\hat{\boldsymbol{f}} = (\hat{\boldsymbol{f}}_1', ..., \hat{\boldsymbol{f}}_T')'$ be the $\hat{r} \times T$ matrix, which contains the first $\hat{r}$ principal component scores of the data matrix $\widetilde{\boldsymbol{X}}$. Then PCA produces the estimates $\hat{\boldsymbol{f}}_t$ and $\hat{\boldsymbol{\Lambda}}$ that satisfy the following two conditions:

1. $\hat{\boldsymbol{f}}_t$ is a linear combination of $\widetilde{\boldsymbol{X}}_t$, i.e., $\hat{\boldsymbol{f}}_t = \hat{\boldsymbol{\Lambda}}' \widetilde{\boldsymbol{X}}_t$ for all t;
2. $\hat{\boldsymbol{\Lambda}}$ minimizes the sum of squared residuals $\sum_{t=1}^{T} (\widetilde{\boldsymbol{X}}_t - \boldsymbol{\Lambda} \boldsymbol{f}_t)' (\widetilde{\boldsymbol{X}}_t - \boldsymbol{\Lambda} \boldsymbol{f}_t)$[5].

Using PCA, the maximal number of factors that can be estimated is $\min(J, T)$. However, these factors may not be all statistically important for crop yield forecasting. Bai and Ng (2002) develop asymptotic properties for consistent estimation of $\boldsymbol{f}_t$ and r when J and $T \to \infty$. Following their framework, we

[5] To make $\boldsymbol{\Lambda}$ and $\boldsymbol{f}_t$ separately identifiable, the normalization $\boldsymbol{\Lambda}' \boldsymbol{\Lambda} = \mathbf{I}_r / J$ is imposed.

determine the optimal number of factors, $\hat{r}$, by the panel information criteria. Specifically, we impose the following assumptions on the true factor representation in Equation (4)[6]:

(i) $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_t \boldsymbol{f}_t^{'} \to \Sigma_{\boldsymbol{f}} = E(\boldsymbol{f}_t \boldsymbol{f}_t^{'})$ as $T \to \infty$, and $\Sigma_{\boldsymbol{f}}$ is positive definite with rank r;

(ii) $||\boldsymbol{\Lambda}^{'}\boldsymbol{\Lambda}/J - \Sigma_{\Lambda}|| \to 0$ as $J \to \infty$ for some positive definite matrix Σ_{Λ};

(iii) The idiosyncratic errors $\omega_{j,t}$ have bounded moments, and covariance across both time and cross-section dimension;

(iv) Weak dependence between factors and idiosyncratic errors, that is, there exists a positive constant, M, such that $E\left(\frac{1}{J}\sum_{j=1}^{J} ||\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \boldsymbol{f}_t \omega_{jt}||^2\right) \leq M$.

Then, the following theorem ensures that the factors $\hat{\boldsymbol{f}}_t$ estimated by PCA, as well as the dimension $\hat{r}$ selected by the panel information criteria, are consistent.

Theorem 1 (Bai and Ng (2002)). *Under assumptions (i)-(iv) above, for each t, there exists a matrix $\boldsymbol{H}$ of rank $\min\{r, \hat{r}\}$, such that as $J, T \to \infty$, we have*

$$\min\{J, T\} \cdot \left(\frac{1}{T}\sum_{t=1}^{T} ||\hat{\boldsymbol{f}}_t - \boldsymbol{H}^{'}\boldsymbol{f}_t||^2\right) = O_p(1), \tag{6}$$

and

$$\lim_{J,T\to\infty} Pr[\hat{r} = r] = 1, \tag{7}$$

where

$$\hat{r} = arg\min_{k} IC(k) = arg\min_{k}\left(\log V(k, \hat{\boldsymbol{f}}_t) + kg(J, T)\right), \tag{8}$$

with $V(k, \hat{\boldsymbol{f}}_t) = \frac{1}{T}\sum_{t=1}^{T}(\widetilde{\boldsymbol{X}}_t - \hat{\boldsymbol{\Lambda}}^{(k)}\hat{\boldsymbol{f}}_t^{(k)})'(\widetilde{\boldsymbol{X}}_t - \hat{\boldsymbol{\Lambda}}^{(k)}\hat{\boldsymbol{f}}_t^{(k)})$. $\hat{\boldsymbol{f}}_t^{(k)}$ and $\hat{\boldsymbol{\Lambda}}^{(k)}$ are the factors and the loading matrix estimated by PCA with dimension k. Finally, $g(J, T)$ is a penalty function satisfying $g(N, T) \to 0$ and $\min\{J, T\} \cdot g(J, T) \to \infty$ as $J, T \to \infty$.

Proof. See Appendix A. □

The above theorem highlights the following results. First, when the estimated factors $\hat{\boldsymbol{f}}_t$ are obtained by PCA, the optimal dimension $\hat{r}$ selected by the information criteria defined by Equation (8) is a consistent estimator of the true dimension r. Second, since the true factors $\boldsymbol{f}_t$ can only be identified up to scale, a rotation of the factors, $\boldsymbol{H}^{'}\boldsymbol{f}_t$, is considered. More importantly, the squared deviations between the estimated factors $\hat{\boldsymbol{f}}_t$ and those lie in the true factor space vanish as $J, T \to \infty$. In the empirical analysis of this paper, we use the panel information criterion $IC_{p1}(k)$ in Bai and Ng (2002), where $g(J, T)$ is defined as:[7]

$$g(J, T) = \left(\frac{J+T}{JT}\right)\log\left(\frac{JT}{J+T}\right). \tag{9}$$

[6] See assumptions A to D in Bai and Ng (2002) for more details.

[7] Using criteria $IC_{p2}(k)$ and $IC_{p3}(k)$ in Bai and Ng (2002) lead to qualitatively similar results.

After estimating the common factors, $\hat{\boldsymbol{f}}_t$, we turn to determine the optimal subset $\hat{\boldsymbol{F}}_t \subseteq \hat{\boldsymbol{f}}_t$ for predicting the crop yields. To this end, different algorithms can be employed. For example, Ludvigson and Ng (2009) form candidates of $\hat{\boldsymbol{F}}_t$ as subsets of $\hat{\boldsymbol{f}}_t$, and/or their functional forms, $g(\hat{\boldsymbol{f}}_t)$, regress the dependent variable on $\hat{\boldsymbol{F}}_t$, and select the optimal set of factors based on Bayesian information criterion (BIC). In this paper, we follow the optimal factor selection procedure advocated in Stock and Watson (2002) and Bai and Ng (2008). Formally, candidates of $\hat{\boldsymbol{F}}_t$ are constructed from all possible subsets of $\hat{\boldsymbol{f}}_t$ (a total of $2^{\hat{r}}$), the dependent variable $\log(y_t)$ is then regressed on each $\hat{\boldsymbol{F}}_t$ (and the time trend bt if $\log(y_t)$ is non-stationary), and the $\hat{\boldsymbol{F}}_t$ that gives the minimal BIC is selected. More specifically, the procedure is outlined below.

Step 1. For $k = 1, \ldots, 2^{\hat{r}}$, perform the following three sub-steps:

1a. Construct a $\tilde{r}$-dimensional candidate $\hat{\boldsymbol{F}}_k = (\hat{\boldsymbol{F}}'_{k,1}, ...\hat{\boldsymbol{F}}'_{k,T})'$ as a subset of $\hat{\boldsymbol{f}}$, where $1 \leq \tilde{r} \leq \hat{r}$,

1b. Denote by $\vec{\mathbf{F}}_k = (\mathbf{1}', \boldsymbol{t}', \hat{\boldsymbol{F}}_k)$ the $(\tilde{r}+2)\times T$ matrix of data, and $\boldsymbol{\Theta} = (a, b, \boldsymbol{\theta})$ the $(\tilde{r}+2)$-dimensional vector of coefficients, estimate the following regression:

$$\log(y_t) = \boldsymbol{\Theta}'\vec{\mathbf{F}}_{k,t} + \epsilon_t, \tag{10}$$

where $\vec{\mathbf{F}}_{k,t}$ is the t-th column of $\vec{\mathbf{F}}_k$,

1c. Compute BIC corresponds to the parameter values estimated from Step 1b and denote its value by BIC_k.

Step 2. The matrix of optimal factor is given by $\hat{\boldsymbol{F}}_{k^*}$ where $k^* \leftarrow \arg\min_k BIC_k$.

Note that by repeating Step 1 for $k = 1, \ldots, 2^{\hat{r}}$, the candidate factor $\hat{\boldsymbol{F}}_k$ identified in Step 1a has dimension ranging from 1 to $\hat{r}$. This implies that when estimating Equation (10), the dimension of the coefficient vector $\boldsymbol{\Theta}$ also changes accordingly so as it is consistent with $\hat{\boldsymbol{F}}_k$. Recall $\hat{\boldsymbol{f}}_t = \hat{\boldsymbol{\Lambda}}'\widetilde{\boldsymbol{X}}_t$ for each t. Then analogously we have $\hat{\boldsymbol{F}}_{k^*,t} = \hat{\boldsymbol{\Lambda}}'_*\widetilde{\boldsymbol{X}}_t$. Here $\hat{\boldsymbol{\Lambda}}_*$ is the $J \times r^*$ matrix with its columns being abstracted from the respective columns of $\boldsymbol{\Lambda}$ and r^* being the dimension of the optimal factor $\hat{\boldsymbol{F}}_{k^*}$ obtained from Step 2. The time vector $\boldsymbol{t}$ should be taken out of $\vec{\mathbf{F}}_k$ in the above algorithm when $\log(y_t)$ is stationary.

