




# Plant growth stages and weather index insurance design

Jing Zou<sup>1</sup> , Martin Odening<sup>2</sup>  and Ostap Okhrin<sup>1,3</sup> 

<sup>1</sup>“Friedrich List” Faculty of Transportation, Chair of Statistics and Econometrics, Technische Universität Dresden, Dresden, Germany; <sup>2</sup>Department of Agricultural Economics, Farm Management Group, Humboldt-Universität zu Berlin, Berlin, Germany; and <sup>3</sup>Center for Scalable Data Analytics and Artificial Intelligence (ScADS.AI), Dresden/Leipzig, Germany  
**Corresponding author:** Martin Odening; Email: [m.odening@agrار.hu-berlin.de](mailto:m.odening@agrار.hu-berlin.de)

(Received 15 September 2022; revised 13 June 2023; accepted 24 June 2023; first published online 3 August 2023)

## Abstract

Given the assumption that weather risks affect crop yields, we designed a weather index insurance product for soybean producers in the US state of Illinois. By separating the entire vegetation cycle into four growth stages, we investigate whether the phase-division procedure contributes to weather–yield loss relation estimation and, hence, to basis risk mitigation. Concretely, supposing stage-variant interaction patterns between temperature-based weather index growing degree days and rainfall-based weather index cumulative rainfall, a nonparametric weather–yield loss relation is estimated by a generalized additive model. The model includes penalized B-spline (P-spline) approach based on nonlinear optimal indemnity solutions under the expected utility framework. The P-spline analysis of variance (PS-ANOVA) method is used for efficient estimation through mixed model re-parameterization. The results indicate that the phase-division models significantly outperform the benchmark whole-cycle ones either under quadratic utility or exponential utility, given different levels of risk aversions. Finally, regarding hedging effectiveness, the expected utility ratio between the phase-division contract and the whole-cycle contract, and the percentage changes of mean root square loss and variance of revenues support the proposed phase-division contract.

**Keywords:** Weather index insurance; nonlinear indemnity; plant growth stages; generalized additive model; PS-ANOVA

## 1. Introduction

Since its introduction in the 1990s, weather index insurance and weather derivatives have been regarded as risk management instruments offering numerous hedging possibilities to farmers and the agricultural sector. Potential advantages attributed to weather index insurance include the remission of loss assessment, transparency of contracts, and mitigation of moral hazard and adverse selection (e.g., Barnett *et al.*, 2008; Kellner & Mußhoff, 2011). However, thus far, weather index insurance products have not met the expectations that they have raised as financial risk management tools in the agricultural sector. As noted by Lin *et al.* (2015), the uptake of these products is relatively low. Several factors have been identified in the literature that may explain the reluctance of farmers to adopt these products, including nontransparent pricing mechanism (e.g., Xu *et al.*, 2008), high costs due to systemic weather risk (Okhrin *et al.*, 2013), and, most importantly, poor hedging effectiveness (e.g., Mußhoff *et al.*, 2011; Pelka & Mußhoff, 2013).

Hedging effectiveness is closely related to basis risk, that is, the difference between actual yield losses and indemnification (Woodard & Garcia, 2008). Following Dalhaus *et al.* (2018), basis risk can be separated into three subcomponents: geographical basis risk, design basis risk, and temporal basis risk. Much research has been conducted to improve the hedging effectiveness of weather index insurance and, thus, the willingness to pay for these products by addressing the

different components of basis risk. For example, geographic basis risk, which arises from various weather conditions at the farm location and the reference station of the index product, can be mitigated by regional diversification of insurance products (Ritter *et al.*, 2014) or by spatial interpolation techniques (Cao *et al.*, 2015). Design basis risk can be controlled by several parameters and methodological choices when modeling the complex weather–yield relationship. First, appropriate weather variables should be selected. Undoubtedly, rainfall and temperature are essential drivers of plant growth. Still, there is no consensus on the number and exact specification of the weather variables to be included in a yield model. Often used weather variables are growing degree days (GDD), cumulative rainfall (CR), dry spells, and the standardized precipitation index (cf. Turvey, 2001; Stoppa & Hess, 2003; Hill *et al.*, 2019; Okpara *et al.*, 2017). Moreover, although not fully adopted, it is widely accepted that the weather–yield relation is nonlinear (e.g., Schlenker & Roberts, 2009). However, the functional form of this relation is plant-specific. It must be adapted to the regional production environment, which calls for a flexible modeling approach, either parametrically or nonparametrically (e.g., Bokusheva, 2011; Delerce *et al.*, 2016). While early work by Vedenov & Barnett (2004) employed traditional regression methods to model the impact of weather variables on agricultural yields, machine learning techniques, specifically artificial neural networks, are increasingly used to accomplish this task (cf. Schmidt *et al.*, 2022). Most recently, B-spline and P-spline methods have been proposed as a nonparametric alternative to model the weather–yield relation by Tan & Zhang (2020) and Bucheli *et al.* (2022). An additional insertion point for reducing design basis risk is the specification of the indemnity function. The majority of existing empirical applications, such as Vedenov & Barnett (2004) and Okhrin *et al.* (2013), assumed a stepwise linear indemnity function, in which the payoff is triggered by a threshold of the weather index, increases linearly with the index, and is often capped at a maximum level. However, Zhang *et al.* (2019) extended Raviv's (1979) seminal work and proved that optimally designed weather index insurance consists of a nonlinear indemnity function. Finally, temporal basis risk can be considered as a particular aspect of design basis risk that refers to the inexpedient choice of the insured period (Conradt *et al.*, 2015).

Agronomy and crop science assert that the vulnerability of crops concerning drought and temperature stress varies during the plant growth cycle. It is well known that the demand for water increases with the plant and leaf area until the plant begins to mature (e.g., Jensen, 1968). Nielsen & Nelson (1998) studied the effects of water deficit at various growth stages of black beans. They concluded that the yield reduction is highly sensitive to limited precipitation at the flowering stage. Moreover, exposure to temperature extremes at the reproductive stage greatly affects the production of plants (Hatfield & Prueger, 2015). So far, agronomic knowledge has been mainly used to determine the beginning and end dates of the accumulation period of weather indices, for example, GDD or CR (Dalhaus *et al.*, 2018). Schierhorn *et al.* (2021) implemented machine learning techniques to explore the contribution of weather and climate to winter wheat yields during different plant development stages. However, only a few articles, such as Shi & Jiang (2016), explored the effect of weather variables and their interaction in different plant growth stages in the context of weather index insurance.

Against this backdrop, this paper aims to investigate whether the basis risk of weather index insurance can be further reduced by a growth-phase-dependent estimation of the weather–yield loss relationship using a flexible statistical modeling approach. To this end, we employ a generalized additive model (GAM) and several B-spline bivariate row tensor product smoothers (de Boor, 1978; Hastie & Tibshirani, 1987), with each smooth function representing a plant growth phase working together to fit the response variable, which is yield loss. So far, applications of B-spline methods in agricultural insurance are rare. Price *et al.* (2019) used the B-spline method to model the relationship between land quality, insurance coverage rate, and premium rate. Tan & Zhang (2020) and Bucheli *et al.* (2022) propagated this method to estimate the weather–yield relation. Although B-splines are attractive for nonparametric modeling, their specification comes with various challenges. Specifically, complicated knot selection schemes, including numbers and

positions, have been the subject of much research (e.g., Kooperberg & Stone, 1991; Ruppert, 2002). In addition, to control overfitting, the penalties are pivotal, and many variations have been developed to address this issue, such as the thin plate penalty function (Wood, 2003; Eilers *et al.*, 2015). Among the penalty variations, the difference penalty on adjacent coefficients is a particular branch of research called penalized B-spline (P-spline) (cf. Eilers & Marx, 1996). Using the equivalence between P-spline GAM and the mixed model, we apply a P-spline analysis of variance (PS-ANOVA) method. By PS-ANOVA, the knot selection can be avoided, and the smoothing parameter can be estimated in a computationally efficient manner by the decomposition of B-spline row tensor product smoothers, as noted by Lee & Durbán (2011) and Lee *et al.* (2013). The contribution of our paper is twofold. First, we explore whether breaking down the whole vegetation period into several growth stages can significantly improve the performance of weather index insurance. Second, we investigate the application of nonparametric methods, namely GAM and PS-ANOVA, in a nonlinear weather–yield loss relation estimation.

The remainder of this paper is organized as follows: section 2 introduces the optimal indemnity framework and the statistical approaches, GAM and PS-ANOVA. Section 3 describes the data, study area, and model estimation. Moreover, this section provides results and discusses the performance of a phase-division contract compared to a benchmark contract that includes the entire vegetation period of the crop. Section 4 draws conclusions on the optimal design of weather index insurance and evaluates the proposed statistical procedure.

## 2. Methodology

### 2.1. Optimal indemnity framework

In this paper, we take up the theoretical framework developed by Zhang *et al.* (2019), who extended Raviv's (1979) seminal work on optimal insurance design. In that framework, an optimal indemnity function is derived from an expected utility maximization problem, in which a risk-averse insurance buyer (a farmer) is exposed to (yield) loss  $Y$  and acquires insurance from an insurer at price  $P$  that grants an indemnity payment  $I$ . Maximizing the insured's expected utility is carried out under a participation constraint of the insurer. Raviv (1979) proved that the optimal indemnity function has the structure of a coinsurance above a deductible. Moreover, in the case of a risk-neutral insurer, the optimal indemnity function is linear. This setting, however, is not directly applicable to the design of index-based insurance, because it does not account for basis risk. To capture this aspect, Zhang *et al.* (2019) modeled indemnity payments as a function of the weather index  $X$ , that is,  $I = I(X)$ . Normalizing the crop market price to one, income loss is equivalent to yield loss  $Y$ . Yield loss and the weather index were assumed by Zhang *et al.* (2019) to have a joint probability density function  $f(y, x)$ . Further, taking a risk-neutral insurer, the optimization problem can be stated as follows:

$$\begin{cases} \max_{I \in \mathcal{I}} J(I) = E[U\{w + I(X) - Y - (1 - \tau)P\}] \\ \text{s.t. } P = \gamma E\{I(X)\} \end{cases}, \quad (1)$$

where  $\mathcal{I} := \{I | I: \mathbb{R}^2 \mapsto [0, M]\}$  is the feasible set of measurable indemnity functions with an exogenous upper limit  $M$ ,  $E$  denotes the expectation,  $U$  is the concave utility function of the insured, and  $w$  denotes initial wealth. A subsidy rate  $\tau \in [0, 1]$  is included in the utility function since the government often supports agricultural insurance, particularly in its pilot stage (Tadesse *et al.*, 2015). Due to the assumption of a risk-neutral insurer, the participation constraint takes the form  $P = \gamma E\{I(X)\}$  with a loading factor  $\gamma \geq 1$ . To further simplify the application, we set  $\gamma = 1$ . In contrast to damage-based insurance, the usual constraint  $I < Y$  does not apply. In fact, an indemnity payment is possible even if no damage occurs. The insurance premium  $P \in [0, \gamma M]$  is assumed to be exogenously specified, for example, by the insurance buyer, and the indemnity payment is optimally adjusted to this premium.

