

Challenges in Actuarial Learning for Modeling of Brazilian Soybean Crops

Eduardo F. L. Melo

FGV, School of Applied Mathematics, Rio de Janeiro, Brazil

Susep - Superintendence of Private Insurance, Rio de Janeiro, Brazil

UERJ - State University of Rio de Janeiro, Brazil

`eduardo.melo@fgv.br`

Helton Graziadei

FGV, School of Applied Mathematics, Rio de Janeiro, Brazil

`helton.carvalho@fgv.br`

Rodrigo S. Targino

FGV, School of Applied Mathematics, Rio de Janeiro, Brazil

`rodrigo.targino@fgv.br`

September 21, 2026

Abstract

In Brazil, the agricultural sector plays a fundamental role, accounting for approximately 25% of the national Gross Domestic Product (GDP). Within this sector, soybean cultivation stands out as the most significant, with a forecast yield exceeding 150 million tons in the 2022/2023 harvest, which is equivalent to the combined production of all Organisation for Economic Co-operation and Development (OECD) nations. Nevertheless, soybean production is subject to considerable climatic vulnerabilities that not only affect the yield but also have a considerable impact on the formulation of crop insurance pricing. The consequences of drought and excessive rainfall in Brazil's South and Central-West regions severely affected the crop yield of 2021/2022, leading to a substantial rise in the loss ratio of insurer's portfolios. These climatic adversities led to a substantial rise in the loss ratio of insurers' portfolios. In this paper, we utilize actuarial learning models to predict losses in soybean crops,

taking into account weather covariates, specifically rainfall, and temperature, given their relevance to crop productivity. Both generalized linear models and random forest have been benchmarked to assess their predictive performance within the two-stage frequency-severity framework. This framework comprises a model dedicated to claim occurrence and a second model focused on determining the severity of individual claims. We also compare the charged premium with the estimated premium by using standard insurance modeling techniques. Additionally, we carry out simulated stress tests to highlight the significance of weather-related variables in evaluating total losses during extreme events. Our assessment encompasses the estimation of losses across diverse scenarios. The implications of these results are significant for (re)insurance pricing, risk management, and solvency. They also bear substantial importance for formulating effective agricultural public policies.

1 Introduction

The agricultural sector is extremely important for Brazil, contributing to approximately 25% of the Brazilian GDP in 2022¹. This sector generated 2.65 trillion BRL (approx. 475 billion USD) through agribusiness, thereby highlighting the importance of sustained productivity for the overall Brazilian economy. As a consequence, any downturn in production directly impacts the income of this critical economic sector. Based on data from the National Supply Company (CONAB)², the planted area generated a total harvest of 312.5 million tons of grains in 2022/2023, which are destined to supply the national and international markets. In drawing a comparison with soybean yields, which is the predominant crop, Brazil's projected output mirrored the total production of all OECD nations, reaching an impressive 153.6 million tons in 2023, as per the same source. In the context of the war in Eastern Europe, rising prices, inflation, and extreme weather, agricultural production in Brazil becomes even more relevant for the whole world.

Given its relevance, adequate insurance coverage is essential for this economic activity. Agricultural insurance plays a critical role in mitigating losses for farmers, making it a crucial component of the country's agriculture and economy. By definition, agricultural insurance is a contract established between a farmer and an insurance institution. Upon the payment of a premium, it provides coverage against claims that affect the insured crops. The risk in-

¹<https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx>

²<https://www.conab.gov.br/ultimas-noticias/4971-producao-de-graos-esta-estimada-em-312-5-milhoes-de-toneladas-na-safra-2022-23>

herent in the agricultural insurance portfolio is tied to losses resulting from weather events and pests, both of which can severely affect crop productivity. The Food and Agriculture Organization of the United Nations (FAO) has highlighted the importance of this insurance for enhancing food production (FAO, 2020, 2021), positioning it as an efficient tool to increase the resilience of rural producers against large-scale events that impact the rural population. Providing insurance against disaster risks in agriculture has the potential to fortify the social safety net, offering protection against a broad spectrum of catastrophic events. Thus, agricultural insurance serves not only as a safety measure for farmers but also as a key player in ensuring the stability and growth of the national agricultural economy. For agricultural operating loans in Brazil, the presence of crop insurance is essential. In fact, the acquisition of insurance coverage is mandatory for agricultural operating loans valued up to R\$300,000 in Brazil.

This means that in 2022, for every 100 units of currency received in premiums, the insurance companies had to pay out about 281.6 units due to crop damages, which is more than three times the amount paid out in the previous year. This increase highlights the significant impact of adverse weather conditions on agriculture in these regions of Brazil."

In the most agriculturally productive regions of Brazil, encompassing both the south and middle-west, recent droughts and excessive rainfall have caused a substantial increase in loss ratios for the crop insurance portfolio. According to the Brazilian regulatory agency SUSEP (Superintendência de Seguros Privados)³, there has been a dramatic rise in the loss ratio for crop insurance. For the first half of 2022, this loss ratio reached approximately 281.6%, a sharp increase from 89.7% during the same period in 2021. This means that in 2022, for every 100 units of currency received in premiums, the insurance companies had to pay out about 281.6 units due to crop damages, which is more than three times the amount paid out in the previous year. This increase highlights the significant impact of adverse weather conditions on agriculture in these regions of Brazil. These years have witnessed significant losses in agricultural insurance, inciting concern among insurers and reinsurers. The pervasive effects of global warming and climate change are becoming increasingly apparent. In agricultural insurance, these changes are reflected in the heightened number of claims and the severity of losses. In this context, it is worth noting the severe losses recently experienced in soybean crops, primarily due to drought and excessive rainfall.

³https://www2.susep.gov.br/menuestatistica/ses/resp_premiosesinistros.asp

One important consequence of this catastrophic scenario is that the country may face a problem of insurance supply shortage. Reinsurers, which historically have assumed most of the insured risks, have retracted sharply in offering contracting limits, which are essential for products to be offered to policyholders⁴⁵. In certain regions of Brazil, it is anticipated that demand may be greater than supply, resulting in a proportion of producers unable to secure insurance for their crops. This constitutes an unfavorable scenario considering the relevance that agribusiness has in the national and international economies.

In recent years, most insurance companies that operate agricultural insurance have identified the need for more sophisticated models to correctly estimate risks and manage them, developing tools for quantifying technical provisions and also for pricing. A more accurate and adequate risk modeling brings more security to the supply side, that is, for insurers and reinsurers, reducing uncertainty in the estimation process and lowering risk margins in premiums, thus increasing penetration and insurance coverage.