We end this subsection by reiterating the difference between our proposed dynamic factor model with principal component regression. In our proposed approach, the optimal dynamic factors are determined following two key model selection steps. First, a large set of dynamic factors are estimated as the first $\hat{r}$ principal components, with $\hat{r}$ determined by the IC_{p_1} criteria in Bai and Ng (2002). This step ensures the consistency of both the dimension $\hat{r}$ and the factors. Second, optimal factors are selected according to the Bayesian Information Criterion.

2.3 Approximating the original relationship

After obtaining the optimal factor $\hat{\boldsymbol{F}}_{k^*}$, the crop yield forecasting regression model becomes

$$\log(y_t) = \boldsymbol{\Theta}'\vec{\mathbf{F}}_{k^*,t} + \epsilon_t, \tag{11}$$

with $\vec{\mathbf{F}} = (\mathbf{1}, \boldsymbol{t}, \hat{\boldsymbol{F}}_{k^*})$. In particular, Bai and Ng (2006) showed that, under the assumption that $J, T \to \infty$ and $\sqrt{T}/J \to 0$, the least squares estimators of $\boldsymbol{\Theta}$ are consistent and asymptotically normal. As a result, not only the point estimates, but also the asymptotic variance can be evaluated as though $\hat{\boldsymbol{F}}_{k^*,t}$ is observed. Empirical analysis assuming the estimated factors as observable is conducted in the existing studies, including Ludvigson and Ng (2009) and Bai and Ng (2008). In particular, assume that $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$, we have the following least squares estimation:

$$\begin{aligned} \hat{\boldsymbol{\Theta}} &= (\vec{\mathbf{F}}\vec{\mathbf{F}}')^{-1}\vec{\mathbf{F}}\left(\log(\boldsymbol{y})\right), \\ \mathrm{Var}(\hat{\boldsymbol{\Theta}}) &= \sigma^2(\vec{\mathbf{F}}\vec{\mathbf{F}}')^{-1}. \end{aligned}$$

Moreover, based on the assumption that $\hat{\boldsymbol{F}}_{k^*,t} = \hat{\boldsymbol{\Lambda}}_*'\widetilde{\boldsymbol{X}}_t$ is treated as observed, we can proceed to assess the importance of $\boldsymbol{X}_t$ for predicting crop yields via $\hat{\boldsymbol{F}}_{k^*,t}$. This type of analysis is useful to better understand the relationship between the original explanatory variables and the crop yields, for example, when designing an IBI policy. Recall the time-series model specification (Model (1) and Model (3)),

$$\log(y_t) = a + bt + \boldsymbol{\gamma}'\boldsymbol{X}_t + \varepsilon_t, \tag{12}$$

with $\varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$.[8] Therefore, replacing $\hat{\boldsymbol{F}}_{k^*,t}$ by $\hat{\boldsymbol{\Lambda}}_*'\widetilde{\boldsymbol{X}}_t$ in (11) gives

$$\begin{aligned} \log(y_t) &= a + bt + \boldsymbol{\theta}'\hat{\boldsymbol{\Lambda}}_*'\widetilde{\boldsymbol{X}}_t + \epsilon_t \\ &= (a - \boldsymbol{\theta}'\hat{\boldsymbol{\Lambda}}_*'\bar{\boldsymbol{X}}) + bt + \boldsymbol{\theta}'\hat{\boldsymbol{\Lambda}}_*'(\widetilde{\boldsymbol{X}}_t + \bar{\boldsymbol{X}}) + \epsilon_t \\ &= \tilde{a} + bt + \boldsymbol{\theta}'\hat{\boldsymbol{\Lambda}}_*'\boldsymbol{X}_t + \epsilon_t, \end{aligned} \tag{13}$$

where $\bar{\boldsymbol{X}}$ is the mean vector of the original data matrix $\boldsymbol{X}$. Hence, the estimator of $\boldsymbol{\gamma}$ and its variance, $\hat{\boldsymbol{\gamma}}$ and $\mathrm{Var}(\hat{\boldsymbol{\gamma}})$, can be approximated by $\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\Lambda}}_*$ and $\hat{\boldsymbol{\Lambda}}_*\mathrm{Var}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\Lambda}}_*'$, respectively.

When estimating Equation (13), it is assumed that the matrix $\hat{\boldsymbol{\Lambda}}_*$ is deterministic. The essence of the dynamic factor model is to approximate the real parameter, $\boldsymbol{\gamma}$, by the alternative parameter, $\boldsymbol{\theta}'\hat{\boldsymbol{\Lambda}}_*$. Recall that the real parameter $\boldsymbol{\gamma}$ is hardly (when $J \approx T$) or even not (when $J \geq T$) identifiable by classical approaches given the available sample. The latter parameter is constructed under a set of data-driven linear constraints described by the matrix $\hat{\boldsymbol{\Lambda}}_*$, which can be estimated more efficiently. We remark that such approximations of the original relationship are not available with the LASSO regression when $J > T - 2$, as the LASSO allows at most T non-zeros coefficients in Equation (12).

[8] Note that ε_t in Equation (12) and ϵ_t in Equation (11) can be different sequences of iid error terms.

3 Empirical results

In this section, the factor model introduced in Section 2 is applied to historical data from the State of Illinois, U.S. Subsection 3.1 describes the dataset. Subsection 3.2 introduces the classic regression models for benchmarking purposes given their popularity in crop yield forecasting literature. The model fitness and the out-of-sample forecasting performance are discussed in Subsection 3.3 and Subsection 3.4, respectively.

3.1 Data

A dataset consisting of county-level and state-level annual crop production experience from the Illinois State (IL), U.S., is utilized for the empirical analysis. The data was a crops survey data over the period from 1981 to 2016 and obtained from the National Agricultural Statistics Service. Three crop types are considered: corn, soybean, and winter wheat. These three crops together comprise more than 80% of the cropland acreage in IL. Figure 1 shows the number of acres harvested for each crop type as a percentage of total harvested cropland acreage. It can be observed that the northern part of IL mainly produces corn, while soybeans and winter wheat are mainly grown in the southern part of the state.

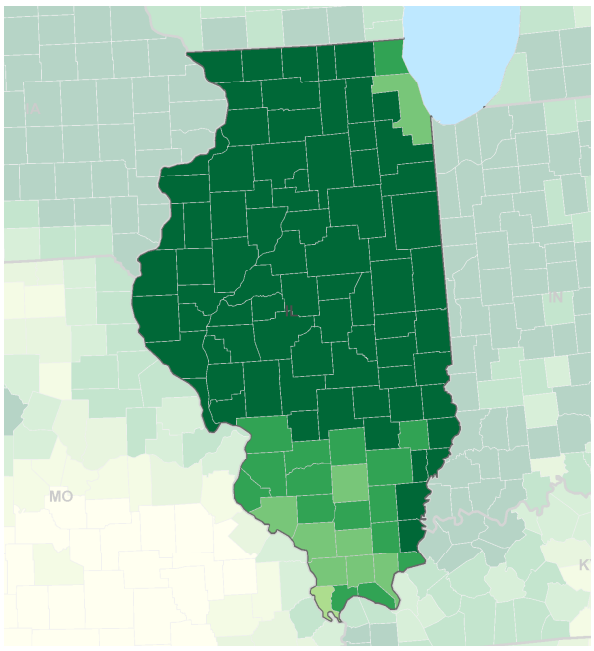
In addition to crop yield data, meteorological variables and soil information are included in the analysis. Weather data, including monthly average temperature and monthly cumulative precipitation, is collected from the PRISM Climate Group. PRISM publishes daily temperature and precipitation data for the conterminous U.S. since 1981 and at a resolution of 4km. Finally, soil information is obtained from the United States of Department of Agriculture (USDA) Natural Resources Conservation Service. The soil dataset includes 56 time-invariant variables representing the soil conditions of IL.[9]

While we have weather data for the entire year, it is customary to only consider the growing season weather variables in the forecasting models in view of their critical effect on the crop production. Table 1 lists the growing seasons for each of the crop types based on USDA (1997).

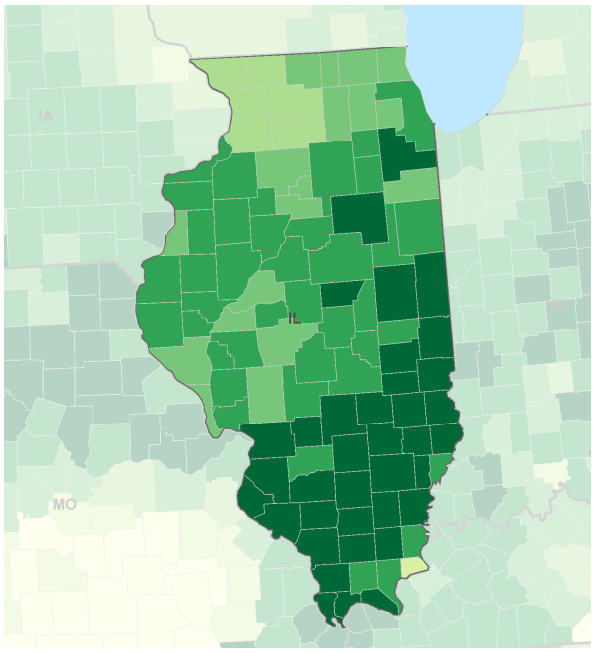
Table 1: The growing season of different crops in Illinois State.