Zhang *et al.* (2019) proved the existence and uniqueness of an optimal solution for the optimization problem (1). They emphasize the finding that the optimal indemnity is generally a nonlinear function of the weather index, though a linear payoff function is usually assumed in the literature. Typically, the optimal indemnity function has to be determined numerically, but there are closed-form solutions for particular utility functions. Specifically, for the quadratic utility  $U_{qua}(r) = ar - br^2$ , with revenue  $r \leq a/(2b)$  and parameters  $a > 0$  and  $b > 0$ ,  $I(X)$  takes the form:

$$I_{qua}^*(X) = \left[ \left\{ E(Y|X) + \eta_{qua}^* \right\} \vee 0 \right] \wedge M, \tag{2}$$

where  $E(Y|X)$  is the conditional expectation of  $Y$  given  $X$ .

For the exponential utility function  $U(r) = -\frac{1}{\alpha}e^{-\alpha r}$ , with risk aversion parameter  $\alpha > 0$ , the solution is

$$I_{exp}^*(X) = \left( \left[ \frac{1}{\alpha} \log \{ E(e^{\alpha Y} | X) \} + \eta_{exp}^* \right] \vee 0 \right) \wedge M. \tag{3}$$

The constants  $\eta^*$  are determined by the constraint  $E[I^*(X)] = P/\gamma$ . Equations (2) and (3) indicate that the optimal indemnity functions under quadratic or exponential utility are affected neither by the initial wealth  $w$  nor subsidy rate  $\tau$ . Moreover, it becomes apparent that the estimation of conditional expectations of yield loss, given the weather index, is pivotal to the practical implementation of this kind of weather insurance. In our empirical application, we will focus on the indemnity functions (2) and (3).

**2.2. Statistical approach**

In the optimal indemnity equation, the conditional expectation  $E(Y|X)$  or  $E(e^{\alpha Y} | X)$  is essentially a regression model of  $Y$  or  $e^{\alpha Y}$ , given weather index  $X$ . As mentioned before, we employ GAM with B-spline row tensor product smoothers to estimate the conditional expectation and then embed it into the optimal indemnity function. We are not restricted to B-splines as GAM allows for flexibility in selecting smoothers, such as kernels or cubic splines (Hastie & Tibshirani, 1987). In this work, we use splines based on our preliminary analysis. Kernel estimation provides similar results, yet the models are often overparameterized. As introduced by Eilers & Marx (1996), P-spline, defined as the combination of B-spline with difference penalties on the estimated coefficients, has a valuable property that it shows no boundary effects, while many types of kernel smoothers do exhibit these. Aydin (2007) also concluded that smoothing spline regression estimators outperform the kernel ones. To the best of our knowledge, the applications of splines in agricultural insurance are rare, as mentioned in the Introduction.

To investigate the interaction between weather indices  $X_1$  and  $X_2$  and their effects on the response  $Y$ , we construct a GAM including one smoother  $S$ , that is,  $Y = S(X_1, X_2)$ , and denote the vectors of these three variables as  $y$ ,  $x_1$ , and  $x_2$ , respectively. Firstly, we obtain the marginal B-spline basis  $B_{i,j}^u$  of order  $j$  for each weather index through de-Boor recursion (de Boor, 1978):

$$B_{i,0}(u) = \begin{cases} 1, & \text{if } u_i \leq u < u_{i+1} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \tag{4}$$

$$B_{i,j} = \frac{u - u_i}{u_{i+j} - u_i} B_{i,j-1}(u) + \frac{u_{i+j+1} - u}{u_{i+j+1} - u_{i+1}} B_{i+1,j-1}(u), \tag{5}$$

where  $u$  denotes the element in the weather index vector  $x_1$  or  $x_2$ ,  $\{u_i\}$  is a uniform knot vector determined by  $k$  which is the number of equally spaced knots over the domain of  $x_1$  or  $x_2$  (strictly  $k - 1$  is the number of internal knots), and  $i = 0, \dots, k + j - 1$ , which represents the  $i$ th basis function (Lee *et al.*, 2013). Throughout this study, we use cubic splines ( $j = 1, 2, 3$ ), which

lead to functions with continuous second derivatives. Secondly, the smooth function  $S(x_1, x_2)$  is constructed by the B-spline tensor product  $B$  and the GAM with one smoother is given by:

$$y = S(x_1, x_2) = B\theta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{6}$$

where  $B$  and  $\theta$  are the “regression basis” and vector of “regression coefficients,” respectively (cf. Lee & Durbán, 2011) and  $\varepsilon$  is a Gaussian error term with variance  $\sigma^2 I$ . For efficient computation, row tensor product or box product  $B = B_2 \square B_1 = (B_2 \otimes 1'_{c_1}) * (1'_{c_2} \otimes B_1)$  instead of the Kronecker product  $B_2 \otimes B_1$  is used, where  $B_q$  for  $q = 1, 2$  denotes the marginal B-spline basis constructed from the weather indices  $x_q$  by (4) and (5) and  $c_q$  is a vector of ones of length  $c_q$  which is the number of columns in  $B_q$  (cf. Eilers *et al.*, 2006; Lee *et al.*, 2013). To control the overfitting and overparameterization, separate difference penalties are imposed on adjacent coefficients in  $\theta$  along the two dimensions, which creates P-splines. For methodological details, we refer interested readers to Eilers & Marx (1996), Lee & Durbán (2011), and Marx & Eilers (1998). The number of knots  $k$  lying between 20 and 40 is considered to be moderate (Ruppert *et al.*, 2003).

We employ the P-spline ANOVA (PS-ANOVA) method to estimate the penalty terms. The PS-ANOVA method originated from the idea that through re-parameterizing the P-spline GAM defined above as (6) into a mixed model that contains both fixed and random effects (e.g., Brumback & Rice, 1998), the smoothing parameter becomes the ratio between the variance of the residuals and the variance of the random effect in a mixed model. Therefore, the smoothing parameter can be estimated using ANOVA decomposition (Currie & Durbán, 2002; Lee & Durbán, 2011). This transformed mixed model is

$$y = F\beta + Z\delta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \delta \sim \mathcal{N}(0, G), \tag{7}$$

where “:” distinguishes the blocks in a matrix and the original box product  $B = [F : Z]$  becomes a block matrix consisting of fixed effects matrix  $F \equiv [1_d : x_1 : x_2 : x_2 \square x_1]$  and random effects matrix  $Z \equiv [Z_1 : Z_2 : Z_2 \square x_1 : x_2 \square Z_1 : Z_2 \square Z_1]$ .  $Z_q = B_q U_q$  for  $q = 1, 2$  and  $U_q$  are eigenvectors corresponding to the positive eigenvalues of the singular value decomposition of  $D'_q D_q$ . The original coefficient  $\theta = (\beta, \delta)^T$  becomes a vector including the fixed effects coefficient  $\beta$  and the random effects coefficient  $\delta$  that is assumed to be Gaussian with covariance matrix  $G$ . Note that the response variable in the mixed model is assumed to be Gaussian in this paper. Moreover, an isotropic penalization is conducted for simplification, which means that the same amount of smoothing is used for all covariates (Rodríguez-Álvarez *et al.*, 2015). There are two main advantages of the PS-ANOVA method. First, one penalty for each covariate is attractive for tuning multiple penalty parameters, especially when the GAM includes several smooth functions (Lee *et al.*, 2013). Second, the variance components of a mixed model can be estimated based on restricted maximum likelihood (REML), which is proved empirically to be preferable to other selection criteria, such as generalized cross-validation (GCV) or Akaike’s information criterion (AIC) (Schall, 1991; Reiss & Todd Ogden, 2009; Wood, 2011).<sup>1</sup>

### 3. Empirical Application: Designing Weather Index Insurance for Soybean Production

#### 3.1. Data and study area

In our empirical study, we use the methodology described in section 2 to design a weather index insurance product for soybean producers in the US state of Illinois. Over one-third of soybeans

<sup>1</sup>Moreover, we compare the PS-ANOVA approach with the benchmark method, that is, the `gam()` function in the “mgcv” package (version 1.8-40), which is the reference R package for GAM estimation in recent years (Rodríguez-Álvarez *et al.*, 2015). In `gam()`, we choose `te()` and set `bs = “ps”` for cubic second-order P-spline tensor product smoothers, which are consistent with PS-ANOVA settings as described in the reference manual for the “mgcv” package (version 1.8-40). In addition, we set `method = “REML”` in `gam()`, where REML is obtained by Laplace approximation and Newton-Raphson iteration (see Wood, 2011 for details), rather than fisher scoring, which is used in the PS-ANOVA method (Lee *et al.*, 2013).

traded on the global market are produced in the USA, and the value of US soybean exports reached a record of \$27.4 billion in 2021 (USDA, 2022). Illinois contributed more than any other US state to soybean output in 2021 (The Illinois Soybean Association, 2022). Thus, significant disruption to soybean production in Illinois will not only affect local farmers but can also influence the international market substantially (Boyer *et al.*, 2013). Soybean production is sensitive to weather conditions and specific weather events like hot-dry summer conditions, which can severely affect US soybean yields (Hamed *et al.*, 2021). Insurance policies are available that cover yield losses for US soybean producers, such as the Supplemental Coverage Option (SCO) and Enhanced Coverage Option (ECO) (Schnitkey *et al.*, 2022). Nevertheless, it appears reasonable to design a weather index insurance product to complement existing contracts given the current situation in which Rainfall Index Insurance is a single-peril insurance that only covers apiculture, annual forage, pasture, and rangeland (USDA, 2021).

The data used in our empirical analysis are publicly available, including yield, meteorological, and phenological data. County-level soybean yield data from 96 counties in Illinois from 1998 to 2020 are from the National Agricultural Statistics Services (NASS) (Fig. 1, upper panel). We detrend raw yield data by a linear time regression model to obtain detrended yields (Shi & Jiang, 2016) (Fig. 1, middle panel). Further, following Zhang *et al.* (2019), yield losses<sup>2</sup>  $Y$  in each county are calculated as the difference between the highest detrended yield during the dataset's 23 years and detrended yields (Fig. 1, bottom panel). The average raw yield is about 48.46 bushels/acre, and the average detrended yield is approximately 38.46 bushels/acre. Yield losses range from 0 to 35.17, and the average yield loss amounts to 9.53 bushels/acre.