In the field of agricultural insurance, a significant body of literature has been dedicated to the development of methodologies for effectively predicting crop insurance claims. Among these, the study conducted by Mota et al. (2020), has primarily focused on predicting the occurrence of claims by using machine learning models. However, there remains a crucial gap in forecasting the financial magnitude of these claims. In contrast, our work extends these methodologies to estimate not only the likelihood but also the potential loss amount of claims. By leveraging regression models and machine learning techniques, we aim to provide a more comprehensive prediction model that helps insurers better allocate resources and design policies that accommodate the actual risk exposure.

In this paper, predictive actuarial learning models are designed to estimate insurance losses for insured and subsidized areas, utilizing databases from a variety of sources which includes the Ministry of Agriculture, the Climate Research Unit at the University of East Anglia (providing meteorological data such as precipitation and temperature, collected from meteorological stations throughout the country, by location – latitude and longitude), National Water Agency from Brazil (soil humidity) and Columbia Climate School data (El Niño and La Niña occurrence indices). These models may provide increased accuracy in pricing as well as a more adequate risk and solvency assessment. Specifically, the proposed models focused on the most relevant culture in Brazil

⁴<https://valor.globo.com/agronegocios/noticia/2023/06/16/perdas-elevadas-reduzem-apetite-de-seguradoras.ghtml>

⁵<https://globorural.globo.com/negocios/noticia/2023/06/aumento-de-indenizacoes-leva-empresas-a-deixar-de-ofertar-seguro-rural.ghtml>

– soybeans. Over the past years, soybean yields have seen considerable declines due to severe droughts and excessive rainfall. These climatic challenges primarily impacted Brazil’s most fertile agricultural areas, notably the southern states of Rio Grande do Sul and Paraná, as well as the central-western states of Mato Grosso and Mato Grosso do Sul.

The results obtained will be highly pertinent to the formulation of public policies, including federal government subsidies, and for designing agricultural catastrophe funds. Therefore, this work aims to generate crucial insights with potential ramifications for both the Brazilian population and the global community, considering the significant role grain production plays in the global supply chain.

We present findings on crop insurance pricing across municipalities. Following a similar methodology described in Liu and Ramsey (2023), we utilize data from the preceding crop year’s weather-related variables to calculate premiums, ensuring a detailed assimilation of these factors. Also, by incorporating total municipal production data, which includes both insured and non-insured areas, we estimate potential total losses in municipalities. Furthermore, we execute simulated stress tests to underscore the influence of weather-related variables and gauge total losses in extreme scenarios.

The structure of the paper is given as follows. Section 2 describes the data used in the study. Section 3 introduces the construction of the variables and the specification of the model utilized in this paper. The extrapolation method used to estimate losses as well as the results of the stress test under different weather scenarios are also described. Section 4 is dedicated to premium calculation techniques and comparative study with the charged premium. Finally, Section 5 provides concluding remarks and insights.

2 Data

In this section, we provide an overview of the datasets employed in the modeling process. To facilitate a clear understanding, we will begin by introducing the crop insurance data, followed by an exploration of the weather and soil variables utilized in our analysis.

2.1 Crop Insurance data

We obtained the crop insurance policies data from the Agricultural Ministry’s website⁶. This dataset was downloaded on January 18th, 2023. The policies

⁶<https://dados.agricultura.gov.br/dataset/sisser3>

are included in the Crop Insurance Subvention Program (PSR), a federal initiative providing subsidies for a portion of insurance premiums for farmers looking for coverage. The primary goal of PSR is to provide financial assistance for the purchase of insurance policies by covering a part of the insurance premiums charged to rural producers. This is crucial because agriculture is highly susceptible to unpredictable natural events, which can cause significant losses. Under the PSR, the subsidy percentage and the maximum subsidizable amount per producer vary depending on the activity. The program also sets financial limits for subsidies per producer, which can vary based on the modality. For agricultural activities, the limit can be up to *R*\$60,000 each. If a producer is involved in multiple modalities, they can potentially receive a maximum subsidy of *R*\$120,000 per year ⁷. Producers interested in accessing this subsidy do not need to apply directly. Instead, the insurance company, after finalizing terms with the producer, submits the insurance policy for review by MAPA’s technical team. If approved and if there are available funds, the subsidy is granted. The subsidy amount is then forwarded directly to the insurance company and deducted from the total amount due from the producer. We restricted our analysis to contracts initiated between March 2016 and February 2022 and selected only yield-based contracts. Furthermore, we focused on policies of soybean crops, which represent 47.28% of all policies in the period analyzed, as depicted in Figure 1. A full description of the main variables contained in this data are given in Table 1.

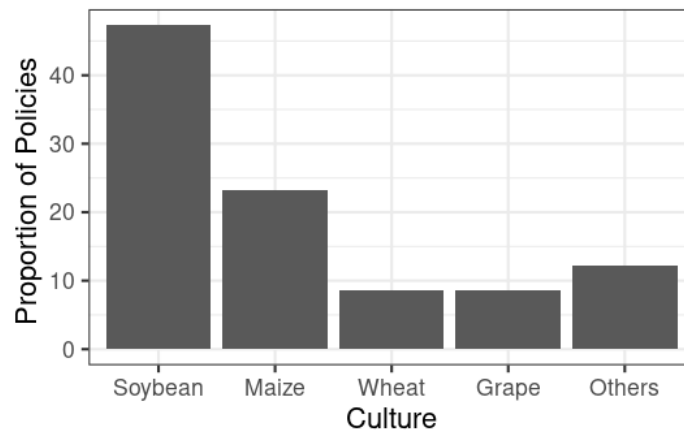


Figure 1: Proportion of policies by culture in the crop insurance dataset.

We define crop year as covering all policies with start dates ranging from March of one year to February of the subsequent year. As an example, the first crop year considered in this study spans from March 2016 to February 2017.

⁷<https://www.gov.br/agricultura/pt-br/assuntos/riscos-seguro/seguro-rural/limites-percentuais-de-subvencao>

This delineation aligns with customary planting schedules, which typically initiate in October and culminate in December⁸. The policies whose start date is between March and October accounts for approximately 95% of the policies.

Variable	Description
Latitude	Latitude of the farm
Longitude	Longitude of the farm
Municipality	Name of the municipality in which the farm is located
ContractDate	Contract start date
State	Name of the state in which the farm is located
Area	Insured area in hectares (ha)
Yield	Insured yield (in kg/ha)
ClaimValue	Claim value in Brazilian reais (BRL)
Premium	Charged premium in BRL (includes the federal subvention)
Subvention	Federal subvention value (in BRL)

Table 1: Variables in the crop insurance dataset obtained from the Ministry of Agricultural from Brazil.