Crop	**Growing Season**
Corn	May - August
Soybeans	May - August
Winter Wheat	October - June (the following year)

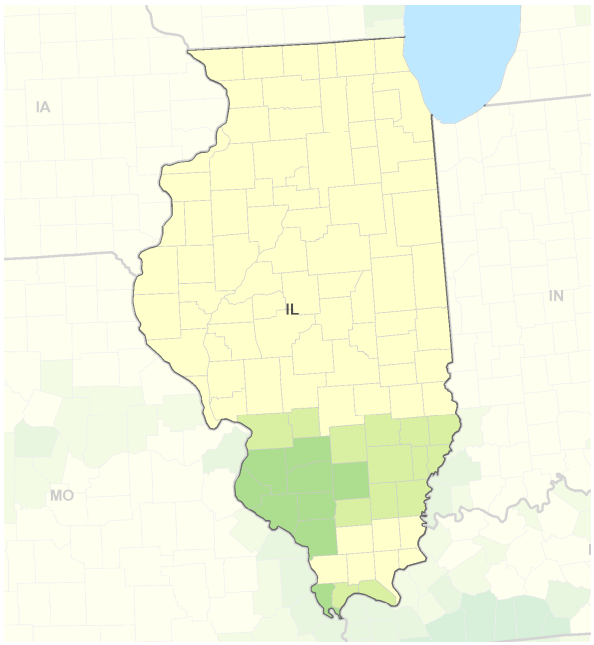
[9]The data sets of the empirical analysis of this paper are all publicly available in Ag-Analytics (`https://ag-analytics.org`). See Woodard (2016a,c,b).



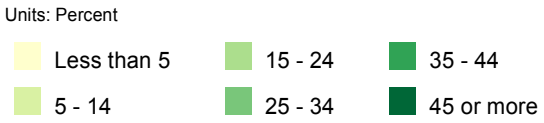
(a) Corn



(b) Soybeans



(c) Spring Wheat



(d) Legend

Figure 1: Acres of harvested crops as percent of harvested cropland acreage (Statistics of 2012). Data sources: USDA National Agricultural Statistics Service, ESRI.

3.2 The Regression Model

A standardized approach to assess Model (1) (Model (2)) is to first specify the explanatory variables in design matrix $\boldsymbol{X}_{i,t}$ ($\boldsymbol{X}_{i,avg}$), and then estimate the models by the least square method. Researches indicate that meteorological variables, such as growing season precipitation and temperature, as well as soil condition variables, are important for predicting crop yields (Kucharik and Serbin, 2008; Lobell and Field, 2007; Lobell and Burke, 2009, 2010; Lobell et al., 2011; Mobarak and Rosenzweig, 2012; Schlenker and Roberts, 106; Tannura et al., 2008; Tao et al., 2008). A widely accepted specification is discussed in Lobell and Burke (2010), which includes growing season average temperature and growing season total precipitation in the time-series specification, as well as their corresponding square terms in the cross-section specification. For ease of referencing, we denote models with these specifications as the *regression models*. These models also serve the benchmark to empirically assess our proposed dynamic factor models. The specifications of the benchmark regression models (1) and (2) are, respectively, given by

(a) Time-series specification:

$$\boldsymbol{X}_{i,t} = (t, T_{i,t}, P_{i,t})', \tag{14}$$

(b) Cross-section specification:

$$\boldsymbol{X}_{i,avg} = (T_{i,avg}, P_{i,avg}, T^2_{i,avg}, P^2_{i,avg})', \tag{15}$$

where t is the year variable, $T_{i,t}$ and $P_{i,t}$ represent the average temperature and total precipitation over the growing season in year t; $T_{i,avg}$ and $P_{i,avg}$ are the average of $T_{i,t}$ and $P_{i,t}$, respectively, over the sample period; and $T^2_{i,avg}$ and $P^2_{i,avg}$ are their corresponding square terms. In the following subsection, an empirical studies is conducted to compare our proposed dynamic models to these benchmark models.

3.3 Empirical Analysis: Model fitness

The advantages of the regression models (14) and (15) lie in their simplicity and parsimony. They contain respectively only three and four regressors with clear interpretation. However, their simplicity and parsimony may cast doubt on their prediction capability. In particular, existing research has suggested that monthly precipitation and temperature (and their squared terms) may have different impacts on crop yield (Tannura et al., 2008). Moreover, soil quality information has also found to be important to explain long-term crop yield variations (Lobell and Burke, 2010; Woodard and Verteramo-Chiu, 2017).

Motivated by the above limitations of the regression models, we consider models taking into account not only the monthly temperature and precipitation in the growing season, their square terms, their interaction terms, but also soil quality information. More specifically, the final design matrix of our proposed dynamic models has in total 116 variables. A complete list of the meteorological variables and soil quality variables is reported in Appendix B (Tables 9 and 10, respectively). The dynamic factor approach is applied to the

design matrix. We refer to model based on the optimally selected factor $\hat{\boldsymbol{F}}_{k^*,t}$ and (11) as the "Optimal Factor" model. In our empirical analysis, three specifications are implemented, namely the cross-section specification at the county level and the time-series specifications at both the county and the state level. The adjusted R^2 obtained from fitting the above three specifications of dynamic optimal factor models are reported in Table 2, along with their optimal r^*. The county-level time-series models are fitted to all counties and hence the reported values should be interpreted as the averages across all counties, except for the values with labels "Min. r^*" and "Max. r^*", which correspond to the minimum and maximum values of r^*.

To evaluate the effectiveness of the proposed factor models, two other types of models are also implemented. The first type is the regression models (14) and (15). The second type is the dynamic factor models with the number of factors constrained to 2 for the time-series specification and 4 for the cross-section specification. These values are chosen so that the factor models have the same number of regressors as the respective regression models. For this reason, the latter implementation of the dynamic factor model is denoted as the "Constrained Factor" model. Their (average) adjusted R^2 are also reported in Table 2.

Table 2: Comparison of adjusted R^2 among the regression models and the dynamic factor models (with county- and state-level time-series specification and county-level cross-section specifications). For the dynamic optimal factor models, r^* is also reported. The r^* values under the county-level time-series specification are averages, and the values with labels "Min. r^*" and "Max. r^*" correspond to the minimum and maximum values of r^* across all counties.

	(Average) Adjusted R^2					
	Regression	**Constrained Factor**	**Optimal Factor**	**Min. r^***	**r^***	**Max. r^***
	County-level Time-Series Model					
Corn	51%	63%	85%	2	8.83	18
Soybean	56%	65%	76%	2	4.72	13
Winter Wheat	32%	46%	53%	2	3.41	11
	State-level Time-Series Model					
Corn	68%	79%	94%	-	12	-
Soybean	71%	79%	93%	-	13	-
Winter Wheat	61%	69%	75%	-	6	-
	County-level Cross-Section Model					
Corn	58%	78%	86%	-	10	-
Soybean	67%	81%	89%	-	8	-
Winter Wheat	33%	85%	99%	-	10	-

An immediate conclusion from these results is that the dynamic factor models (regardless of "optimal" or "constrained") outperform the regression models based on the (average) adjusted R^2. Specifically, the (average) adjusted R^2 from the dynamic constrained factor models are approximately 10% higher than the respective regression models with the same number of factors. It is also reassuring that it is rewarding to optimally determine the factor $\hat{\boldsymbol{F}}_{k^*,t}$. The (average) adjusted R^2 from the dynamic optimal factor models are consistently higher than the constrained ones. This is not surprising since the latter can be considered as sub-optimal implementations of the dynamic optimal factor models. With the exception of winter wheat for the county-level time-series specification, the corresponding (average) adjusted R^2 is at least 75%, and reaches as high as 99%. To further appreciate the efficiency of the dynamic models, let us consider the county-level corn data. For this particular crop, the average adjusted R^2 for the regression model, the dynamic constrained factor model, and the dynamic optimal factor model are 51%, 63%, and 85%, respectively. These translate into an increment of about 23% from the regression model to the dynamic constrained factor model and another 35% to the optimal one. To conclude the discussion on the results reported in Table 2, we point out that the optimal values of r^* for the county-level time-series models are 9, 5, 4 for corn, soybean, and winter wheat respectively. In general, the optimal values of r^* are higher for the state-level specifications.