Since temperature and precipitation are major determinants of plant growth (e.g., Walter, 1985; Donoghue, 2008), we use GDD and CR as weather-related indices in our analysis. We confine our analysis to a two-dimensional case, since this research aims to investigate the interaction between temperature and precipitation and their co-impacts on yield losses in different growth stages. We average the gridded daily weather data collected from the *Daymet* data set (Thornton *et al.*, 2020) for the weather indices calculation in each county. *GDD* and *CR* are defined as:

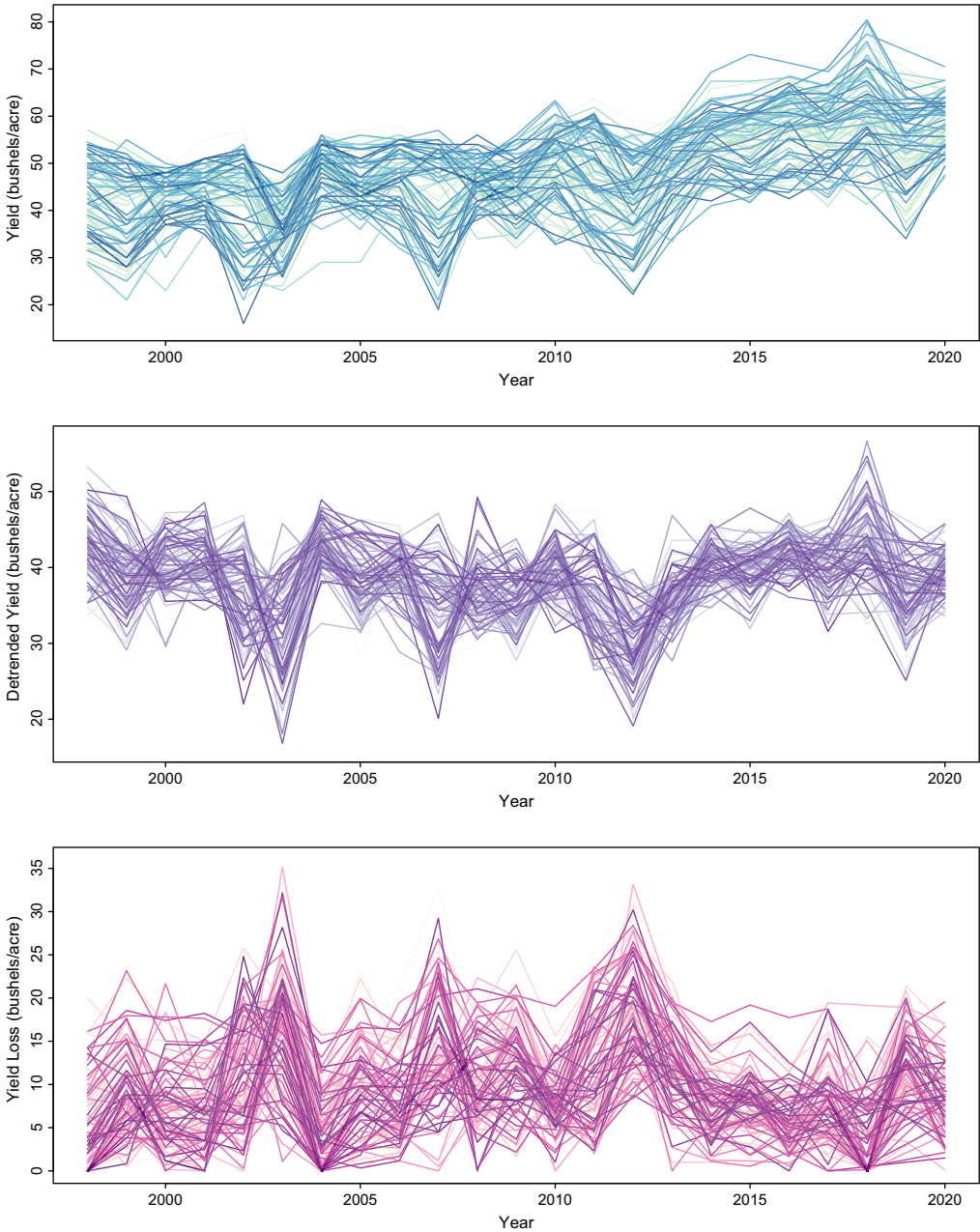
$$GDD_{m,l} = \sum_{t=sd_m}^{ed_m} \max \left\{ 0, \frac{(TMA_{m,t,l} + TMI_{m,t,l})}{2} - TB \right\} \text{ and} \quad (8)$$

$$CR_{m,l} = \sum_{t=sd_m}^{ed_m} R_{m,t,l}, \quad (9)$$

where  $m$  and  $l$  denote the years and the region (county), respectively,  $t$  is the  $t$ -th day in a year,  $sd$  and  $ed$  denote the start and the end date of the accumulation period, respectively, and  $TMA$ ,  $TMI$ , and  $R$  are daily maximum temperature, minimum temperature, and rainfall, respectively. The baseline temperature  $TB$  of soybean growth is defined as 10° Celsius (Scholtes *et al.*, 2019). GDD and CR indices have been criticized because they do not account for the temporal distribution of weather events within the accumulation period. This weakness, however, is attenuated because we divide the accumulation period into shorter periods. In addition, we implement a Rainfall Deficit Index (RDI) that has been suggested as an alternative to CR (e.g., Odening *et al.*, 2007). A definition of this index as well as model results are presented in the Appendix.

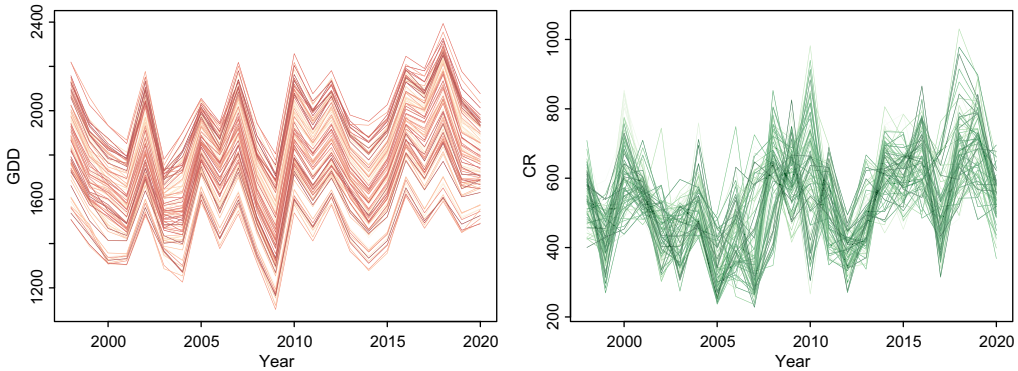
Agronomic research asserts that the weather–yield relationship is time-variant, that is, meteorological factors affect plant development differently during each growth stage (e.g., Delerce *et al.*, 2016). Thus, we divide the whole growth cycle of soybeans into four phases. According to the terms and definitions of the NASS (2018), these four phases are described as “emerged,” “blooming,” “setting pods,” and “dropping leaves.” The phase-division procedure is based on the weekly state-level *Crop Progress Report* (USDA-Economics, Statistics and Market Information System

<sup>2</sup>The definition of yield loss in this article deviates from the convention to measure losses as difference of actual yields from the mean; see Schmidt *et al.* (2022).



**Figure 1.** Yield (upper panel), detrended yield (middle panel), and yield loss (bottom panel) for 96 counties in Illinois. Each line represents one county.

(ESMIS), 2022). For the first stage, “emerged,” for example, we selected the dates for which germination rates are closest to 1% and 100% as the start and end dates, respectively, for this phase. As a result, these dates may change slightly from year to year. For example, in 2020, the beginning and end dates of the first phase are May 3 and June 28, respectively, which deviate about 1 week from those in 1998. Moreover, temporal overlapping between two subsequent growth stages can occur. For example, “blooming” started on June 28, 1998, before the complete germination of



**Figure 2.** Time series of growing degree days (GDD) and cumulative rainfall (CR) in the whole cycle. Each line represents one county.

soybeans ended. To explore the performance gain from dividing the entire vegetation period, we also estimate a benchmark model without decomposition whose beginning and end dates of the accumulation period are defined as the start date of the first growth phase and the end date of the last growth phase, respectively. In our study, we obtained the stage division points from the Crop Progress Report, which was only available after all the necessary information had been collected. Unfortunately, this means that the direct application of insurance contracts is not feasible because we do not know the division points for future years that will be insured. The challenge is compounded by our limited understanding of the underlying mechanism behind plant phenology, as discussed by Tang *et al.* (2016). Given these uncertainties, we opted to rely on ground observations as the basis for our division procedures.

Figs. 2, 3, and 4 depict the GDD and CR values of 96 counties over the observation period for the entire growth cycle at each growth stage. In Figs. 2, 3, and 4, the values are annual and accumulative, and each line is smoothed by observation points across 23 years. Each subplot reflects accumulative values at a particular phase.

For example, in Fig. 3, upper left plot, each line represents the GDD values in the “Emerged” stage across 23 years of one specific county, where stage 1 of the first year is immediately followed by stage 1 of the second year. Some facts are notable. First, both weather indices show a considerable variation over time, reflecting the prevalence of weather risk. Second, the graphs reveal spatial heterogeneity of weather conditions across Illinois, though all counties follow similar patterns. Third, unsurprisingly, spatial correlation is less pronounced for rainfall than for temperature. Fourth, the patterns for the two weather indices differ across the four growth stages, which motivates our proposed division. Actually, though the observation period is rather short, some changes in weather patterns can be observed which may be interpreted as a manifestation of climate change. There are several aspects to be mentioned in this context. First of all, it is essential to distinguish between changes in levels (e.g., “global warming”) and changes in weather variability. From an insurance perspective, the latter is more important. The question of whether weather becomes more volatile over time has been intensively discussed in the literature (e.g., Wang *et al.*, 2013). Accordingly, models have been developed that capture time-varying volatility of temperature and rainfall processes (e.g., Okhrin *et al.*, 2013). A related question is whether yields become more volatile due to climate change and what amount of historical data should be used to estimate models (e.g., Shen *et al.*, 2018). Admittedly, our modeling approach does not account for changing parameters of the stochastic weather processes. However, this seems acceptable, because our paper focuses on the optimal design insurance contract that depends on the weather–yield relationship and this relationship is not directly affected by climate change.