Figure 2 indicates that drought is responsible for 73.0% of the claims related to soybean crops. The second main cause is excessive rain, representing 15.0% of the claims. These figures emphasize the importance of our further examination of these types of claims. While hail and flood are less common, they represent 5.78% and 4.86% of the total claims, respectively. All other claim causes make up the remaining 1.31% of the claims.

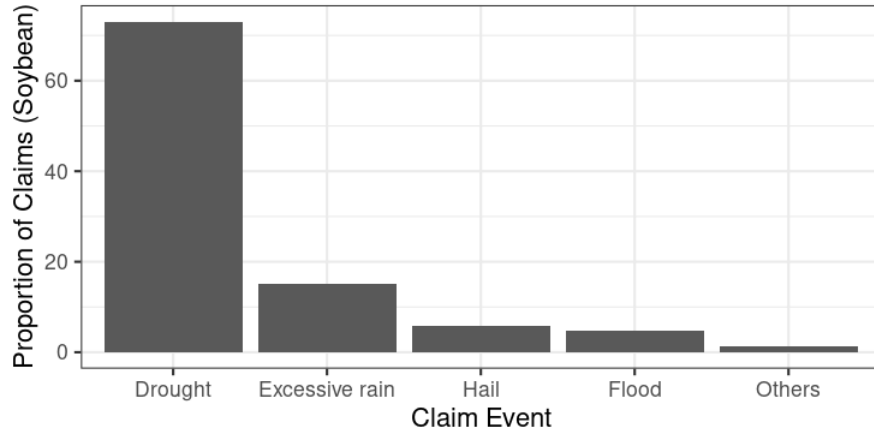


Figure 2: Proportion of soybeans claims by claim cause.

⁸https://www.conab.gov.br/outras-publicacoes/item/download/45710_77f0a5ee35e765e2248999f4c61e70c8

2.2 Weather and soil data

Several studies underscore the importance of historical weather data, especially precipitation and temperature, in determining crop insurance yield and, as a consequence, influencing insurance losses. For instance, Liu and Ramsey (2023) employed Bayesian methodologies to predict losses based on these weather variables. Similarly, Zhu et al. (2019) highlighted the role of weather data in improving the accuracy of crop yield predictions. Furthermore, Berlato and Fontana (2001) investigates the effects of the ENSO phenomena on rainfall and relates them to the specific requirements of crops.

The precipitation and temperature data are sourced from the Climate Research Unit gridded Time Series (see Harris et al. (2020)) dataset from the University of East Anglia. This dataset compiles monthly observations from land-based meteorological stations and maps them onto a 0.5° regular grid utilizing angular distance weighting. This particular dataset is renowned for its extensive use across diverse domains and applications, especially in models that rely heavily on localized weather and climate data, as discussed in Harris et al. (2020).

The maps depicted in Figures 8, 9, 10, and 11 showcase the rainfall patterns across different Brazilian municipalities for the crop years spanning from 2016/2017 to 2021/2022. Notably, the 2021/2022 crop year recorded lower cumulative precipitations in southern Brazil. Conversely, Figures 12, 13, 14, and 15 depict temperature variations during the same period. Particularly in December and January of the 2021/2022 crop year, southern Brazil experienced some of the highest temperatures compared to preceding crop years.

The phenological soybean cycle can be broadly segmented into four phases: establishment (≈ 12 days); vegetative phase (≈ 40 days); reproductive phase (≈ 40 days); and maturation (≈ 28 days). Zanatta (2019) observed that soybean crops are particularly vulnerable to water scarcity during the reproductive phase. As outlined in Section 2.1, soybean planting in Brazil usually begins between October and December. Furthermore, the Climate Risk Agricultural Zoning (ZARC)⁹ recommends this same period — October to December — for most municipalities as the optimal window for sowing soybean seeds. A limitation of the crop insurance data, described in 2, is the absence of individual farmer planting dates. Nonetheless, considering that the soybean’s most water-sensitive period initiates approximately 52 days post-seeding, we utilize the cumulative monthly rainfall data from November to February as a representative gauge for when the soybean crops are most susceptible to water

⁹<https://mapa-indicadores.agricultura.gov.br/publico/extensions/Zarc/Zarc.html>

shortages.

Several studies have consistently highlighted a relationship between the climatic variability of the El Niño-Southern Oscillation (ENSO) phenomenon and anomalies in precipitation (Berlato and Fontana, 2001; Diaz et al., 1998). Specifically, for the southern, southeastern, and southern central-western regions of Brazil, El Niño tends to generate precipitation that surpasses the climatological average, while La Niña typically results in rainfall that is below this average (Berlato and Fontana, 2001). The most pronounced effects of both El Niño and La Niña on rainfall are observed in the spring and early summer — coincidentally, this is also the period when soybeans have the highest water requirements. Conversely, the ENSO’s warm phases (El Niño) lead to decreased rainfall in the northeastern and north-central western parts of Brazil. In contrast, during La Niña episodes, these regions experience elevated precipitation levels.

In this context, we consider two key indices: the Oceanic Niño Index (ONI), and the Southern Oscillation Index (SOI). The ONI is defined as the three-month moving average of temperature anomalies—deviations in the surface waters of the east-central tropical Pacific, near the International Dateline. In summary, when ONI exceeds a threshold of 0.5 or rises above it, it suggests the occurrence of an El Niño event. Conversely, a value falling beneath -0.5 indicates the occurrence of La Niña (see Banholzer and Donner (2014)). In contrast, the SOI is calculated from the standardized observed sea level pressure (SLP) differences between Tahiti and Darwin, Australia. An SOI value below -1.5 is indicative of El Niño conditions, whereas a value surpassing 1.5 points to La Niña. The time series for SOI and ONI are provided in Figure 17. In our proposed model, we incorporate the number of times in which SOI and ONI exceed 1.5 (or fall below -1.5) within a specific crop year as covariates to represent the intensity and continuity of El-Niño/La-Niña. These counts serve as covariates, representing the intensity and persistence of El Niño and La Niña events.