The original regression

Recall that the point estimates and the standard deviations of the coefficients in the original regression can be approximated according to Subsection 2.3. For illustration, Table 3 shows the implied point estimates and the corresponding t-statistics for the original relationship for the state-level corn time-series specification. From Table 3, we draw the following remarks:

- The time trend is positive and significant, indicating a substantial and persistent impact on crop yield;
- Many of the considered meteorological variables are found to have a significant impact on crop yield, and the impact of each meteorological variable can be interpreted. For example, one unit increase in precipitation in August precipitation will lead to $(-2.798 + 0.12 \times \text{August Temperature})$ percentage change in corn yield. That is, given the sample average August temperature of $24.26°C$, one unit increase in precipitation in August would on average lead to 0.11 percentage increase in corn yield, *ceteris paribus*.

The estimation results obtained by directly estimating the regression model is shown in Table 4 for comparison. The point estimates are, in general, similar to those implied by the dynamic factor model in Table 3, especially for coefficients that are significantly different from zero. More importantly, directly estimating the original regression provides smaller t-statistics, i.e., larger standard deviations, of the estimators. This is not surprising, since compared to the dynamic factor model, direct regression has a smaller degree-of-freedom.

Table 3: Approximating the original relationship. Point estimates for the coefficients in the original regression, implied from the dynamic factor model, are displayed, with the corresponding t-statistics in parentheses. This table shows the results for state-level time-series specification for corn as an illustration. "*" and "**" represent estimated coefficients that are significant at 5% and 1% levels, respectively. Coefficient significance levels are calculated using 22 degrees of freedom.

Regressor			**Temperature**		
			May	-0.158**	(-4.01)
Intercept	4.668**	(203.67)	**June**	-0.667*	(-2.15)
Time trend	0.015**	(13.36)	**July**	0.220**	(2.98)
Pre-season	0.001	(1.24)	**August**	-0.537**	(-2.52)
Precipitation ($\times 10^{-2}$)			**Squared Temperature** ($\times 10^{-2}$)		
May	-0.078	(-0.72)	**May**	0.400**	(3.67)
June	0.548	(0.92)	**June**	1.531*	(2.05)
July	0.632	(0.91)	**July**	-0.518**	(-3.02)
August	-2.798**	(-3.68)	**August**	0.849*	(2.04)
Squared Precipitation ($\times 10^{-4}$)			**Precipitation×Temperature** ($\times 10^{-2}$)		
May	-0.080*	(-1.93)	**May**	0.013**	(2.78)
June	-0.156**	(-3.74)	**June**	-0.005	(-0.23)
July	-0.352**	(-3.70)	**July**	0.008	(0.35)
August	0.040	(0.72)	**August**	0.120**	(3.98)

Table 4: The original relationship (estimated directly). Point estimates for the coefficients in the original regression are displayed with the corresponding t-statistics in parentheses. This table shows the results for state-level time-series specification for corn as an illustration. "*" and "**" represent estimated coefficients that are significant at 5% and 1% levels, respectively. Coefficient significance levels are calculated using 12 degrees of freedom.

Regressor			**Temperature**		
			May	-0.153	(-0.58)
Intercept	4.668**	(113.32)	**June**	-0.887*	(-1.90)
Time trend	0.014**	(6.7)	**July**	0.310	(0.69)
Pre-season	0.000	(0.27)	**August**	-0.373	(-1.14)
Precipitation ($\times 10^{-2}$)			**Squared Temperature** ($\times 10^{-2}$)		
May	-0.517	(-0.75)	**May**	0.348	(0.49)
June	0.029	(0.03)	**June**	1.970*	(1.82)
July	1.048	(0.63)	**July**	-0.684	(-0.78)
August	-2.809*	(-2.42)	**August**	0.512	(0.75)
Squared Precipitation ($\times 10^{-4}$)			**Precipitation×Temperature** ($\times 10^{-2}$)		
May	0.000	(-0.22)	**May**	0.031	(0.91)
June	-0.001*	(-2.12)	**June**	0.017	(0.47)
July	-0.004*	(-2.36)	**July**	-0.004	(-0.07)
August	0.001	(0.93)	**August**	0.113*	(2.37)

For example, the dynamic model in Table 3 is estimated with 12 factors (see first row of the second panel in Table 2), and hence the degree-of-freedom is 22 (calculated as $35 - 12 - 1$). In contrast, in the direct regression model in Table 4, there are 22 regressors, including 21 meteorological variables and a time trend, and therefore, the degree-of-freedom of this model is 12 (calculated as $35 - 22 - 1$). Therefore, the dynamic factor model is more likely to identify variables that have a significant impact on crop yield. Moreover, recall that in the cross-section specification, there are more explanatory variables than the number of observations, and direct regression is no longer possible in this case. The dynamic factor model, in contrast, is still feasible here, and the original relationship, such as the one shown in Table 3, can accordingly be obtained.

3.4 Cross-validation

In order to test the out-of-sample forecasting ability of the proposed dynamic factor approach, leave-one-out cross validation is also conducted, and the relative root mean square error (RMSE) is calculated. The cross-validation process can be summarized in the following steps:

1. Remove one year at a time;
2. Estimate each model with the remaining years;
3. Predict the removed year yield;
4. Repeat Steps 1-3 for all years, and calculate the RMSE as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\log y_t - \widehat{\log y_t}}{\log y_t}\right)^2}, \tag{16}$$

where $\widehat{\log y_t}$ is the estimated yield (in logarithmic scale) that is used to predict the removed yield $\log y_t$ (in logarithmic scale).

Table 5 shows the cross-validation RMSEs under each model specification for each crop. The dynamic factor approach provides satisfactory cross-validation results. The optimal factor approach from Algorithm **??** on average reduces cross-validation RMSE by 50%. The constrained factor method also shows improvements, with a reduction in the cross-validation RMSE of 20% on average.

Table 5: Cross-validation results. This table displays the RMSEs for each model specification, including state and county level time-series specification and the cross-section specification, for each crop.

	Regression	Constrained Factor	Dynamic Factor
State level			
Corn	2.5%	2.2%	0.9%
Soybean	2.3%	2.1%	0.9%
Winter Wheat	3.1%	2.8%	2.3%
County level			
Corn	4.4%	3.8%	1.9%
Soybean	3.5%	3.0%	2.2%
Winter Wheat	4.5%	3.9%	3.6%
Cross-section			
Corn	1.6%	1.1%	0.9%
Soybean	1.9%	1.4%	1.1%
Winter Wheat	1.4%	0.7%	0.01%

4 Application to Index-Based Insurance (IBI)

The purpose of this section is to demonstrate how the dynamic factor approach developed in this paper can be applied in the design of an IBI policy. As pointed out earlier, a main challenge in the success of index insurance is basis risk, which refers to a mismatch between the index that determines the indemnities and the actual loss experienced on the farm. Subsection 4.1 discusses the construction of IBI based on the proposed dynamic factor model relative to the benchmark regression model. Subsection 4.2 compares and contrasts the efficiency of these IBIs with the indemnity-based insurance program.

4.1 Index Insurance Design

We begin this subsection by presenting a generic crop insurance contract given by

$$\mathcal{P} = Area \times Price \times \max(K - y, 0), \tag{17}$$

where $Area$ is the predetermined number of insured acres for each contract, $Price$ is the unit price of the insured crop determined at the beginning of the growing season, K is a predetermined constant and y is a random variable that relates in some ways to the yield of the insured crop. The payoff structure (17) resembles a put option with K interpret as the strike price.

Depending on the specification of the random variable y, the above crop insurance contract can be classified as either indemnity-based insurance or index-based insurance (IBI). The multiple peril crop insurance (MPCI) contract is an example of an indemnity based insurance policy. More precisely, if y represents the actual yield at the end of the growing season and $K = c \cdot \overline{Y}_t$ where c is the coverage level of the contract and $\overline{Y}_t$ is the expected crop yield in year t (i.e. the normal production level), then the payoff of a MPCI insurance contract can we expressed as (see Chantarat et al. (2013) and Woodard et al. (2016))

$$\mathcal{P}^M\left(\overline{Y}_t, c\right) = Area \times Price \times \max\left(c \cdot \overline{Y}_t - y_t, 0\right). \tag{18}$$

If y_t is an index that depends on some other underlying variables such as weather, soil, etc., then the resulting insurance contract (17) is an IBI. The effectiveness of such IBI depends on how well it mimics the payoff of the indemnity-based MPCI. If the discrepancy between these two payoffs is large, then the producers who insure their crop losses via IBI are exposed to large basis risk. Therefore, basis risk is a measure of how well the chosen index models the actual yield.

In this section, we evaluate the IBI constructed by the proposed dynamic factor model. In particular, an IBI can be constructed by setting $y = \hat{y}_t^{\text{Dynamic}}$, where $\hat{y}_t^{\text{Dynamic}}$ is the yield forecasted based on dynamic optimal factor model (11). With this design, then $\hat{y}_t^{\text{Dynamic}}$ becomes the index of the IBI. As a comparison, if y corresponds to the yield forecasted by the benchmark regression model, then it produces another IBI and we denote its projected yield as $\hat{y}_t^{\text{Regression}}$. To differentiate between these two types of IBI, we use $\mathcal{P}^I_{\text{Dynamic}}$ and $\mathcal{P}^I_{\text{Regression}}$ to denote the payoff of the IBI based on the two models, respectively. In the following subsection, the efficiencies of both IBIs are assessed against the benchmark indemnity-based MPCI contract $\mathcal{P}^M$.