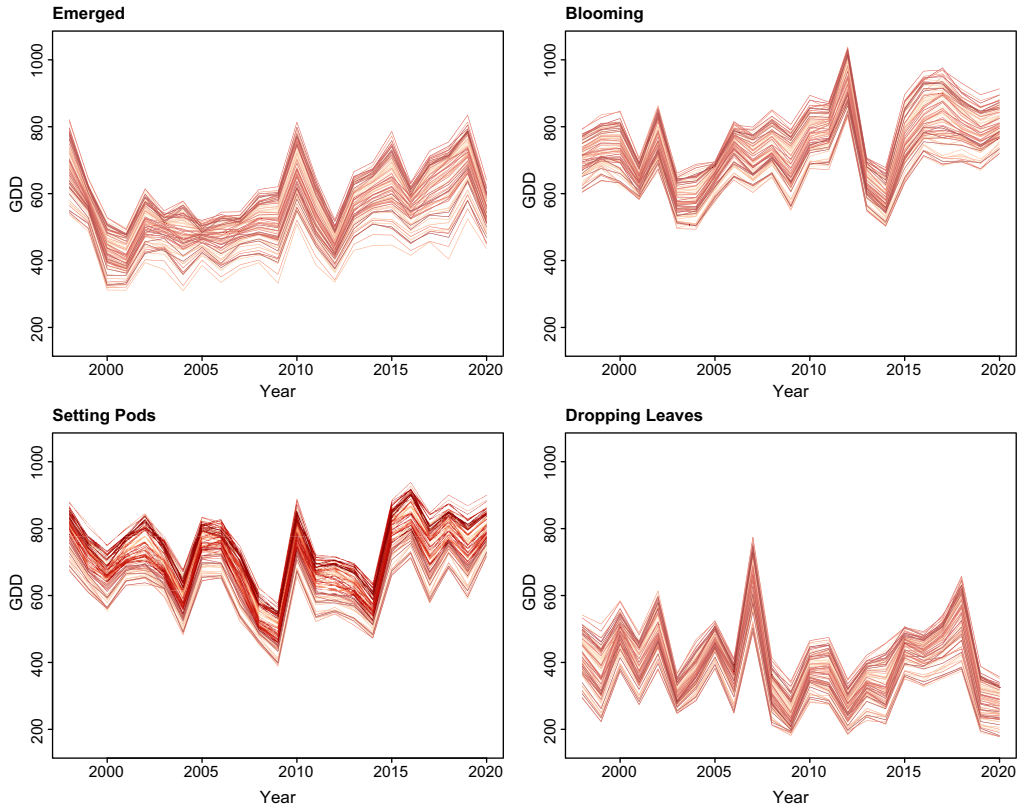


Figure 3. Time series of growing degree days (GDD) in separate phases. Each line represents one county.

3.2. Model estimation

We determine the optimal indemnity function under quadratic and exponential utility functions, which requires estimating conditional expectations of yield loss  $Y$  and  $e^{\alpha Y}$ , respectively, given the weather indices  $X$  and a specific combination of  $P$  and  $M$  according to equations (2) and (3), respectively. The estimation of the conditional expectation is conducted for two settings. First, we estimate conditional yield loss based on the weather–yield loss relation for the entire growth period, for example,  $Y_{m,l} = S_0(GDD_{m,0,l}, CR_{m,0,l}) + \varepsilon_{m,0,l}$  in the quadratic utility case, where the subscript zero denotes the entire growth cycle. This GAM serves as a benchmark in our analysis. Second, we divide the whole growth period into four phases and estimate the weather–yield loss relation for the phase-division GAM consisting of four interaction smoothers, namely:

$$Y_{m,l} = S_1(GDD_{m,1,l}, CR_{m,1,l}) + S_2(GDD_{m,2,l}, CR_{m,2,l}) + S_3(GDD_{m,3,l}, CR_{m,3,l}) + S_4(GDD_{m,4,l}, CR_{m,4,l}) + \varepsilon'_{m,s,l} \tag{10}$$

where the subscripts 1, 2, 3, and 4 represent the phases of “emerged,” “blooming,” “setting pods,” and “dropping leaves,” respectively, and  $s$  is an overall symbol denoting phase division. Moreover, to consider interactions among weather variables within different stages, we attempted to model the interactions between all four phases (one eight-dimensional smoother). Unfortunately, it was computationally highly demanding. Therefore, we allow for interactions within the four variables of the first two phases, and within the four variables of the third and the fourth phase, that is, a two-segment model containing two four-dimensional smoothers as a benchmark model:

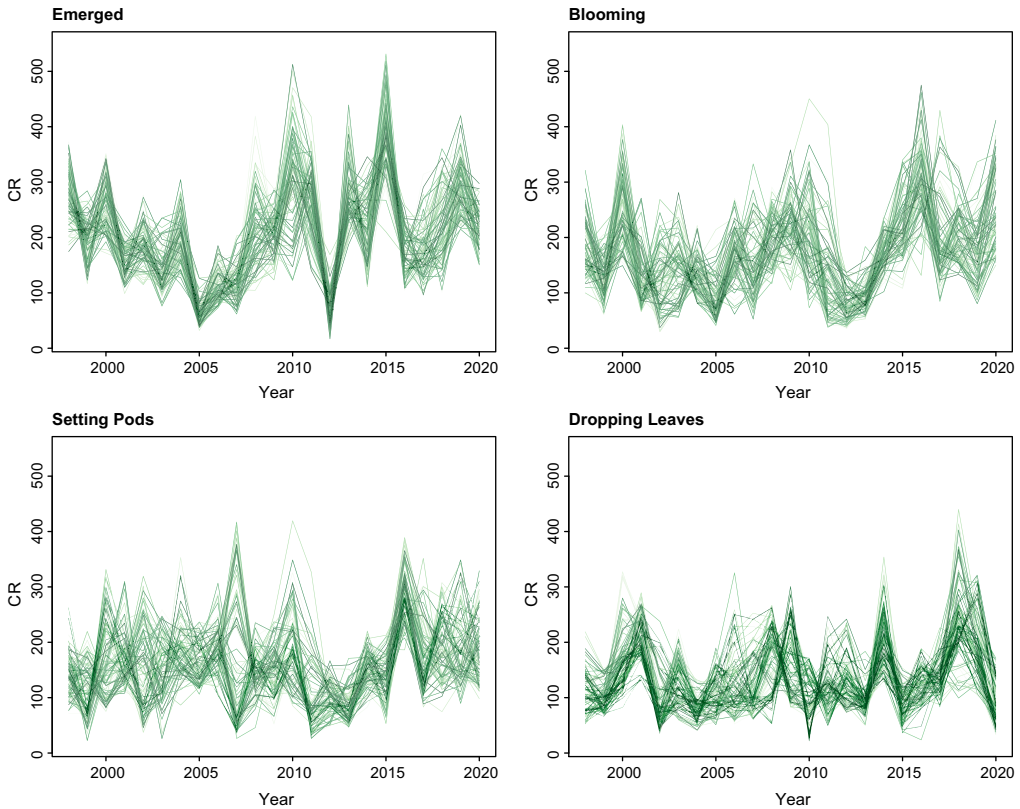


Figure 4. Time series of cumulative rainfall (CR) in separate phases. Each line represents one county.

$$Y_{m,l} = S_I(GDD_{m,1,l}, CR_{m,1,l}, GDD_{m,2,l}, CR_{m,2,l}) + S_{II}(GDD_{m,3,l}, CR_{m,3,l}, GDD_{m,4,l}, CR_{m,4,l}) + \varepsilon''_{m,s,l}. \tag{11}$$

Only the benchmark method `mgcv::gam()` is used for the estimation of the model (11) due to computational burden. There might be more efficient ways to handle high-dimensional weather variables, such as neural network models (Crane-Droesch, 2018; Chen *et al.*, 2020) or regression tree-based models. However, in our research, we specifically chose the GAM to prioritize interpretability for each separate growth phase. While neural networks may offer potential improvements in model performance, the relatively small dimensions of our study (only eight variables) suggest that nonparametric regression, as employed in the components of GAM, provides qualitatively similar results. Additionally, neural networks often require extensive hyperparameter tuning, which can be time-consuming and challenging. Thus, the investigation of the performance of other methods in phase-division crop yield modeling is beyond the scope of this study.

In the case of exponential utility, the response variable of GAM is  $e^{\alpha Y}$ , where  $\alpha$ — is the (absolute) risk aversion coefficient. In our analysis, we consider three levels of risk aversion for US soybean farmers: low-, moderate-, and high-risk aversion. To determine appropriate levels for the risk aversion coefficient, we follow Tan & Zhang (2020), who derived  $\alpha$ — by dividing a relative risk aversion coefficient of 2, 3, and 4, respectively, by an estimate of the initial wealth of corn producers in Illinois. This results in the following absolute risk aversion coefficients  $\alpha$ — = 0.0052 (low), 0.008 (moderate), and 0.0103 (high).

**Table 1.** RMSE and adjusted  $R^2$  of GDD-CR GAMs estimated by PS-ANOVA & mgcv : : gam( ).

PS-ANOVA					
Whole-cycle	RMSE	Adj. $R^2$	Phase-division	RMSE	Adj. $R^2$
Quadratic	5.14	0.27	Quadratic	4.04	0.55
Exp-low	5.14	0.27	Exp-low	4.04	0.55
Exp-moderate	5.10	0.29	Exp-moderate	4.04	0.55
Exp-high	5.10	0.29	Exp-high	4.04	0.55
gam( ) in "mgcv"					
Whole-cycle	RMSE	Adj. $R^2$	Phase-division	RMSE	Adj. $R^2$
Quadratic	5.24	0.24	Quadratic	4.40	0.46
Exp-low	5.24	0.24	Exp-low	4.39	0.47
Exp-moderate	5.24	0.25	Exp-moderate	4.39	0.47
Exp-high	5.24	0.25	Exp-high	4.39	0.47

Notes: RMSE is the root mean squared error.