We also took into account the available water capacity (AWC), which is a measure of the soil’s ability to hold water for plants. Soil humidity plays a crucial role in various fields such as agronomy and hydrology, influencing irrigation planning, water balance studies, plant water availability assessments, infiltration analyses, drainage conditions, and more. Typically, two critical benchmarks are used to determine the range of moisture that plants can use from the soil: the field capacity (FC), which is the upper limit, and the permanent wilting point (PWP), the lower threshold. The difference between these two values gives the AWC. Essentially, the AWC represents the maximum vol-

ume of water plants can use when the soil’s moisture level is at field capacity, excluding the moisture deemed unusable at the permanent wilting point. This dataset, maintained by Agencia Nacional de Águas (ANA) and Universidade Federal do Paraná¹⁰, provides AWC data for Brazilian municipalities in 2021. The AWC data is visually represented in Figure 16.

3 Modeling

In this section, we outline the modeling process for predicting insurance losses in soybean crops. We start by defining the response variables representing municipality losses and the covariates for prediction. We then present several choices for the two-stage models, consisting of generalized linear models, generalized additive models, and random forests. Next, we discuss the evaluation criteria used to compare the frequency and loss models. Finally, we cover graphical techniques for interpreting our proposed models.

3.1 The Actuarial Model

For developing the actuarial model, we aggregate the original dataset, described in Section 2, to the municipality level. First, let $A_{k,j}$ be the insured area for farm k in crop year j , and $P_{k,j}$ denote the insured crop yield for farm k in crop year j . Since the claim values stem from different crop years, to account for claims inflation, we adjust these amounts based on the date on which the soybean producer contracted the insurance product. Specifically, S_k stands for the soybean spot price in BRL on the contract’s signing date, while S_J is the soybean spot price in BRL as of December 30th, 2022. This date serves as our reference point, marking the final working day of December 2022. Hence, the aggregated loss variable L_{ij} is defined as

$$L_{i,j} = \frac{\sum_{k \in i} Y_{k,j} \frac{S_J}{S_k}}{\sum_{k \in i} A_{k,j} P_{k,j}},$$

where $Y_{k,t}$ represents the claim amount for insured and subsidized farm k located in municipality i during crop year j , $L_{i,j}$ is the loss for municipality j in year-crop t relative to the insured production. We also define $Z_{i,j}$ which is an indicator of at least one claim in municipality i at crop year j . Hence, the set of variables we consider for modeling is given in Table 2.

Once our loss variable is defined, assume the following data-generating

¹⁰https://metadados.snirh.gov.br/geonetwork/srv/api/records/28fe4baa-66f3-4f6b-b0d2-890abf5910c4/attachments/CAD_AWC_media_municipal.xlsx

Variable	Description
Municipality	Name of the municipality
Crop	Crop year, numbered from 1 to 6
Latitude	Latitude of the municipality
Longitude	Longitude of the municipality
AWC	Available water capacity of the municipality
PrecNov	Total precipitation (in mm) in November of the crop year.
PrecDec	Total precipitation (in mm) in December of the crop year.
PrecJan	Total precipitation (in mm) in January of the crop year.
PrecFev	Total precipitation (in mm) in February of the crop year.
TempNov	Average temperature (in Celsius) in November of the crop year
TempDec	Average temperature (in Celsius) in December of the crop year
TempJan	Average temperature (in Celsius) in January of the crop year
TempFev	Average temperature (in Celsius) in February of the crop year
ElNinoCount	The number of times the SOI index fell below -1.5 during the crop year
LaNinaCount	The number of times the SOI index exceeded 1.5 during the crop year
FlagClaim	Flag indicating there was at least a claim in the municipality
RelativeLoss	Relative loss, ratio of deflated total claims amounts to production

Table 2: Variables in the crop insurance dataset (municipality-level).

process $\mathcal{D} = \{(X_{ij}, Z_{ij}, L_{ij}) : i = 1, \dots, n_j, j = 1, \dots, J\}$, where n_j is the number of municipalities with a positive insured area in crop year j and J is the total number of crop years. Also, $X_{ij} \in \mathbb{R}^d$ is a vector of predictors in municipality i at crop year j , $Z_{ij} \in \{0, 1\}$ is equal to 1 in case the municipality has at least one claim in crop year j and 0, otherwise, and L_{ij} is the relative loss defined in 3.1. Our primary goal is to predict expected losses L_{ij} based on covariates X_{ij} . This is estimated by the conditional expectation $E(L_{ij} | X_{ij})$ which can be factored, by the law of total expectation, as

$$E[L_{ij} | X_{ij}] = E[L_{ij} | X_{ij}, Z_{ij} = 1] \cdot P(Z_{ij} = 1 | X_{ij}), \quad (1)$$

under the assumption that $E(L_{ij} | X_{ij}, Z_{ij} = 0) = 0$. This assumption simply states that if there are no claims in the municipality i at crop year j , then the relative loss is equal to zero.

Hence, from (1), we begin by fitting a classification model for Z_{ij} given X_{ij} , using the entire dataset. Next, we fit a regression model for L_{ij} restricted to instances where $Z_{ij} = 1$. Hence, we consider the following two-stage model hierarchy:

$$\begin{aligned} Z_{ij} | \mathbf{X}_{ij} &\sim F_{ij} \\ Y_{ij} | Z_{ij}, \mathbf{X}_{ij} &\sim G_{ij} \end{aligned}$$

In the subsequent subsections, we review the statistical and machine learning models employed for predicting frequency and aggregate losses. First, we discuss the generalized linear model approach, renowned for its widespread application. Following this, we examine decision trees and random forests, which have the capability to account for interactions and non-linearities.

3.1.1 Generalized Linear Models (GLM)

We start considering a generalized linear model framework (Nelder and Wedderburn, 1972), where F_{ij} is the cumulative distribution function (cdf) a $Binomial(\theta_{ij})$ in which

$$g_1(\theta_{ij}) = X_{ij}^T \beta,$$

where $g_1(x) = \frac{e^{-x}}{1+e^{-x}}$ is a logistic link function, and β is a d -dimensional parameter vector. This specific type of generalized linear model is commonly referred to as the *Logistic regression model*. Also, G_{ij} is assumed to be the cdf of a $Gamma(\mu_{ij}, \phi)$, where

$$g_2(\mu_{ij}) = X_{ij}^T \gamma,$$

with $g_2(x) = \log(x)$ is a logarithmic link function and γ is a d -dimensional parameter vector and $\phi > 0$. This type of generalized linear model is referred to as the *Gamma regression model* with a logarithmic link function.