4.2 Basis Risk Analysis

The underlying model of the IBI policy designed in the previous subsection is based on the contract parameters estimated from the optimally selected forecasting model. Therefore, it is of interest to analyze the resulting basis risk from this approach to understand the potential improvements. As previously discussed, basis risk arises from the difference between the hedging instrument and the actual loss incurred on the farm. Therefore, basis risk generally induces two types of mismatches, including the scenario where the insurer fails to pay the producer when an actual loss occurs, and the situation where the producer receives a payment when in fact no loss occurs. Basis risk can be more formally decomposed into the following two types of error. Given a null hypothesis that a producer experiences a loss:

Type I Error: the incorrect rejection of a true null hypothesis, where the producer does not receive an indemnity despite a loss incurred on the farm.

TYPE II ERROR: the failure to reject a false null hypothesis, where the producer receives an indemnity despite no loss on the farm.

Note that both errors are expressed in terms of probabilities. Figure 2 provides a graphical visualization of the two errors. In particular, τ_1 and τ_2 are deductibles of the indemnity-based and index-based contract, respectively. The dot-shaded area refers to the Type I error and the line-shaded area refers to the Type II error. In particular, in the upper right area of the square, where neither Type I error nor Type II error occurs, an IBI policy may still overestimate or underestimate the actual loss, and the situation where the index perfectly matches the actual loss on the farm, is along the diagonal. On the other hand, in the lower left area of the square, where the yield loss does not excess τ_1 and index does not excess τ_2, i.e., payments based on both contracts are not triggered, is an un-triggered area.

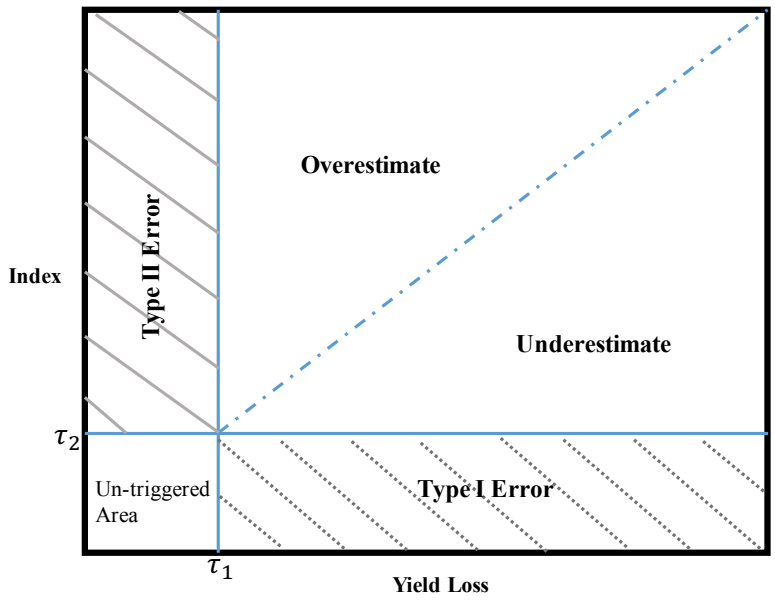


Figure 2: Illustration of basis risk.

For IBI policy, the payoff of the insurance index, $\mathcal{P}^I_{\text{Dynamic}}$ (or $\mathcal{P}^I_{\text{Regression}}$), is compared with the indemnity-based MPCI contract $\mathcal{P}^M$ given by (18). The basis risk of the proposed IBI policy is tested through a backtesting procedure, where in-sample county- and state-level historical data is used to calibrate the model, calculate the payoff based on the designed index, $\mathcal{P}^I_{\text{Dynamic}}$ (or $\mathcal{P}^I_{\text{Regression}}$), and compare to the actual losses based on the indemnity-based insurance, $\mathcal{P}^M$. The comparison procedure is described in greater details below. For $t = 2001, 2002, \ldots 2016$, we

(1) use in-sample historical data to calibrate the forecasting model, $\hat{y}_s = \hat{y}_s^{\text{Dynamic}}$ (or $\hat{y}_s^{\text{Regression}}$), where $s = 1981, 1982, \ldots, t$;

(2) calculate the payoff of the index, $\mathcal{P}^I_{\text{Dynamic}}$ (or $\mathcal{P}^I_{\text{Regression}}$), for year t, according to Equation (17);

(3) calculate the actual loss of the producer, the payoff of the indemnity-based insurance contract, $\mathcal{P}^M$, for year t, according to Equation (18);

(4) compare $\mathcal{P}^I_{\text{Dynamic}}$ (or $\mathcal{P}^I_{\text{Regression}}$) and $\mathcal{P}^M$ by calculating the Type I & Type II errors, as well as the mismatch in the cases of correctly triggering the payoff;

(5) update the information set by letting $t = t+1$, reestimate all parameters and repeat the steps outlined above.

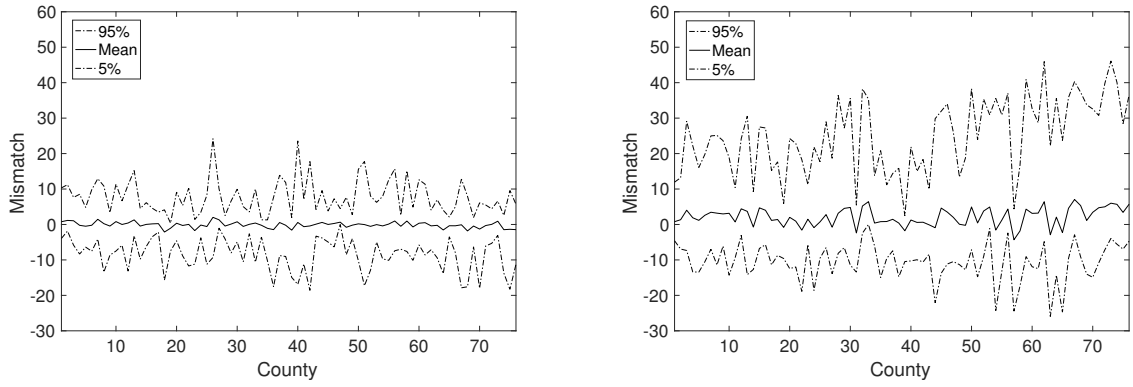
The analysis of basis risk is summarized in Table 6 for state-level data and in Table 7 for county-level data. The results are based on setting *Area* and *Price* as 1 in Equations (17) and (18) and coverage level c as 100% in Equation (18). Recall that Type I error corresponds to the cases when the producers fail to receive a payment despite crop losses are incurred. This situation is particularly worrisome as the producers should have been compensated for the incurred loss, given that they have already paid for the insurance coverage. If this occurs, the producers will perceive that the insurance has no value to them and eventually the demand on such IBI program will diminish. Tables 6 and 7 indicate that the Type I errors of the IBI constructed from $\mathcal{P}^I_{\text{Regression}}$ are quite large in most cases, while those from the $\mathcal{P}^I_{\text{Dynamic}}$ based IBI are much smaller. These observations support the design of an IBI based on the dynamic factor approach.

We now consider the Type II error. Recall that this error arises when producers receive insurance payments while there is no incurred losses. Paying benefits to producers in such situation would increase the costs to the insurer and hence raises concerns on the sustainability of the program. Results in Tables 6 and 7 confirm the superiority of the $\mathcal{P}^I_{\text{Dynamic}}$ based IBI. In all but one case the $\mathcal{P}^I_{\text{Dynamic}}$ based IBIs have smaller Type II errors. If we gauge the efficiency by comparing the "Total Error Prob.", which is the sum of both Type I and Type II error, the $\mathcal{P}^I_{\text{Dynamic}}$ based IBI again outperforms the $\mathcal{P}^I_{\text{Regression}}$ based IBI.

While comparing probabilities associated with Type I and Type II errors provides one way of assessing the relative efficiency of IBI from the payment frequency perspective, comparing the claim amounts from the IBI and the MPCI gives another way of gauging IBI from the loss severity perspective. In particular, small basis risk implies small discrepancy in indemnity amounts between these contracts. The averages of these discrepancies are reported in Tables 6 and 7 under the row labeled "Mismatch Mean". These values are calculated for cases where the respective IBI correctly identifies positive payout. Hence small absolute "Mismatch Mean" implies, on average, small basis risk. Again, we see that the dynamic factor model outperforms the regression model. Moreover, variability of the mismatch, as measured by its standard deviation (labeled as "Mismatch Std."), is also smaller for the $\mathcal{P}^I_{\text{Dynamic}}$ based IBI.

The efficiency of the $\mathcal{P}^I_{\text{Dynamic}}$ based IBI relative to the $\mathcal{P}^I_{\text{Regression}}$ based IBI is further highlighted in Figures 3 and 4. Figure 3 displays the mean, the 5%, and the 95% quantiles of the mismatches for IBI

constructed from $\mathcal{P}^I_{\text{Dynamic}}$ and $\mathcal{P}^I_{\text{Regression}}$. The results are computed for each county over the backtesting period and for corn.[10] Observe that the dynamic factor approach gives more stable and smaller mean payoff mismatches for each county (the mean mismatches are closer to zero, with smaller variability). Moreover, the 5% and 95% quantiles from the dynamic factor approach are much narrower than those from the regression model. Figure 4 shows the similar statistics but for each backtesting year across all counties. Similarly, we see that the dynamic factor model leads to smaller basis risk. These results support the design of IBI based on the dynamic factor model.