To circumvent potential overfitting, cross-validation is conducted to select the hyperparameter  $k$ , that is, the number of knots. To this end, the entire data set was divided into three. The training data set contains data from 1998 to 2013, and the validation data set includes data from 2014 to 2017, while the test data set contains data from 2018 to 2020. For computational stability, we normalize the values of all variables into the range [0, 1]. After estimation, we convert the fitted values back to the original scale to obtain the scale-dependent root mean squared error:

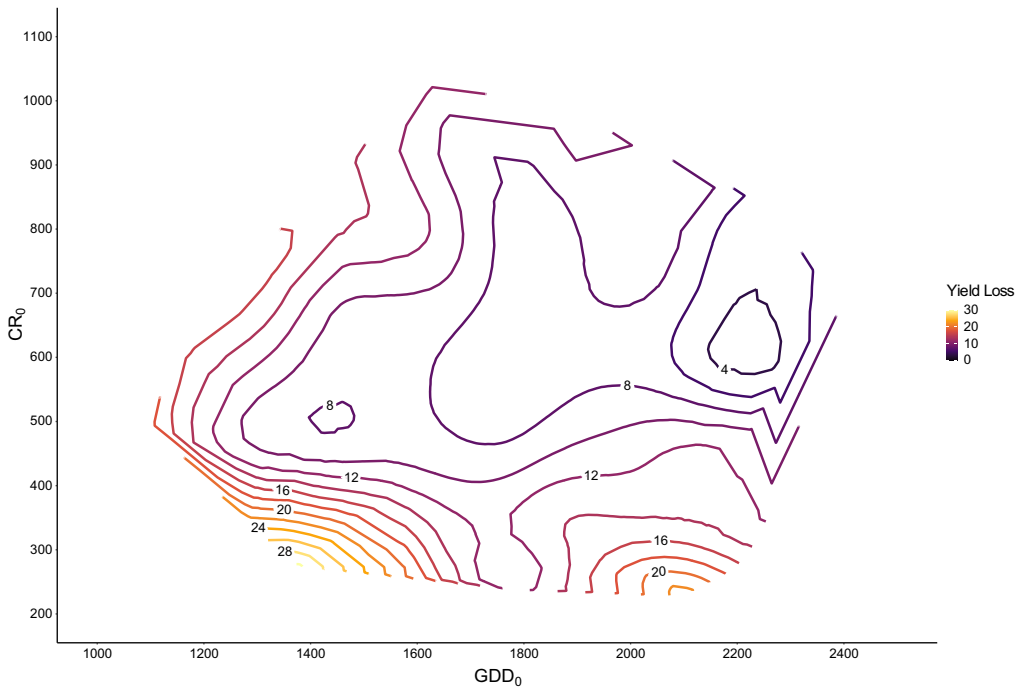
$RMSE = \sqrt{\frac{\sum_{m=1}^n \sum_{l=1}^{96} (\widehat{y}_{m,l} - y_{m,l})^2}{96 * n}}$ , where  $n$  is the time series length of each county in the data set, for example,  $n = 16$  for the training set.  $RMSE$  measures the average deviation between predicted and observed values and serves as our study’s model evaluation criterion (Schmidt *et al.*, 2022). We select the appropriate  $k$  values in the range [20, 40] by cross-validation. For whole-cycle models, cross-validation resulted in  $k = 21$  for the one under quadratic utility and the one under exponential utility given  $\alpha = 0.0052$ ;  $k = 30$  for the models under exponential utility given  $\alpha = 0.008$  and  $\alpha = 0.0103$ . For all phase-division models, cross-validation resulted in  $k = 39$ . After cross-validation, we estimate the models with the full sample size from 1998 to 2020 using selected  $k$  values.

### 3.3. Results from the weather–yield loss relation

Table 1 shows that in all cases, the models based on phase division significantly outperform the whole-cycle models in terms of  $RMSE$  and adjusted  $R^2$ .<sup>3</sup>

Compared with the benchmark method mgcv : : gam( ), PS-ANOVA shows a slightly better fit.  $RMSE$  of the two-segment model (11) is 3.79 bushels/acre, and the Adj. $R^2$  is 0.60, indicating performance gain compared to the original four-phase-division model (10). There is clearly a trade-off between model performance and model parsimony. Despite this, model (10) effectively isolates the impact of each phase on yield losses, as we will see later. In contrast, the interpretability of the two-segment model (11) is limited not only by the arbitrary nature of segment numbers – as we could theoretically conduct a three-segment model containing two three-dimensional and one two-dimensional tensor products – but also by the challenges of visualizing interactions between more than three variables. As a result, the specific effects of each phase may not be clear. Through our analysis and the results presented in the figures and the discussion that we will see in

<sup>3</sup>The adjusted  $R^2$  is calculated by the classical Ezekiel estimator (Ezekiel, 1930).



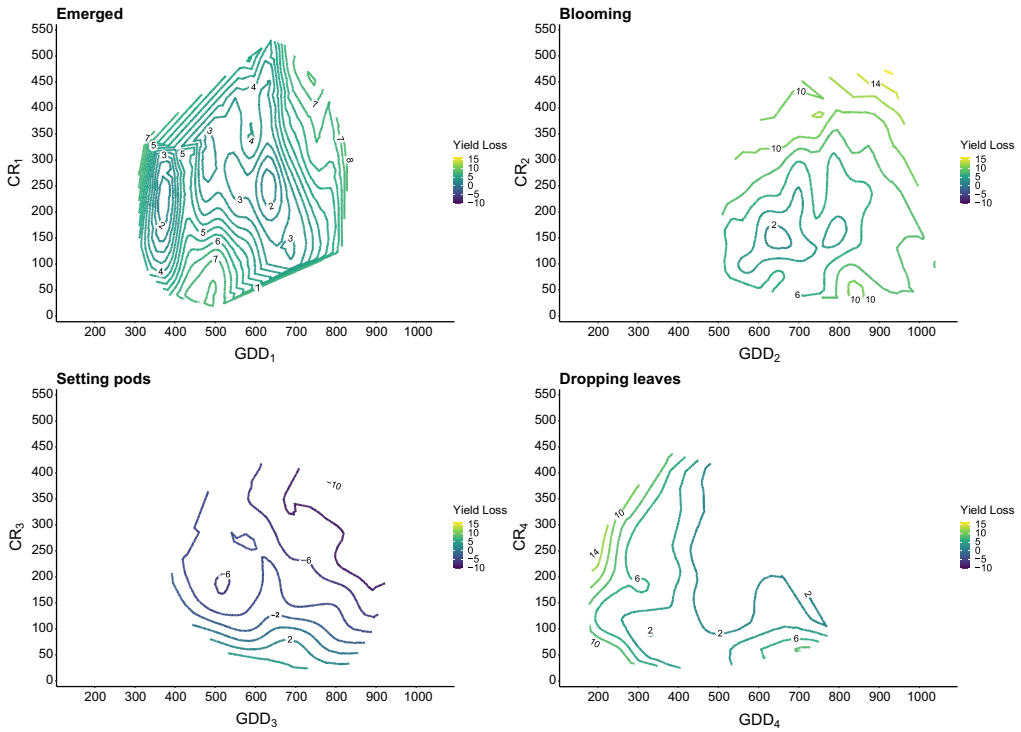
**Figure 5.** Whole-cycle weather–yield loss relation.

Notes: CR is cumulative rainfall and is given in millimeters (mm). GDD is growing degree days and is given in degrees Celsius. The subscript 0 after CR and GDD indicates that the whole-cycle model is utilized rather than the models that separate the four growing phases.

section 3.3, we could clearly separate and quantify the contributions of each phase to yield losses. This interpretability is crucial for understanding the underlying factors affecting crop yield and designing effective index insurance.

A contour plot of the weather–yield loss relation is shown in Fig. 5 for the whole-cycle model. It visualizes the complex interaction of variables  $GDD$  and  $CR$  as determinants of yield losses in soybean production. Overall, soybean yield loss decreases with increasing  $GDD$  because higher temperatures induce a faster crop development provided that a sufficient amount of rainfall is available, in this case, more than 500 mm (Peiris *et al.*, 1996). When the  $GDD$  is as high as  $2,200^{\circ}$  Celsius and the  $CR$  amounts to approximately 600 mm, yield loss is at its minimum (about four bushels/acre), which is less than half of the average yield loss (9.53 bushels/acre). However, when the  $CR$  is below 500 mm, the contour lines appear symmetric. It means that under a deficit in rainfall, both insufficient and excessive  $GDD$  cause yield losses. Specifically, when  $GDD$  exceeds the threshold of  $1,800^{\circ}$  Celsius, the joint occurrence of rainfall deficiency and heat stress causes significant yield loss. Thus, our analysis confirms the negative impact of hot-dry conditions on US soybean, which has been found in previous studies (e.g., Hamed *et al.*, 2021). Furthermore, Fig. 5 depicts that excessive rainfall also induces yield losses, particularly under cold conditions. However, soybean yield losses are less sensitive to excessive rainfall than drought.

Fig. 6 depicts the impact of temperature and rainfall on soybean yield losses in each growth phase. In each subplot, the displayed values measure the contribution to total yield loss, where negative values indicate a reduction of yield losses. Comparing the four subplots reveals that the weather–yield loss relation differs significantly between the four growth phases. The level of the loss contribution, sensitivity to weather events, and the interaction between temperature and precipitation show pronounced differences. In the first phase (“emerged”), small losses occur at the



**Figure 6.** Phase-division weather–yield loss relation.

Notes: Each subplot represents the yield loss contribution from the interaction between growing degree days (GDD) and cumulative rainfall (CR) at the corresponding growth phase. The four growing phases are shown in the subscripts on the labels of the axes: 1 indicates emerged, 2 blooming, 3 setting pods, and 4 dropping leaves. CR is in mm and GDD is in degrees Celsius.

GDD–CR combination of around (650° Celsius, 200 mm), whereas large losses take place when CR is lower than 100 mm, and GDD is below 550° Celsius. Both CR and GDD influence yield production positively within some range but result in a negative impact after exceeding a certain level. In 2012, for example, the CR of 32 counties added up to less than 50 mm in this development phase. As a result of this countrywide drought, a considerable production disruption was reported, with an estimated soybean yield loss of around 170 million bushels (Boyer *et al.*, 2013; Rippey, 2015). The contour plots of the second phase (“blooming”) show that potential yield losses are comparably large, that is, soybean plants are vulnerable in this growth phase. Minimal yield losses are realized at 650° Celsius and 150 mm. Notably, plants are more sensitive to excessive rainfall than drought in this phase. High temperatures advance blooming (Cooper, 2003), and more precipitation will likely cause a longer blooming period and high outcrossing rate (Qu *et al.*, 2020). Another possible explanation is that under persistent wet weather, the disease *sclerotinia stem rot* could attack soybeans during reproductive stages and, in turn, cause yield loss (Wrather & Koenning, 2009).