3.1.2 Generalized Additive Models (GAM)

To account for interactions and non-linear effects, we fitted a generalized additive model for both classification and regression stages, where we replaced the linear predictor by

$$g_1(\theta_{ij}) = \sum_{k=1}^d f_k^{(1)}(X_{ij})$$

$$g_2(\mu_{ij}) = \sum_{k=1}^d f_k^{(2)}(X_{ij})$$

Here, $f_k^{(1)}$ and $f_k^{(2)}$, where $k = 1, \dots, d$, represent a set of penalized cubic splines (Hastie et al., 2009). Also, g_1 and g_2 are the same as in the generalized linear model, that is, the logistic and logarithmic link functions, respectively. These splines allow us to flexibly model the non-linear relationships and interactions present in the data, enabling us to capture more complex patterns beyond

linear relationships.

3.1.3 Random Forests

A second strategy for tackling interaction and non-linear effects is to employ a random forest model. The foundation of this method lies in the development of regression trees by the Classification and Regression Trees (CART) algorithm (Breiman et al., 1984). This algorithm systematically partitions the predictor variable space through a recursive process, seeking to identify optimal splits in the training data by greedily minimizing a quadratic cost function.

The process starts by repeatedly drawing samples with replacement from the training dataset, generating what is known as a bootstrap sample (Efron and Tibshirani, 1994). This bootstrap sample serves as the basis for constructing individual tall regression trees using the CART algorithm. By repeating this procedure, a collection of standalone regression trees is obtained.

To enhance the predictive power further, these individual trees are aggregated using a technique called bagging (Breiman, 1996). In bagging, multiple regression functions are trained from bootstrap samples, and their predictions are averaged to create a composite regression function. To optimize this aggregation process, a refinement is introduced. At each decision point within each regression tree of the ensemble, a random subset of predictors is selected without replacement. This enhancement gives rise to what is commonly referred to as a *Random Forest* (Breiman, 2001).

We applied a random forest model for both classification and regression, with 10^4 trees and set the number of predictors considered at each split of the trees¹¹ to be equal to the square root of the number of predictors.

3.2 Model evaluation

Here we compare the results of the three proposed models.

To assess the predictive performance of these algorithms, we employed an iterative procedure, described as follows. In each iteration, the training set includes policies from the crop year 1 to the crop year $i - 1$, while the test set includes policies from the crop year i . This procedure is carried out for $i = 2, \dots, 6$, where $i = 2$ is the 2017/2018 crop year, and $i = 6$ is the 2021/2022 crop year. Next, we measure the area under the ROC curve (Hastie et al., 2009) for the claims classification model and the mean absolute error for the loss model. A visual representation of this iterative method can be found in Table 3.

¹¹`mtry` in the R package `randomForest` or `max_features` in Python's `sklearn`

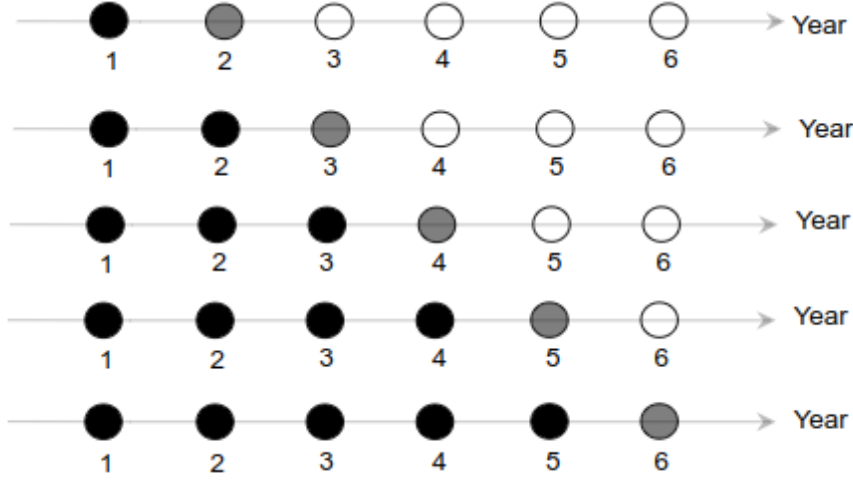


Figure 3: Illustration of the iterative model assessment procedure. The black nodes indicate the crop years considered in the training set, the gray nodes represent the crop year included in the test set and the white nodes mark crop years omitted in each iteration.

Table 8 displays the average of the areas under the ROC curve for the classification model, as well as this metric only for the last iteration of the procedure (crop 2021/2022 as the test set). For both predictive performance measures, the Random Forest demonstrated superior performance over the Logistic Regression model. Furthermore, Table 9 displays the average of the mean absolute errors for the loss model, taken across the various test sets. It also provides, in the third column, the mean absolute error from the final iteration. In line with the results of the classification model, the Random Forest algorithm also surpassed the performance of the Gamma regression. Therefore, we utilize a first random forest model to compute claims probabilities, followed by a second Random Forest to the loss stage.

3.3 Interpretability

We verify the importance of the predictors as well as the impact of weather variables for predicting losses.

To delve into the black box of the chosen loss model, we first verify the importance of each variable in predicting losses. In the context of a regression problem with the Random Forest model, feature importance is typically computed based on the average decrease in the model’s performance (often measured as the mean squared error) when a particular feature’s values are randomly permuted. This method offers an intuitive approach: if a variable is important for accurate predictions, permuting its values should significantly

lead to a decrease in the predictive performance, otherwise, its permutation should have little to no effect on the performance. The importance of the random forest algorithm is given in Figure 7. This indicates the importance of the rainfall in December to predict losses on soybean crops. Unsurprisingly, December coincides with the phase wherein the majority of soybean crops are in the maturation phase, the most susceptible period to water scarcity. Also, latitude is the second most important variable which is also consistent with expectations, considering that the southern region of Brazil is associated with higher-risk municipalities.

Also, to understand in greater detail the marginal contribution of each weather variable on the predictive loss model, we explore partial dependence plots (Friedman, 2001; Zhao and Hastie, 2021). The partial dependence plot (PDP) offers a visual representation of the relationship between a predictor variable and the outcome of interest, after accounting for the average effects of all other variables in the model (see figure 7). We evaluate the PDPs for precipitation variables in December and January, which are the most important weather variables in the loss model. A close examination of the PDP reveals a pronounced trend, illustrating a non-linear correlation between relative loss and rainfall. As December’s rainfall volume rises, there’s a decline in the projected relative loss. However, once rainfall surpasses 250mm, the relative loss begins to ascend. This pattern indicates that moderate rainfall in December, up to a certain limit, might act as a buffer against losses, perhaps due to advantages like ideal soil moisture, reduced irrigation demands, or other agronomic conditions that enhance crop health and yield. However, rainfall levels surpassing 250mm monthly appear to amplify losses. It is crucial to underscore that the loss variable incorporates claims related to both drought and excessive rainfall. In Figure 18, one can observe the joint PDP delineated with level curves. This representation reaffirms the conclusions previously drawn. However, we need to treat this finding with circumspection. While the PDP can indicate a relationship between precipitation and relative loss, causation cannot be directly inferred. Further studies or domain knowledge might be required to elucidate the underlying mechanisms driving this relationship.