(a) Dynamic Factor Model

(b) Benchmark Regression

Figure 3: The mean, 5%, and 95% quantiles of payoff mismatches for IBI constructed from $\mathcal{P}^I_{\text{Dynamic}}$ and $\mathcal{P}^I_{\text{Regression}}$, relative to $\mathcal{P}^M$ (results for corn). The results are computed for each county over the backtesting period.

5 Conclusion

This paper investigates a new approach to forecasting yields using a dynamic factor model, and proposes a weather index-based insurance (IBI) policy based on the proposed forecasting model. The factor model starts with a rich set of explanatory variables, and achieve the first stage of dimension reduction using a much smaller number of latent common factors. At the second stage, an optimal subset of these factors for crop yield forecasting is selected based on information criteria. A dataset consisting of county-level and state-level annual crop production experience for three crops, corn, soybeans and wheat, from the state of Illinois in the U.S. is utilized for the empirical analysis. Empirical results show that the dynamic factor model developed in this paper is able to substantially improve crop yield forecasting relative to the classic regression specifications in the literature.

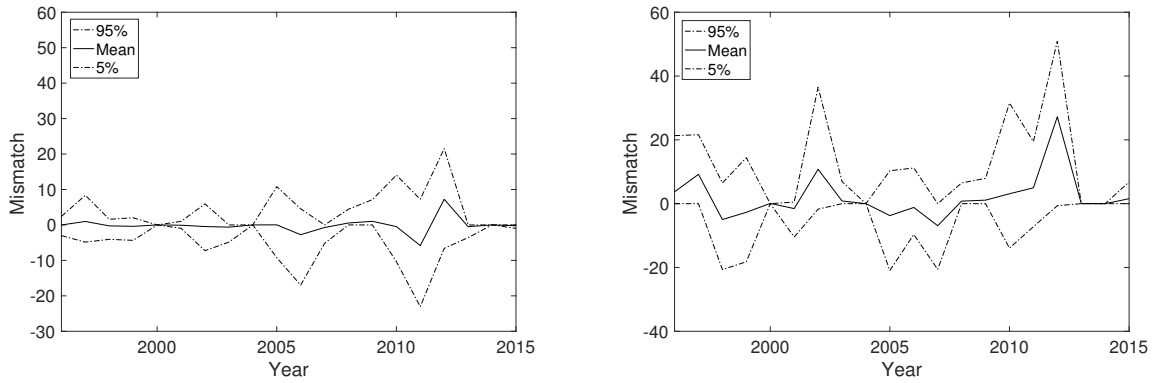
[10] Figures for soybeans and winter wheat are qualitatively similar, and are available upon request.

Table 6: Summary of basis risk analysis for state level. The first panel gives the summary statistics of the actual losses based on MPCI. The second panel shows summary of Type I & Type II errors, as well as the mismatch means and standard deviations.

	Corn		**Soybean**		**Winter Wheat**	
Summary of Actual Losses Based on MPCI						
Loss Prob.	28.57%		33.33%		28.57%	
Loss Mean	5.46		0.72		1.57	
Loss Std.	14.05		1.64		2.91	
Basis Risk Analysis						
	Dynamic Factor	**Regression**	**Dynamic Factor**	**Regression**	**Dynamic Factor**	**Regression**
Type I Error	0.00%	33.33%	14.29%	57.14%	0.00%	0.00%
Type II Error	0.00%	26.67%	14.29%	7.14%	26.67%	33.33%
Total Error Prob.	0.00%	60.00%	28.58%	64.28%	26.67%	33.33%
Mismatch Mean	0.33	0.63	0.00	0.35	-0.14	-0.22
Mismatch Std.	1.86	6.85	0.53	1.53	1.92	2.77

Table 7: Summary of basis risk analysis for county level. The first panel gives the summary statistics of the actual losses based on MPCI. The second panel shows summary of Type I & Type II errors, as well as the mismatch means and standard deviations for each county.

	Corn		Soybean		Winter Wheat	
Summary of Actual Losses Based on MPCI						
Loss Prob.	34.14%		33.60%		35.00%	
Loss Mean	7.41		1.45		2.23	
Loss Std.	16.96		2.95		4.30	
Basis Risk Analysis						
	Dynamic Factor	**Regression**	**Dynamic Factor**	**Regression**	**Dynamic Factor**	**Regression**
Type I Error	10.30%	28.52%	18.84%	48.95%	23.72%	31.22%
Type II Error	6.50%	20.91%	11.66%	14.30%	17.59%	41.40%
Total Error Prob.	16.8%	49.43%	30.50%	63.25%	41.31%	72.62%
Mismatch Mean	-0.12	2.12	0.77	0.77	0.33	0.88
Mismatch Std.	4.66	10.02	1.17	2.44	2.29	3.79



(a) Dynamic Factor Model

(b) Benchmark Regression

Figure 4: The mean, 5%, and 95% quantiles of payoff mismatches for IBI constructed from $\mathcal{P}^I_{\text{Dynamic}}$ and $\mathcal{P}^I_{\text{Regression}}$, relative to $\mathcal{P}^M$ (results for corn). The results are computed for each backtesting year across all counties.

Next, based on the proposed dynamic factor approach, a weather index-based insurance (IBI) policy is designed to improve producer risk management through reducing basis risk, which has historically been a major limitation in the more wide-spread use of index-based insurance solutions for agriculture. Type I and Type II errors were considered in order to assess the basis risk of the underlying IBI product. We show that the index constructed from the proposed factor approach matches more accurately the indemnity payments relative to the actual loss experienced on the farm, compare to the existing popular regression models.

There are several interesting directions for future research. First, it would be interesting to extend the set of explanatory variables to include more recent measurements relevant to crop growth, such as index representing global warming, as well as remote sensing data (Justice et al., 1998; Doraiswamy et al., 2003; White et al., 2011; Becker-Reshef et al., 2010; Turvey and Mclaurin, 2012). Second, other techniques for variable selection and dimension reduction, such as the LASSO, ridge regression, and elastic net, could be explored as alternatives to improve the performance of the forecasting model and reduce basis risk of the agriculture IBI design. Third, the current factor selection procedure involves evaluation of $2^{\hat{r}}$ sets of factors, and could be computationally intensive when $\hat{r}$ is large. A more efficient way to determine the optimal subset of factor for forecasting could be considered in future research.

References

Aït-Sahalia, Y. and D. Xiu (2019). Principal component analysis of high-frequency data. Journal of the American Statistical Association 114(525), 287–303.

Asseng, S., F. Ewert, C. Rosenzweig, J. Jones, J. Hatfield, A. Ruane, K. J. Boote, P. J. Thorburn, R. P. Rötter, D. Cammarano, et al. (2013). Uncertainty in simulating wheat yields under climate change. Nature Climate Change 3(9), 827–832.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. Econometrica 70(1), 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74(4), 1133–1150.

Bai, J. and S. Ng (2008). Forecasting economic time series using targeted predictors. Journal of Econometrics 146(2), 304–317.

Barnett, B. J. and O. Mahul (2007). Weather index insurance for agriculture and rural areas in lower-income countries. American Journal of Agriculture Economics 89(5), 1241–1247.

Basso, B., D. Cammarano, A. Troccoli, D. Chen, and J. T. Ritchie (2010). Long-term wheat response to nitrogen in a rainfed mediterranean environment: Field data and simulation analysis. European Journal of Agronomy 33(2), 132–138.

Becker-Reshef, I., E. Vermote, M. Lindeman, and C. Justice (2010). A generalized regression-based model for forecasting winter wheat yields in kansas and ukraine using modis data. Remote Sensing of Environment 114(6), 1312–1323.

Bondeau, A., P. C. Smith, S. Zaehle, S. Schaphoff, W. Lucht, W. Cramer, D. Gerten, H. LOTZE-CAMPEN, C. Müller, M. Reichstein, et al. (2007). Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biology 13(3), 679–706.

Bornn, L. and J. V. Zidek (2012). Efficient stabilization of crop yield prediction in the canadian prairies. Agricultural and forest meteorology 152, 223–232.

Chantarat, S., A. G. Mude, C. B. Barrett, and M. R. Carter (2013). Designing index-based livestock insurance for managing asset risk in northern Kenya. Journal of Risk and Insurance 80(1), 205–237.

Cole, S., X. Giné, J. Tobacman, P. Topalova, R. Townsend, and J. Vickery (2013). Barriers to household risk management: Evidence from india. American Economic Journal: Applied Economics 5(1), 104–135.

Conradt, S., R. Finger, and M. Spörri (2015). Flexible weather index-based insurance design. Climate Risk Management 10, 106–117.

Cummins, J. D., D. Lalonde, and R. D. Phillips (2004). The basis risk of catastrophic-loss index securities. Journal of Financial Economics 71, 77–111.