The subplot of the third stage (“setting pods”) suggests a pattern that can be summarized as “the warmer and wetter, the better.” When GDD is in the range [400, 650], yield loss remains unchanged if the CR level exceeds 125 mm. This indicates that temperature conditions restrict the soybean growth rate. Once GDD exceeds 650° Celsius, more rainfall means higher yield gain. The plot also shows that GDD affects yield loss only slightly under rainfall deficit conditions. This can be explained by pods’ abortion under drought stress, since sufficient water is critical to fill

the soybean pods (Liu *et al.*, 2004). In the fourth phase (“dropping leaves”), the soybean yield is relatively robust against varying rainfall. For instance, when *GDD* lies in the range [175, 500] and *CR* is above 150 mm, the contour lines are approximately vertical, which implies that rainfall increments have little impact on yield loss. This result reflects the fact that the plant’s demand for water declines when reaching maturity (Jensen, 1968). On the other hand, soybean’s temperature demand seems to be relatively strong until *GDD* achieves 500° Celsius. The bottom right area of the plot visualizes that yield losses begin to increase under the hot-dry weather. One possible interpretation for this phenomenon is that drought stress increases crop vulnerability to aphids, a transmitter of the *Soybean Mosaic Virus*, and further suppresses soybean yields (Rice *et al.*, 2007; van Munster *et al.*, 2017).

### 3.4. Hedging effectiveness

Though estimating the conditional loss expectation for different plant growth phases provides exciting insights into the weather–yield relationship, it is not yet clear if the suggested procedure can improve the design of weather index insurance. To answer this question, we calculate the hedging effectiveness of a phase-division contract and a standard contract based on the entire vegetation period. Hedging effectiveness is defined in terms of the expected utility of the insurance contract. We determine the optimal indemnity functions under the assumption of an exponential utility function and parametrize the level of risk aversion. Under exponential utility, *EU* is  $E(-\frac{1}{\alpha}e^{-\alpha r})$  and revenue is  $r = w + I^*(X) - Y - (1 - \tau)P$  after the indemnity payment. As we are interested in the relative performance of the phase-division contract compared with the standard contract, we determine the ratio of the expected utility of the two contracts. Thus, the initial wealth *w*, premium *P*, and subsidy rate  $\tau$  are canceled out, and the expectation of the response variable  $e^{\alpha Y}$  is equivalent to the product of the fixed effects matrix and fixed effects coefficient, that is,  $E(e^{\alpha Y}) = F\beta$  given the zero-mean assumption of the random effect  $\delta$  and the model error term  $\varepsilon$  in the mixed model (7). We conduct the calculation for each county, that is, we consider each county as a representative farmer in Illinois.

Up to this point, estimating a distribution of weather variables was not required since calculating the optimal indemnity function is based on the conditional loss expectation given the weather conditions. To arrive at the expected utility, we simply average the observed realizations of weather variables in the 23 years in the observation period. With this simplified historical simulation, the *EU* ratio under exponential utility takes the following form:

$$\frac{\widehat{EU}_{s,l}}{\widehat{EU}_{0,l}} = \frac{1}{23} \sum_{m=1}^{23} \frac{e^{-\alpha \widehat{I}_{m,s,l}^*(X_{m,s,l})} F_{m,s,l} \widehat{\beta}_{m,s,l}}{e^{-\alpha \widehat{I}_{m,0,l}^*(X_{m,0,l})} F_{m,0,l} \widehat{\beta}_{m,0,l}}, \tag{12}$$

where the subscripts *s* and zero denote “phase-division” and “whole-cycle,” respectively, as defined in section 3.2, and *m* and *l* are year and county indices, respectively. Considering that the average yield loss is about 9.53 bushels/acre and the maximum yield loss is approximately 35.17 bushels/acre, we assume that the premium *P* is 9 \$/acre and the maximum indemnity *M* is \$35 per acre, which normalizes the soybean price to \$1/bushel without the loss of generality. Under these assumptions, the optimal indemnities  $I^*(X)$  of phase-division and whole-cycle contracts are obtained according to equation (3). The 96 counties are clustered into nine groups according to the agricultural district attributes defined by the NASS (2022). The nine groups are “Northwest” (10), “Northeast” (20), “West” (30), “Central” (40), “East” (50), “West Southwest” (60), “East Southeast” (70), “Southwest” (80), and “Southeast” (90). Fig. 7 depicts which counties are included in these groups and where they are located.

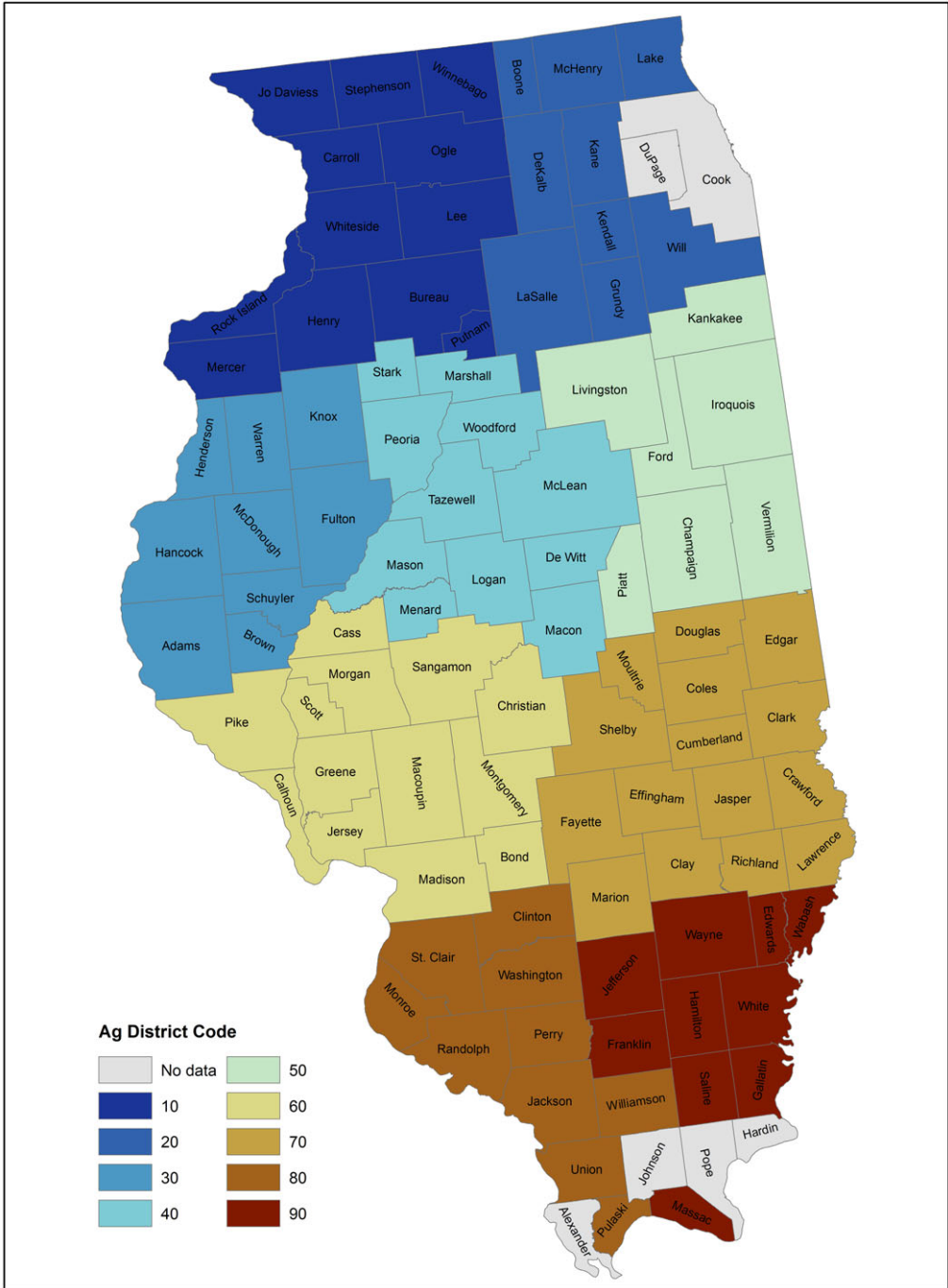


Figure 7. County-level agricultural district classification in Illinois.

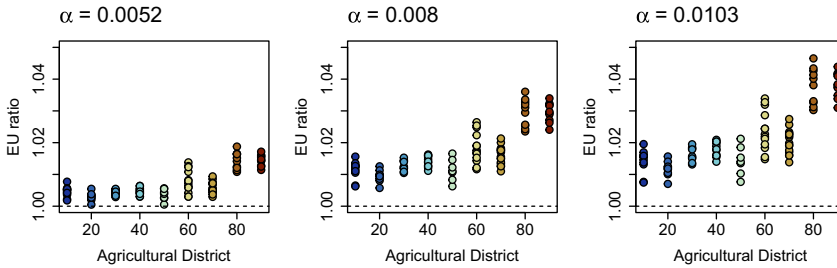


Figure 8. EU ratio under three levels of risk aversion.

Fig. 8 presents the results of the hedging effectiveness calculation. We find that the EU ratios are above one across all nine regions, which indicates that the phase-division contract outperforms the whole-cycle one. However, the gain is modest. For instance, when  $\alpha = 0.008$ , the mean of all the 96 EU ratios is about 1.017, that is, on average, the expected utility of the phase-division contract is only 1.7% higher compared to the whole-cycle contract. For more risk-averse policyholders ( $\alpha = 0.0103$ ), the relative superiority of the phase-division contract increases: the mean of the EU ratios is about 1.022. Furthermore, we observe that EU ratios vary across agricultural districts. For example, in the East District (50), the EU ratio is considerably lower than in the Southwest district (80) for all three levels of risk aversion. In addition, we implement the mean root square loss (MRSL) (e.g., Vedenov & Barnett, 2004) and variance of revenue (VAR) as alternative performance measures and calculate its percentage change with and without insurance for both the whole-cycle and the phase-division contract.

For each county, the MRSL, VAR, and the percentage changes of MRSL and VAR after insurance contract purchasing are defined as:

$$MRSL = \sqrt{\frac{1}{23} \sum_{m=1}^{23} \left\{ \max \left( 0, price * \overline{yield} - r_m \right) \right\}^2}, \tag{13}$$

$$VAR = \sqrt{\frac{1}{23} \sum_{m=1}^{23} \left( price * \overline{yield} - r_m \right)^2}, \tag{14}$$

$$MRSLPC = \frac{MRSL_{with} - MRSL_{without}}{MRSL_{with}}, \tag{15}$$

$$VARPC = \frac{VAR_{with} - VAR_{without}}{VAR_{with}}, \tag{16}$$

where *price* is set to be \$1/bushel to represent the soybean price; average yield across 23 years is *yield*; *m* is the year, and  $r_m$  is the yearly revenue of each county. In the case of “with” contract, the revenue  $r_m = price * \overline{dyield}_m$ , where  $\overline{dyield}_m$  is detrended yield in each year. In the case of “without” contract, the revenue  $r_m = price * \overline{dyield}_m + I_{exp,m}(X_m) - P$ , where the premium  $P = 9\$/acre$  is equivalent to expected payoff in this paper.