4 Premium Calculation and Extrapolation

This section outlines our premium calculation methods for estimating the pure premium of Brazilian municipalities. We also demonstrate a method for extrapolating losses, based on the homogeneity assumption, to encompass all soybean crop farmers.

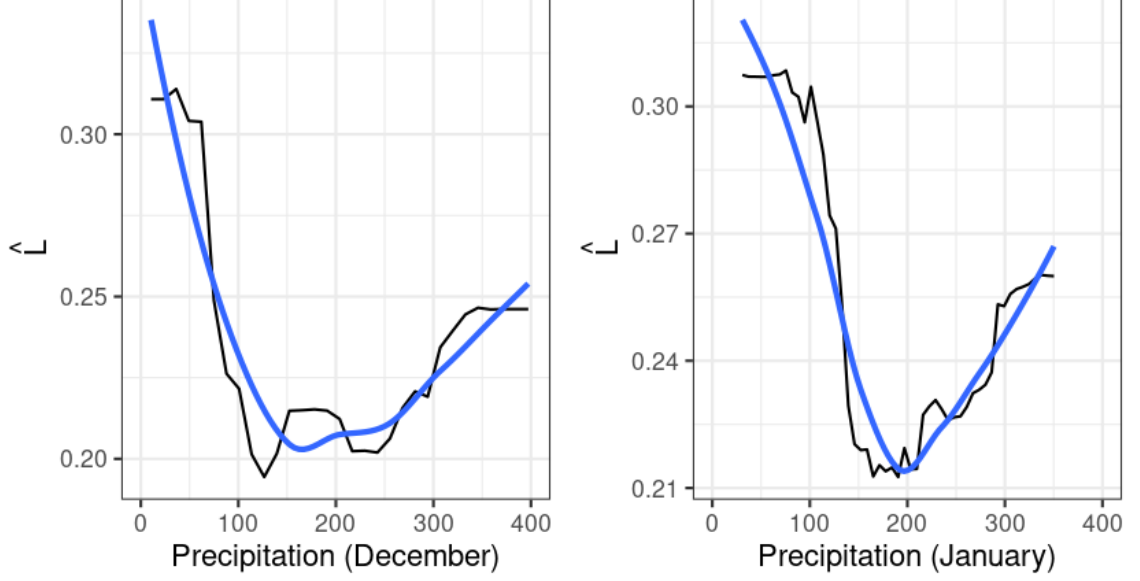


Figure 4: Partial dependence plots showing the average predictions of the relative losses based on December and January total precipitation values (with smoothed curves in blue).

4.1 Premium Calculation

Essentially, this involves averaging over multiple climate scenarios, taking into account the likelihood of each one of them. Consider $X_{ij} = (W_{ij}, X_i^*)$ as the covariate vector for a municipality i during crop year j . We denote by W_{ij} the vector of dynamic and weather-related covariates, while X_i^* denotes the vector of static covariates whose values are constant across crop years. Then, we approximate the expected value of the relative loss in municipality i in crop year j using the following conditional expectation:

$$E(L_{iJ} | X_i^*) = E\left[E(L_{iJ} | X_i^*, W_{iJ})\right]$$

which can be approximated by the empirical average

$$E(L_{iJ} | X_i^*) \approx \frac{1}{J-1} \sum_{j=1}^{J-1} \hat{E}[L_{ij} | X_i^*, W_{ij}, Z_{ij} = 1] \cdot \hat{P}(Z_{ij} = 1 | X_i^*, W_{ij})$$

Thus, this approximation enables us to determine the expected indemnity for a particular municipality at the onset of a crop year, as it factors out the climatic variables that are not available when setting the price. Figure 5 shows the QQ-plots comparing the estimated pure premium and total indemnity distributions in the 2021/2022 crop year, by using climate data from the previous crop years (2017/2018 until 2020/2021). Utilizing climate data from the 2018/2019 and

Variable	Available at the beginning of the crop year?
Municipality	Yes
Crop	Yes
Latitude	Yes
Longitude	Yes
AWC	Yes
PrecNov	No
PrecDec	No
PrecJan	No
PrecFev	No
TempNov	No
TempDec	No
TempJan	No
TempFev	No
ElNinoCount	No
LaNinaCount	No
FlagClaim	No
RelativeLoss	No

Table 3: Indication of variables available at the beginning of the crop year.

2019/2020 crop years tends to elevate the estimated premium. In contrast, data from the 2017/2018 and 2020/2021 crop years often yield more precise premium estimates.

In our subsequent analysis, we derived estimated premiums based solely on weather data from the 2020/2021 crop years. We also assess the loss ratio using both the charged and the estimated premiums. Let Y_{ij} be the total observed loss in municipality i during crop year j . Then, the loss ratios are defined as follows:

$$LR_j^{(charged)} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} \pi_{ij}},$$

for $j = 1, \dots, J$ and $\pi_{ij}^{(charged)}$ is the total charged premium in municipality i in crop year j . Moreover,

$$LR_j^{(estimated)} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} \hat{\pi}_{ij}},$$

is the *loss ratio* for the estimated premium. In the 2021/2022 crop year, the loss ratio for charged premiums was $LR_j^{(charged)} = 0.4399$, whereas the estimated premium resulted in a loss ratio of $LR_j^{(estimated)} = 0.9973$. This illustrates that the estimated premiums exhibited a more favorable alignment with observed

indemnity compared to the charged premiums.