Deng, X., B. J. Barnett, and D. V. Vedenov (2007). Is there a viable market for area-based crop insurance? American Journal of Agricultural Economics 89(2), 508–519.

Doraiswamy, P. C., S. Moulin, P. W. Cook, and A. Stern (2003). Crop yield assessment from remote sensing. Photogrammetric Engineering & Remote Sensing 69(6), 665–674.

Eberhardt, M., C. Helmers, and H. Strauss (2013). Do spillovers matter when estimating private returns to r&d? Review of Economics and Statistics 95(2), 436–448.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. The Annals of statistics 32(2), 407–499.

Fan, J., F. Han, and H. Liu (2014). Challenges of big data analysis. National science review 1(2), 293–314.

FAO (2009). Global agriculture towards 2050. Available at www.fao.org/leadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdf. Accessed: May 2018.

French, D. and C. O'Hare (2013). A dynamic factor approach to mortality modeling. Journal of Forecasting 32(7), 587–599.

Goodwin, B. K. (1993). An empirical analysis of the demand for multiple peril crop insurance. American Journal of Agricultural Economics 75(2), 425–434.

Goodwin, B. K. and V. H. Smith (1995). The economics of crop insurance and disaster aid. American Enterprise Institute.

Hainaut, D. (2019). Hedging of crop harvest with derivatives on temperature. Insurance: Mathematics and Economics 84(1), 98–114.

Hao, M., A. S. Macdonald, P. Tapadar, and R. G. Thomas (2018). Insurance loss coverage and demand elasticities. Insurance: Mathematics and Economics 79(1), 15–25.

Hazell, P., J. Anderson, N. Balzer, A. Hastrup Clemmensen, U. Hess, and F. Rispoli (2010). The potential for scale and sustainability in weather index insurance for agriculture and rural livelihoods. Technical Report 1-153, World Food Programme (WFP).

Hoogenboom, G., J. W. White, and C. D. Messina (2004). From genome to crop: integration through simulation modeling. Field Crops Research 90(1), 145–163.

Hyndman, R. J., H. Booth, and F. Yasmeen (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. Demography 50(1), 261–283.

Iturrioz, R. (2009). Agricultural insurance. Primer Series on Insurance. The World Bank. No. E20-77, 1–35.

Jensen, N. D., C. B. Barrett, and A. G. Mude (2016). Index insurance quality and basis risk: Evidence from northern kenya. American Journal of Agriculture Economics 98(5), 1450–1469.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd Edition ed.). John Wiley & Sons, Ltd.

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. American Economic Review 105(3), 1177–1216.

Justice, C. O., E. Vermote, J. R. Townshend, R. Defries, D. P. Roy, D. K. Hall, V. V. Salomonson, J. L. Privette, G. Riggs, A. Strahler, et al. (1998). The moderate resolution imaging spectroradiometer (modis): Land remote sensing for global change research. IEEE Transactions on Geoscience and Remote Sensing 36(4), 1228–1249.

Kucharik, C. J. and S. P. Serbin (2008). Impacts of recent climate change on Wisconsin corn and soybean yield trends. Environmental Research Letters 3(3), 1–10.

Lam, C., Q. Yao, et al. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. The Annals of Statistics 40(2), 694–726.

Lee, J.-P. and M.-T. Yu (2002). Pricing default-risky cat bonds with moral hazard and basis risk. Journal of Risk and Insurance 69(1), 25–44.

Lobell, D. B. and M. Burke (2009). Climate Change and Food Security: Adapting Agriculture to a Warmer World. Springer Science & Business Media.

Lobell, D. B. and M. B. Burke (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology 150(11), 1443–1452.

Lobell, D. B. and C. B. Field (2007). Global scale climate–crop yield relationships and the impacts of recent warming. Environmental research letters 2(1), 014002.

Lobell, D. B., W. Schlenker, and J. Costa-Roberts (2011). Climate trends and global crop production since 1980. Science 333(6042), 616–620.

Ludvigson, S. C. and S. Ng (2007). The empirical risk–return relation: A factor analysis approach. Journal of Financial Economics 83(1), 171–222.

Ludvigson, S. C. and S. Ng (2009). Macro factors in bond risk premia. Review of Financial Studies 22(12), 5027–5067.

Maestro, T., B. J. Barnett, K. H. Coble, A. Garrido, and M. Bielza (2016). Drought index insurance for the central valley project in california. Applied Economic Perspectives and Policy 38(3), 521–545.

Mahul, O. and J. Skees (2007). Managing Agricultural Risk at the Country Level: The Case of Index-Based Livestock Insurance in Mongolia, Volume 4325. Washington, DC: World Bank Publications.

Masshoff, O., M. Odening, and W. Xu (2011). Management of climate risks in agriculture - will weather derivatives permeate? Applied Economics 43(9), 1067–1077.

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. Journal of Business & Economic Statistics 34(4), 574–589.

Miranda, M. and J. Glauber (1997). Systemic Risk, Reinsurance and the Failure of Crop Insurance Market. American Journal of Agriculture Economics 79(1), 206–215.

Miranda, M. J. (1991). Area yield crop insurance reconsidered. American Journal of Agricultural Economics 73, 233–242.

Mobarak, A. M. and M. R. Rosenzweig (2012). Selling formal insurance to the informally insured. Yale Economics Department Working Paper No. 97; Yale University Economic Growth Center Discussion Paper No. 1007.

Ozaki, V. A., B. K. Goodwin, and R. Shirota (2008). Parametric and nonparametric statistical modelling of crop yield: implications for pricing crop insurance contracts. Applied Economics 40(9), 1151–1164.

Pesaran, M. H., L. V. Smith, and T. Yamagata (2013). Panel unit root tests in the presence of a multifactor error structure. Journal of Econometrics 175(2), 94–115.

Phélippé-Guinvarc'h, M. V. and J. E. Cordier (2006). A private management strategy for the crop yield insurer: A theoretical approach and tests. Insurance: Mathematics and Economics 39(1), 35–46.

Rowell, D. and L. B. Connelly (2012). A history of the term "moral hazard". Journal of Risk and Insurance 79(4), 1051–1075.

Schlenker, W. and M. J. Roberts (106). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. Proceedings of the National Academy of Sciences 37(15594-15598), 2009.

Sharif, B., D. Makowski, F. Plauborg, and J. E. Olesen (2017). Comparison of regression techniques to predict response of oilseed rape yield to variation in climatic conditions in denmark. European Journal of Agronomy 82, 11–20.

Shen, Z. and M. Odening (2013). Coping with systemic risk in index-based crop insurance. Agricultural Economics 44(1), 1–13.

Shi, W., F. Tao, and Z. Zhang (2013). A review on statistical models for identifying climate contributions to crop yields. Journal of Geographical Sciences 23(3), 567–576.

Skees, J. R., J. R. Black, and B. J. Barnett (1997). Designing and rating an area yield crop insurance contract. American Journal of Agriculture Economics 79(2), 430–438.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20(2), 147–162.

Tannura, M. A., S. H. Irwin, and D. L. Good (2008). Weather, technology, and corn and soybean yields in the us corn belt. Technical report, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.

Tao, F., M. Yokozawa, J. Liu, and Z. Zhang (2008). Climate–crop yield relationships at provincial scales in china and the impacts of recent climate trends. Climate Research 38(1), 83–94.

Thompson, L. M. (1969). Weather and technology in the production of corn in the us corn belt. Agronomy Journal 61(3), 453–456.

Thompson, L. M. (1988). Effects of changes in climate and weather variability on the yields of corn and soybeans. Journal of Production Agriculture 1(1), 20–27.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1), 267–288.

Turvey, C. G. and M. K. Mclaurin (2012). Applicability of the normalized difference vegetation index (ndvi) in index-based crop insurance design. Weather, Climate, and Society 4(4), 271–284.

USDA (1997). Usual planting and harvesting dates for U.S. field crops. In Agricultural Handbook. Agricultural Statistics Board, National Agricultural Statistics Service, United States Department of Agriculture.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. International journal of forecasting 30(4), 1030–1081.

White, J. W., G. Hoogenboom, P. W. Wilkens, P. W. Stackhouse, and J. M. Hoel (2011). Evaluation of satellite-based, modeled-derived daily solar radiation data for the continental united states. Agronomy journal 103(4), 1242–1251.

Woodard, J. D. (2016a). Ag-analytics.org data warehouse. Available at ag-analytics.org. Accessed: May, 2017.

Woodard, J. D. (2016b). Big data and ag-analytics: an open source, open data platform for agricultural & environmental finance, insurance, and risk. Agricultural Finance Review 76(1), 15–26.

Woodard, J. D. (2016c). Data science and management for large scale empirical applications in agricultural and applied economics research. Applied Economic Perspectives and Policy 38(3), 373–388.

Woodard, J. D. and P. Garcia (2008). Weather derivatives, spatial aggregation, and systemic risk: Implications for reinsurance hedging. Journal of Agricultural and Resource Economics 33(1), 34–51.

Woodard, J. D., A. Shee, and A. Mude (2016). A spatial econometric approach to designing and rating scalable index insurance in the presence of missing data. The Geneva Papers on Risk and Insurance - Issues and Practice 41(2), 259–279.