Averages of the MRSLPC and VARPC of GDD-CRI models are provided in Table 2. The results show a significantly better performance of the phase-division contracts. For example, given  $\alpha = 0.0103$ , the average decrease in MRSL after purchasing the whole-cycle contract is 18.4%, while the average decline in MRSL for a phase-division contract is as much as 41.9%. Moreover, the average decrease in variance after purchasing the whole-cycle contract is 16.6% given  $\alpha = 0.0103$ , and the average decrease of the phase-division one is 37.8%.



**Table 2.** Average *MRSLPC* and *VARPC* of GDD-CRI models.

Whole-cycle	<i>MRSLPC</i> mean	Phase-division	<i>MRSLPC</i> mean
Exp-low	−17.9%	Exp-low	−41.9%
Exp-moderate	−18.3%	Exp-moderate	−41.9%
Exp-high	−18.4%	Exp-high	−41.9%
Whole-cycle	<i>VARPC</i> mean	Phase-division	<i>VARPC</i> mean
Exp-low	−16.2%	Exp-low	−37.8%
Exp-moderate	−16.6%	Exp-moderate	−37.8%
Exp-high	−16.6%	Exp-high	−37.8%

#### 4. Conclusions

This paper utilizes a nonlinear indemnity function framework to optimize weather index insurance. We suggest a GAM with P-splines box product smoothers to estimate conditional yield loss functions as a flexible alternative to commonly used regression models or neural networks that recently became popular for modeling weather–yield relations. The statistical framework is applied to the case study of soybean production in the US state of Illinois. Rainfall and temperature and their complex interaction are the main drivers of soybean yield losses. Our analysis contributes to decomposing the whole growth cycle into four stages. Our results indicate that the proposed phase-division contract outperforms the whole-cycle one regarding model fit and hedging effectiveness. This finding is in line with earlier studies emphasizing informational gains from disaggregated weather data (e.g., Schmidt *et al.*, 2022). The division of the growth cycle allows a state-dependent analysis of the collective impact of temperature and rainfall on crop yields. This is not only beneficial for the design of weather index insurance, but it may also support managerial decisions, such as the timing of irrigation measures (Katyal & Pandian, 2019). Moreover, the model performance gain is significant in terms of the percentage change of mean root square loss and VARs despite the relatively minor outperformance in EU ratio. The extent to which our results can be generalized to other regions, crops, and weather variables is suggested for future research.

We must address some limitations of our analysis. First, aggregated county-level soybean yield data are used instead of individual farm-level yield data due to data availability, which inevitably underestimates yield variability and basis risk of weather index insurance (Popp *et al.*, 2005). We conjecture that the relative advantage of the proposed statistical procedure will increase if it were applied to farm-level data with higher variability. Second, weekly state-level data from the *Crop Progress Report* cause temporal overlapping of plant growth phases. This dilutes the phase division and may result in an estimation error. Finally, the distribution of weather variables and the joint density of yield losses and weather indices have not been estimated, and thus the random nature of weather perils has not been explicitly considered. While this is not necessary for calculating the indemnity payments in our setting, it is essential for evaluating weather-related insurance products. Linking the proposed P-spline method with a parametric or nonparametric estimate of the density of weather indices would be an exciting direction for future research.

**Acknowledgments.** This research is supported by the China Scholarship Council. We thank the editor and two anonymous reviewers for their comments and suggestions that help improve the quality of this paper.

**Data availability statement.** The data used in our empirical analysis including yield, meteorological, and phenological data as well as computer codes (in R) of our calculations are provided at: [https://github.com/BaerchenJ/AAS\\_Plant-Growth-Stages-and-Weather-Index-Insurance-Design.git](https://github.com/BaerchenJ/AAS_Plant-Growth-Stages-and-Weather-Index-Insurance-Design.git)

## References

- Aydin, D. (2007). A comparison of the nonparametric regression models using smoothing spline and kernel regression. *World Academy of Science, Engineering and Technology*, **36**, 253–257.
- Barnett, B.J., Barrett, C.B. & Skees, J.R. (2008). Poverty traps and index-based risk transfer products. *World Development*, **36**(10), 1766–1785.
- Bokusheva, R. (2011). Measuring dependence in joint distributions of yield and weather variables. *Agricultural Finance Review*, **71**(1), 120–141.
- Boyer, J.S., Byrne, P., Cassman, K.G., Cooper, M., Delmer, D., Greene, T., Gruis, F., Habben, J., Hausmann, N., Kenny, N., Lafitte, R., Paszkiewicz, S., Porter, D., Schlegel, A., Schussler, J., Setter, T., Shanahan, J., Sharp, R.E., Vyn, T.J., Warner, D., & Gaffney, J. (2013). The US drought of 2012 in perspective: a call to action. *Global Food Security*, **2**(3), 139–143.
- Brumback, B.A. & Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**(443), 961–976.
- Bucheli, J., Dalhaus, T. & Finger, R. (2022). Temperature effects on crop yields in heat index insurance. *Food Policy*, **107**, 102214.
- Cao, X., Okhrin, O., Odening, M. & Ritter, M. (2015). Modelling spatio-temporal variability of temperature. *Computational Statistics*, **30**(3), 745–766.
- Chen, Z., Lu, Y., Zhang, J. & Zhu, W. (2020). Managing Weather Risk with a Neural Network-Based Index Insurance. Nanyang Business School Research Paper No. 20–28. <https://dx.doi.org/10.2139/ssrn.3539811>
- Conradt, S., Finger, R. & Spörri, M. (2015). Flexible weather index-based insurance design. *Climate Risk Management*, **10**, 106–117.
- Cooper, R.L. (2003). A delayed flowering barrier to higher soybean yields. *Field Crops Research*, **82**(1), 27–35.
- Cornell University Albert R. Mann Library (2022). United States Department of Agriculture (USDA) Economics, Statistics and Market Information System (ESMIS). Crop Progress. Available online at the address <https://usda.library.cornell.edu/concern/publications/8336h188j?locale=en&page=96#release-items> [accessed 21 July 2022].
- Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, **13**(11), 114003.
- Currie, I.D. & Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, **2**(4), 333–349.
- Dalhaus, T., Mußhoff, O. & Finger, R. (2018). Phenology information contributes to reduce temporal basis risk in agricultural weather index insurance. *Scientific Reports*, **8**(1), 1–10.
- de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.
- Delerce, S., Dorado, H., Grillon, A., Rebollo, M.C., Prager, S.D., Patiño, V.H., Varón G. G. & Jiménez, D. (2016). Assessing weather-yield relationships in rice at local scale using data mining approaches. *PLoS ONE*, **11**(8), e0161620.
- Donoghue, M.J. (2008). A phylogenetic perspective on the distribution of plant diversity. *Proceedings of The National Academy of Sciences of The United States of America*, **105**(supplement\_1), 11549–11555.
- Eilers, P.H., Currie, I.D. & Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, **50**(1), 61–76.
- Eilers, P.H. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–121.
- Eilers, P.H., Marx, B.D. & Durbán, M. (2015). Twenty years of P-splines. *SORT: Statistics and Operations Research Transactions*, **39**(2), 149–186.
- Ezekiel, M. (1930). *Methods of Correlational Analysis*. New York: Wiley.
- Hamed, R., Van Loon, A.F., Aerts, J. & Coumou, D. (2021). Impacts of hot-dry compound extremes on US soybean yields. *Earth System Dynamics Discussions*, **12**, 1–26.
- Hastie, T. & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, **82**(398), 371–386.
- Hatfield, J.L. & Prueger, J.H. (2015). Temperature extremes: effect on plant growth and development. *Weather and Climate Extremes*, **10**, 4–10.
- Hill, R.V., Kumar, N., Magnan, N., Makhija, S., de Nicola, F., Spielman, D.J. & Ward, P.S. (2019). Ex ante and ex post effects of hybrid index insurance in Bangladesh. *Journal of Development Economics*, **136**, 1–17.
- Jensen, M.E. (1968). Water consumption by agricultural plants. In: T.T. Kozłowski (Ed.), *Water Deficits and Plant Growth. Vol. II. Plant Water Consumption and Response* (pp. 1–22). New York & London: Academic Press.
- Katyal, N. & Jaganatha Pandian, B. (2020). A comparative study of conventional and smart farming. In *Emerging Technologies for Agriculture and Environment* (pp. 1–8). Singapore: Springer.
- Kellner, U. & Mußhoff, O. (2011). Precipitation or water capacity indices? An analysis of the benefits of alternative underlyings for index insurance. *Agricultural Systems*, **104**(8), 645–653.
- Kooperberg, C. & Stone, C.J. (1991). A study of log-spline density estimation. *Computational Statistics & Data Analysis*, **12**(3), 327–347.
- Lee, D.J. & Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**(1), 49–69.