4.2 Extrapolation of Losses

Since the random forest demonstrated superior prediction performance in classification and regression tasks, we employed these models for the following analyses. We propose a method to estimate losses across the entire planted area within a municipality. For this purpose, assume that the relative loss in the subsidized area mirrors the relative loss across the entirety of the planted region in a given municipality. To calculate the total loss, we multiply the predicted relative loss by the total soybean production (in kilograms) sourced from the Municipal Agricultural Production database of the Brazilian Institute of Geography and Statistics. This database presents a comprehensive collection of data and statistics related to agricultural production in Brazil. IBGE, being the main provider of data and statistics in Brazil, compiles this information to provide insights into various aspects of the country’s agricultural sector. From this dataset, we used the total soybean production within a specific municipality.¹² Hence, we obtain an estimated total soybean loss due to drought across Brazilian municipalities. As a result, we have aggregated these loss estimates at the national level as well as the observed losses that only considered the subsidized areas. More specifically, the total soybean loss due to drought and excessive rain in a municipality i at crop year j , denoted as $\hat{\tau}_{ij}$, is given by

$$\hat{\tau}_{ij} = \hat{E}[L_{ij} \mid X_{ij}, Z_{ij} = 1] \cdot \hat{P}(Z_{ij} = 1 \mid X_{ij}) \cdot prod_{ij}$$

where $prod_{ij}$ is the total soybean production in municipality i at crop year j .

Table 4 gives examples for some municipalities, where L shows the total indemnity (relative losses), $\hat{L}$ shows the estimated losses, $prod$ is the total production collected from IBGE, and $\hat{\tau}$ is the total estimated loss (insured and non-insured).

Table 5 displays the extrapolated results and the observed loss derived from the subsidized data. The highest observed loss for insured and subsidized areas reached R\$0.856 billion in the 2019/2020 crop year. This coincides with the year that recorded an extrapolated total loss of R\$ 23.39 billion.

5 Stress test under different weather conditions

In this section, we conduct a stress test on the climate variables to understand how the model performs under four environmental scenarios. The procedure

¹²<https://sidra.ibge.gov.br/tabela/5457>

Municipality	L	$\hat{L}$	$prod$ (tons)	$\hat{\tau}$
Cascavel (PR)	851,382	661,582	300,000	24,900,000
Paranatinga (MT)	4,244,852	3,262,853	742,560	90,502,001
Machadinho (RS)	180,220	147,912	70,800	2,619,000
Anaurilandia (MS)	2,293,388	2,885,280	81,670	6,895,000

Table 4: Total observed and predicted losses (in BRL) of soybean cultures in four Brazilian municipalities in 2021/2022 crop year, accounting insured and non-insured areas, and total insured observed loss (in BLR) of soybean only for subsidized areas and yield-based contracts.

Crop	Total Extrapolated Loss	Observed Loss
2017/2018	R\$ 7.335 bi	R\$ 0.117 bi
2018/2019	R\$ 18.388 bi	R\$ 0.458 bi
2019/2020	R\$ 23.390 bi	R\$ 0.856 bi
2020/2021	R\$ 6.817 bi	R\$ 0.299 bi
2021/2022	R\$ 14.288 bi	R\$ 0.386 bi

Table 5: Total predicted loss (in BRL) of soybean cultures in Brazil for each crop year accounting insured and non-insured areas, and observed losses (in BLR) of soybean only for subsidized areas and yield-based contracts.

starts by setting a stress criterion which is used to impute values that simulate extreme climate conditions for the 2021/2022 crop year. These imputed values are designed to mimic challenging scenarios that the crops might face in a changing climate. The scenarios we consider for the stress tests are given as follows.

- **Scenario I (strong drought):** This scenario is designed to simulate a year of significantly reduced rainfall. The value for the precipitation in the 2021/2022 crop year in each municipality is set to the minimum precipitation in all municipalities, which is equal to 30.6 mm in December and 27.1 mm in January.
- **Scenario II (extreme drought):** This scenario takes the conditions of Scenario I a step further by assuming a complete absence of rainfall. This is represented by setting the precipitation in all municipalities to zero during the key growing months of December, and January for the 2021/2022 crop year.
- **Scenario III (excessive rainfall):** This scenario imputes high precipitation values (300 mm) in the months of December and January.
- **Scenario IV (extreme rainfall):** This scenario imputes high precipitation (500mm) values in December and January.

Table 6 presents the estimated loss for insured areas across four municipalities, taking into account the four distinct scenarios. Scenarios I and II produce the most elevated extrapolated values. Scenario IV, characterized by significant rainfall, follows closely behind, while Scenario III registers the lowest estimates. The nearly identical loss estimates observed between Scenarios I and II can be attributed to the inherent behavior of Random Forests, which tends to provide flat extrapolations outside the observed data range. When predicting values outside the observed range of the training data, a Random Forest will typically provide an output that is close to the extreme values it has seen during training. This behavior leads to a flat extrapolation, meaning it does not account for possible variations or trends outside the observed range, but simply gives an almost constant prediction.

Município	UF	Scenario I	Scenario II	Scenario III	Scenario IV
Cascavel	PR	3,596,958	3,596,958	863,470	1,553,803.7
Paranatinga	MT	8,366,718	8,388,983	3,130,357	4,804,123
Machadinho	RS	652,495	652,495	460,127	510,050.10
Anaurilandia	MS	9,368,487	9,368,487	2,620,789	3,977,720

Table 6: Total predicted loss (in BRL) of soybean cultures in Brazil for in 2021/2022 according to each scenario in four municipalities from Brazil. Estimates are given for insured areas.

We estimate the total loss in each scenario also for the non-insured area by using the total production municipality data from Sidra/IBGE, as given in Table 7. Scenario II, which simulates an extreme drought condition, results in the highest extrapolated loss estimate.

Scenario	Total Loss (Insured + Non-insured)
I	14,391,066
II	16,895,125
III	9,215,977
IV	14,987,086

Table 7: Total predicted loss (in BRL) of soybean cultures in Brazil for 2021/2022 according to each scenario. Estimates are given for insured and non-insured areas using the municipal production data (Sidra/IBGE).

6 Concluding remarks

In this paper, we proposed soybean loss prediction models using data from subsidized insurance policies provided by the Brazilian Federal Government's

Rural Subsidy Program (PSR). By merging policy data with climate data on precipitation and temperature from CRU-TS, soil moisture capacity, and El-Niño and La-Niña indices, we constructed predictive models based on Generalized Linear Models (GLM) and Random Forest to forecast aggregate losses at the municipal level in Brazil. The two-stage model (frequency-severity) based on Random Forest yielded favorable results for both classification and regression tasks. Additionally, due to the absence of climate data at the beginning of the crop season, we used data from previous seasons to estimate losses while marginalizing climate variables. We also proposed methods to extrapolate losses from insured to uninsured entities, assuming risk homogeneity, using municipal production totals data. A stress analysis was conducted to assess the sensitivity of precipitation variables. As a future research direction, exploring a farm-level model is recommended, utilizing farm-specific data to differentiate between two farms within the same municipality.