Woodard, J. D. and L. J. Verteramo-Chiu (2017). Efficiency impacts of utilizing soil data in the pricing of the federal crop insurance program. American Journal of Agriculture Economics 99(3), 757–772.

Zhang, J., K. S. Tan, and C. Weng (2018). Index insurance design. ASTIN Bulletin: The Journal of the IAA forthcoming, 1–33.

Zhu, J., Z. Ge, and Z. Song (2017). Distributed parallel pca for modeling and monitoring of large-scale plant-wide processes with big data. IEEE Transactions on Industrial Informatics 13(4), 1877–1885.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67(2), 301–320.

Appendix

A Proof of Theorem 1

The proofs of this theorem are based on Theorem 1, Theorem 2, and Corollary 1 in Bai and Ng (2002). For the sake of completeness, we will lay out the key steps of the theorem as follows. First, under assumptions (1) to (4), and for any $\hat{r} \geq 1$, Equation (6) follows Theorem 1 in Bai and Ng (2002). The inclusion of matrix $\boldsymbol{H}$ is because the true factors $\boldsymbol{f}_t$ can only be identified up to scale, and thus the rotation of $\boldsymbol{f}_t$ is considered. Equation (6) indicates that, for a given $\hat{r}$, the time average of the squared deviations between the estimated factors $\hat{\boldsymbol{f}}_t$ and the rotation $\boldsymbol{H}^{'}\boldsymbol{f}_t$ that lies in the true factor space vanishes as $J, T \to \infty$.

Second, Equation (7) and (8) follow Theorem 2 and Corollary 1 in Bai and Ng (2002). These two equations together guarantee that the optimal dimension $\hat{r}$ selected by the criteria $IC(k)$ is consistent. The conditions on the penalty function $g(J,T)$ is to ensure that it vanishes at an appropriate rate, such that under and overparameterized models will be ruled out.

B Data Description

This appendix contains a complete list of the meteorological variables and soil quality information included in the design matrix, including their short names and meanings. In particular, Table 9 lists the set of meteorological variables and Table 10 displays the definitions and names of soil quality data included in the design matrix for each crop. For Table 9, the symbol "O" and "N" means the corresponding variable takes value from the previous year or the current year, respectively. The pre-season is from September in the previous year to April in the current year for corn and soybean, and from July to September in the previous year for winter wheat.

Table 8: The set of meteorological variables included in the design matrix for each crop. "O" and "N" means the corresponding variable takes value from the previous year or the current year, respectively. The pre-season is from September in the previous year to April in the current year for corn and soybean, and from January to August in the previous year for winter wheat.

Variable	Corn	Soybean	Winter Wheat	Variable	Corn	Soybean	Winter Wheat
Pre. P	N	N	N				
Jan. P			O	July T	O	O	
Feb. P			O	Aug. T	O	O	
March P			O	Sep. T			
April P			O	Oct. T			N
May P	O	O	O	Nov. T			N
June P	O	O	O	Dec.T			N
July P	O	O		Jan. T^2			O
Aug. P	O	O		Feb. T^2			O
Sep. P				March T^2			O
Oct. P			N	April T^2			O
Nov. P			N	May T^2	O	O	O
Dec. P			N	June T^2	O	O	O
Jan. P^2			O	July T^2	O	O	
Feb. P^2			O	Aug. T^2	O	O	
March P^2			O	Sep. T^2			
April P^2			O	Oct. T^2			N
May P^2	O	O	O	Nov. T^2			N
June P^2	O	O	O	Dec. T^2			N
July P^2	O	O		Jan. $P \times T$			O
Aug. P^2	O	O		Feb. $P \times T$			O
Sep. P^2				March $P \times T$			O
Oct. P^2			N	April $P \times T$			O
Nov. P^2			N	May $P \times T$	O	O	O
Dec. P^2			N	June $P \times T$	O	O	O
Jan. T			O	July $P \times T$	O	O	
Feb. T			O	Aug. $P \times T$	O	O	
March T			O	Sep. $P \times T$			
April T			O	Oct. $P \times T$			N
May T	O	O	O	Nov. $P \times T$			N
June T	O	O	O	Dec. $P \times T$			N

Series Number	Variable Name	Description
1	aws0_5	Available water storage estimate (AWS) in standard layer 1 or standard zone 1 (0-5 cm depth).
2	aws5_20	AWS in standard layer 2 (5-20 cm depth).
3	aws20_50	AWS in standard layer 3 (20-50 cm depth).
4	aws50_100	AWS in standard layer 4 (50-100 cm depth).
5	aws100_150	AWS in standard layer 5 (100-150 cm depth)
6	aws150_999	AWS in standard layer 6 (150 cm to the reported depth of the soil profile).
7	aws0_20	AWS in standard zone 2 (0-20 cm depth).
8	aws0_30	AWS in standard zone 3 (0-30 cm depth).
9	aws0_100	AWS in standard zone 4 (0-100 cm depth).
10	aws0_150	AWS in standard zone 5 (0-150 cm depth).
11	aws0_999	AWS in total soil profile (0 cm to the reported depth of the soil profile),.
12	tk0_5a	Thickness of soil components used in standard layer 1 or standard zone 1 (0-5 cm).
13	tk5_20a	Thickness of soil components used in standard layer 2 (5-20 cm) .
14	tk20_50a	Thickness of soil components used in standard layer 3 (20-50 cm) .
15	tk50_100a	Thickness of soil components used in standard layer 4 (50-100 cm) .
16	tk100_150a	Thickness of soil components used in standard layer 5 (100-150 cm).
17	tk150_999a	Thickness of soil components used in standard layer 6 (150 cm to the reported depth of the soil profile).
18	tk0_20a	Thickness of soil components used in standard zone 2 (0-20 cm).
19	tk0_30a	Thickness of soil components used in standard zone 3 (0-30 cm).
20	tk0_100a	Thickness of soil components used in standard zone 4 (0-100 cm).
21	tk0_150a	Thickness of soil components used in standard zone 5 (0-150 cm).
22	tk0_999a	Thickness of soil components used in total soil profile (0 cm to the reported depth of the soil profile).
23	musumcpcta	The sum of the comppct_r* values used in the available water storage calculation for the map unit.
24	soc0_5	Soil organic carbon stock estimate (SOC) in standard layer 1 or standard zone 1 (0-5 cm depth).

25	soc5_20	SOC in standard layer 2 (5-20 cm depth).
26	soc20_50	SOC in standard layer 3 (20-50 cm depth).
27	soc50_100	SOC in standard layer 4 (50-100 cm depth).
28	soc100_150	SOC in standard layer 5 (100-150 cm depth).
29	soc150_999	SOC in standard layer 6 (150 cm to the reported depth of the soil profile).
30	soc0_20	SOC in standard zone 2 (0-20 cm depth).
31	soc0_30	SOC in standard zone 3 (0-30 cm depth).
32	soc0_100	SOC in standard zone 4 (0-100 cm depth).
33	soc0_150	SOC in standard zone 5 (0-150 cm depth).
34	soc0_999	SOC in total soil profile (0 cm to the reported depth of the soil profile).
35	tk0_5s	Thickness of soil components used in standard layer 1 or standard zone 1 (0-5 cm).
36	tk5_20s	Thickness of soil components used in standard layer 2 (5-20 cm).
37	tk20_50s	Thickness of soil components used in standard layer 3 (20-50 cm).
38	tk50_100s	Thickness of soil components used in standard layer 4 (50-100 cm).
39	tk100_150s	Thickness of soil components used in standard layer 5 (100-150 cm).
40	tk150_999s	Thickness of soil components used in standard layer 6 (150 cm to the reported depth of the soil profile).
41	tk0_20s	Thickness of soil components used in standard zone 2 (0-20 cm).
42	tk0_30s	Thickness of soil components used in standard zone 3 (0-30 cm).
43	tk0_100s	Thickness of soil components used in standard zone 4 (0-100 cm).
44	tk0_150s	Thickness of soil components used in standard zone 5 (0-150 cm).
45	tk0_999s	Thickness of soil components used in total soil profile (0 cm to the reported depth of the soil profile).
46	musumcpcts	The sum of the comppct_r* values used in the soil organic carbon calculation for the map unit.
47	nccpi2cs	NCCPI** for Corn and Soybeans (weighted average) for major earthy components.
48	nccpi2sg	NCCPI for Small Grains (weighted average) for major earthy components.
49	nccpi2all	NCCPI that has the highest value among Corn and Soybeans, Small Grains, or Cotton for major earthy components.
50	pctearthmc	The NCCPI map unit percent earthy is the map unit summed comppct_r* for major earthy components.

51	rootznemc	Depth within the soil profile that commodity crop roots can effectively extract water and nutrients for growth.
52	rootznaws	Root zone (commodity crop) available water storage estimate.
53	droughty	Indicator of available water storage within the root zone for commodity crops being less than or equal to 6 inches (152 mm
54	pwsl1pomu	The percentage of the map unit that meets the PWSL criteria.
55	musumcpct	The sum of the comppct_r* values for all listed components in the map unit.

Table 10: Definitions and names of soil quality data included in the design matrix for each crop.

* From the SSURGO component table.

** National Commodity Crop Productivity Index.