- Lee, D.J., Durbán, M. & Eilers, P. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Computational Statistics & Data Analysis*, **61**, 22–37.
- Lin, J., Boyd, M., Pai, J., Porth, L., Zhang, Q. & Wang, K. (2015). Factors affecting farmers' willingness to purchase weather index insurance in the Hainan Province of China. *Agricultural Finance Review*, **75**(1), 103–113.
- Liu, F., Jensen, C.R. & Andersen, M.N. (2004). Drought stress effect on carbohydrate concentration in soybean leaves and pods during early reproductive development: its implication in altering pod set. *Field Crops Research*, **86**(1), 1–13.
- Marx, B.D. & Eilers, P.H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**(2), 193–209.
- Mußhoff, O., Odening, M. & Xu, W. (2011). Management of climate risks in agriculture – will weather derivatives permeate? *Applied Economics*, **43**(9), 1067–1077.
- Nielsen, D.C. & Nelson, N.O. (1998). Black bean sensitivity to water stress at various growth stages. *Crop Science*, **38**(2), 422–427.
- Odening, M., Mußhoff, O. & Xu, W. (2007). Analysis of rainfall derivatives using daily precipitation models: opportunities and pitfalls. *Agricultural Finance Review*, **67**(1), 135–156.
- Okhrin, O., Odening, M. & Xu, W. (2013). Systemic weather risk and crop insurance: the case of China. *Journal of Risk and Insurance*, **80**(2), 351–372.
- Okpara, J.N., Afiesimama, E.A., Anuforom, A.C., Owino, A. & Ogunjobi, K.O. (2017). The applicability of standardized precipitation index: drought characterization for early warning system and weather index insurance in West Africa. *Natural Hazards*, **89**(2), 555–583.
- Peiris, D.R., Crawford, J.W., Grashoff, C., Jefferies, R.A., Porter, J.R. & Marshall, B. (1996). A simulation study of crop growth and development under climate change. *Agricultural and Forest Meteorology*, **79**(4), 271–287.
- Pelka, N. & Mußhoff, O. (2013). Hedging effectiveness of weather derivatives in arable farming – is there a need for mixed indices? *Agricultural Finance Review*, **73**(2), 358–372.
- Popp, M., Rudstrom, M. & Manning, P. (2005). Spatial yield risk across region, crop and aggregation method. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroéconomie*, **53**(2-3), 103–115.
- Price, M.J., Yu, C.L., Hennessy, D.A. & Du, X. (2019). Are actuarial crop insurance rates fair?: an analysis using a penalized bivariate B-spline method. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**(5), 1207–1232.
- Qu, Y., Wang, K., Kang, J. & Liang, F. (2020). Effects of rainfall, temperature and illumination on outcrossing rate of male sterile line in soybean. *Oil Crop Science*, **5**(1), 17–21.
- Raviv, A. (1979). The design of an optimal insurance policy. *The American Economic Review*, **69**(1), 84–96.
- Reiss, P.T. & Todd Ogden, R. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 505–523.
- Rice, M., O'Neal, M. & Pedersen, P. (2007). Soybean Aphids in Iowa – 2007. Iowa State University, University Extension, SP 247.
- Rippey, B.R. (2015). The US drought of 2012. *Weather and Climate Extremes*, **10**, 57–64.
- Ritter, M., Mußhoff, O. & Odening, M. (2014). Minimizing geographical basis risk of weather derivatives using a multi-site rainfall model. *Computational Economics*, **44**(1), 67–86.
- Rodriguez-Álvarez, M.X., Lee, D.J., Kneib, T., Durbán, M. & Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, **25**(5), 941–957.
- Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**(4), 735–757.
- Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**(4), 719–727.
- Schierhorn, F., Hofmann, M., Gagalyuk, T., Ostapchuk, I. & Müller, D. (2021). Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. *Climatic Change*, **169**(3), 1–19.
- Schlenker, W. & Roberts, M.J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of The National Academy of Sciences of The United States of America*, **106**(37), 15594–15598.
- Schmidt, L., Odening, M., Schlanstein, J. & Ritter, M. (2022). Exploring the weather-yield nexus with artificial neural networks. *Agricultural Systems*, **196**, 103345.
- Schnitkey, G., Zulauf, C., Paulson, N., Swanson, K. & Baltz, J. (2022). 2021 Corn and Soybean Yields: Implications for Crop Insurance and Commodity Title payments.” farmdoc daily (12):37, Department of Agricultural and Consumer Economics, the University of Illinois at Urbana-Champaign. Available online at the address <https://farmdocdaily.illinois.edu/2022/03/2021-corn-and-soybean-yields-implications-for-crop-insurance-and-commodity-title-payments.html>.
- Scholtes, A.B., Sperry, B.P., Reynolds, D.B., Irby, J.T., Eubank, T.W., Barber, L.T. & Dodds, D.M. (2019). Effect of soybean growth stage on sensitivity to sublethal rates of dicamba and 2, 4-D. *Weed Technology*, **33**(4), 555–561.
- Shen, Z., Odening, M. & Okhrin, O. (2018). Adaptive local parametric estimation of crop yields: implications for crop insurance ratemaking. *European Review of Agricultural Economics*, **45**(2), 173–203.
- Shi, H. & Jiang, Z. (2016). The efficiency of composite weather index insurance in hedging rice yield risk: evidence from China. *Agricultural Economics*, **47**(3), 319–328.

- Stoppa, A. & Hess, U. (2003). Design and use of weather derivatives in agricultural policies: the case of rainfall index insurance in Morocco. In *International Conference "Agricultural Policy Reform and the WTO: Where are We Heading", Capri (Italy)*
- Tadesse, M.A., Shiferaw, B.A. & Erenstein, O. (2015). Weather index insurance for managing drought risk in smallholder agriculture: lessons and policy implications for sub-Saharan Africa. *Agricultural and Food Economics*, 3(1), 1–21.
- Tan, K.S. & Zhang, J. (2020). Flexible Weather Index Insurance Design with Penalized Splines. Nanyang Business School Research Paper, (21–13).
- Tang, J., Körner, C., Muraoka, H., Piao, S., Shen, M., Thackeray, S. J., & Yang, X. (2016). Emerging opportunities and challenges in phenology: a review. *Ecosphere*, 7(8), e01436.
- The Illinois Soybean Association (ISA). (2022). How much soy does Illinois produce? Available online at the address <https://www.ilsoy.org/faq-items/how-much-soy-does-illinois-produce/> [accessed 13 June 2023].
- Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S. & Wilson, B.E. (2020). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4*. Oak Ridge, Tennessee, USA: ORNL DAAC.
- Turvey, C.G. (2001). Weather derivatives for specific event risks in agriculture. *Applied Economic Perspectives and Policy*, 23(2), 333–351.
- United States Department of Agriculture (USDA). (2021). Rainfall Index Insurance Standards Handbook. Available online at the address <https://www.rma.usda.gov/-/media/RMA/Handbooks/Coverage-Plans—18000/Rainfall-and-Vegetation-Index—18150/2022-18150-Rainfall-Index-Handbook.ashx?la=en>.
- United States Department of Agriculture (USDA). (2022). Foreign Agricultural Service. US Soybeans Exports in 2021. <https://www.fas.usda.gov/commodities/soybeans> [accessed 21 July 2022].
- United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS). (2018). National Crop Progress—Terms and Definitions. Available online at the address [https://www.nass.usda.gov/Publications/National\\_Crop\\_Progress/Terms\\_and\\_Definitions/index.php](https://www.nass.usda.gov/Publications/National_Crop_Progress/Terms_and_Definitions/index.php).
- United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS). Quick Stats. Available online at the address <https://quickstats.nass.usda.gov/> [accessed 21 July 2022].
- van Munster, M., Yvon, M., Vile, D., Dader, B., Fereres, A. & Blanc, S. (2017). Water deficit enhances the transmission of plant viruses by insect vectors. *PLoS ONE*, 12(5), e0174398.
- Vedenov, D.V. & Barnett, B.J. (2004). Efficiency of weather derivatives as primary crop insurance instruments. *Journal of Agricultural and Resource Economics*, 29, 387–403.
- Walter, H. (1985). *Vegetation of the Earth and Ecological Systems of the Geo-Biosphere*. 3rd edition, Heidelberg: Springer.
- Wang, W., Bobojonov, I., Härdle, W.K. & Odening, M. (2013). Testing for increasing weather risk. *Stochastic Environmental Research and Risk Assessment*, 27(7), 1565–1574.
- Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95–114.
- Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3–36.
- Woodard, J.D. & Garcia, P. (2008). Basis risk and weather hedging effectiveness. *Agricultural Finance Review*, 68(1), 99–117.
- Wrather, A. & Koenning, S. (2009). Effects of diseases on soybean yields in the United States 1996 to 2007. *Plant Health Progress*, 10(1), 24.
- Xu, W., Odening, M. & Mußhoff, O. (2008). Indifference pricing of weather derivatives. *American Journal of Agricultural Economics*, 90(4), 979–993.
- Zhang, J., Tan, K.S. & Weng, C. (2019). Index insurance design. *ASTIN Bulletin: The Journal of the IAA*, 49(2), 491–523.

## Appendix A

### A.1. Results for the Rainfall Deficit Index (RDI)

For each phase, RDI is calculated as:

$$RDI_{m,l} = \sum_{\varphi=1}^{\pi} \min \left\{ 0, \sum_{t=(\varphi-1)*7+1}^{7*\varphi} R_{m,l,t} - R_{m,l,\varphi}^{\min} \right\}$$

where  $m$  and  $l$  represent the year and location,  $\varphi$  is the ordinal number of week at certain growth stage of Illinois soybeans,  $\pi$  is the total number of weeks within a certain growth phase,  $t$  is the  $t$ -th day in a year, and  $R_{m,l,\varphi}^{\min}$  represents the desired rainfall amount (“strike level”). We take the average rainfall amount per week in 2018 to determine this strike level, since the 96 counties reach the highest overall yield level in that year (cf. Fig. 1). The four stages include 7, 8, 8, and 7 weeks in 2018.

**Table A.1.** RMSE and adjusted  $R^2$  of GDD-RDI GAMs estimated by PS-ANOVA & `mgcv::gam()`.

PS-ANOVA					
Whole-cycle	RMSE	Adj. $R^2$	Phase-division	RMSE	Adj. $R^2$
Quadratic	5.58	0.24	Quadratic	4.16	0.52
Exp-low	5.24	0.24	Exp-low	4.16	0.52
Exp-moderate	5.25	0.24	Exp-moderate	4.16	0.53
Exp-high	5.24	0.24	Exp-high	4.16	0.53
gam() in "mgcv"					
Whole-cycle	RMSE	Adj. $R^2$	Phase-division	RMSE	Adj. $R^2$
Quadratic	5.34	0.21	Quadratic	4.49	0.44
Exp-low	5.34	0.21	Exp-low	4.49	0.45
Exp-moderate	5.34	0.22	Exp-moderate	4.49	0.45
Exp-high	5.34	0.22	Exp-high	4.49	0.45

For the GDD-RDI models, we again split the data set into three parts and select the suitable  $k$  values in the range [20, 40] by cross-validation. For the whole-cycle models, cross-validation resulted in  $k = 36$  for the one under quadratic utility and the one under exponential utility given  $\alpha = 0.0052$ ;  $k = 39$  for the models under exponential utility given  $\alpha = 0.008$  and  $\alpha = 0.0103$ . For all phase-division models, cross-validation resulted in  $k = 33$ . Comparing Table A.1 with Table 1, similar results indicate model robustness, which is again reflected in the comparison between EU ratio means in Table A.2.

**A.2. EU ratio results comparison**

**Table A.2.** Average EU ratios of GDD-CR and GDD-RDI contracts.

GDD-CR	EU ratio mean
Exp-low	1.007
Exp-moderate	1.017
Exp-high	1.022
GDD-RDI	EU ratio mean
Exp-low	1.013
Exp-moderate	1.021
Exp-high	1.026