References

- Banholzer, S. and Donner, S. (2014). The influence of different El Niño types on global average temperature. *Geophysical Research Letters*, 41(6):2093–2099.
- Berlato, M. A. and Fontana, D. C. (2001). Impacts of El Niño and La Niña on agricultural production in southern brazil and the use of climate forecasts in agriculture. *Applications of climate forecasting for better decision-making processes in agriculture. Passo Fundo, Embrapa Trigo*, pages 217–241.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Routledge.
- Diaz, A. F., Studzinski, C. D., and Mechoso, C. R. (1998). Relationships between precipitation anomalies in uruguay and southern brazil and sea surface temperature in the pacific and atlantic oceans. *Journal of Climate*, 11(2):251–271.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.
- FAO (2020). Protecting livelihoods: Linking disaster risk insurance and social protection in agriculture.

- FAO (2021). Protecting livelihoods – linking agricultural insurance and social protection.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Harris, I., Osborn, T. J., Jones, P., and Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific data*, 7(1):109.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.
- Liu, Y. and Ramsey, A. F. (2023). Incorporating historical weather information in crop insurance rating. *American Journal of Agricultural Economics*, 105(2):546–575.
- Mota, A. L., Miquelluti, D. L., and Ozaki, V. A. (2020). Predição de sinistros agrícolas: uma abordagem comparativa utilizando aprendizagem de máquina. *Economia Aplicada*, 24(4):533–554.
- Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384.
- Zanatta, M. A. (2019). *Variabilidade espaço-temporal da precipitação e efeito sobre a produtividade da soja*. PhD thesis, Universidade de São Paulo.
- Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.
- Zhu, W., Porth, L., and Tan, K. S. (2019). A credibility-based yield forecasting model for crop reinsurance pricing and weather risk management. *Agricultural Finance Review*, 79(1):2–26.

A Model performance evaluation

We present the predictive performance results for both classification and regression models described in 3.1.

B Maps of weather variables

Model	AUC (average)	AUC (2021-2022)
Logistic Regression	0.777	0.830
GAM	0.759	0.815
Random Forest	0.805	0.844

Table 8: Area under the ROC curves (AUC) for the classification models. The second column presents the average area under the curves computed across the series of test sets, whereas the third column shows the AUC specifically for the 2021-2022 crop year.

Model	MAE (average)	MAE (2021-2022)
Gamma Regression	0.290	0.299
GAM	0.272	0.318
Random Forest	0.253	0.183

Table 9: Mean absolute errors of the loss models. The second column presents the average mean absolute error computed across the series of test sets, whereas the third column shows the mean absolute error specifically for the 2021-2022 crop year.

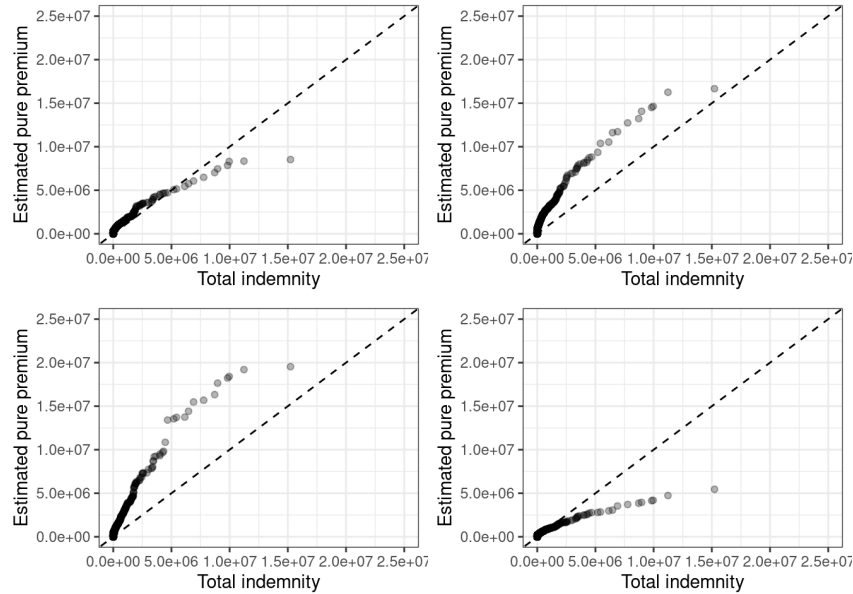


Figure 5: Quantile-quantile plots comparing the total indemnities with the estimated pure premium distributions for municipalities.

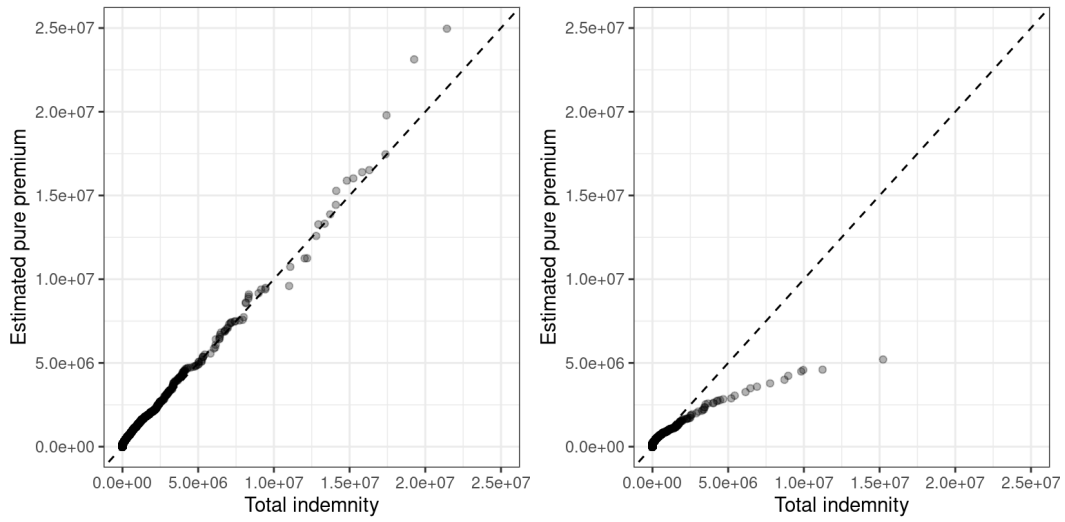


Figure 6: Quantile-quantile plots comparing the total indemnities with the estimated pure premium distributions for municipalities. The left panel shows the in-sample plot, while the right panel shows the out-of-sample plot, specifically from the crop year 2021/2022.

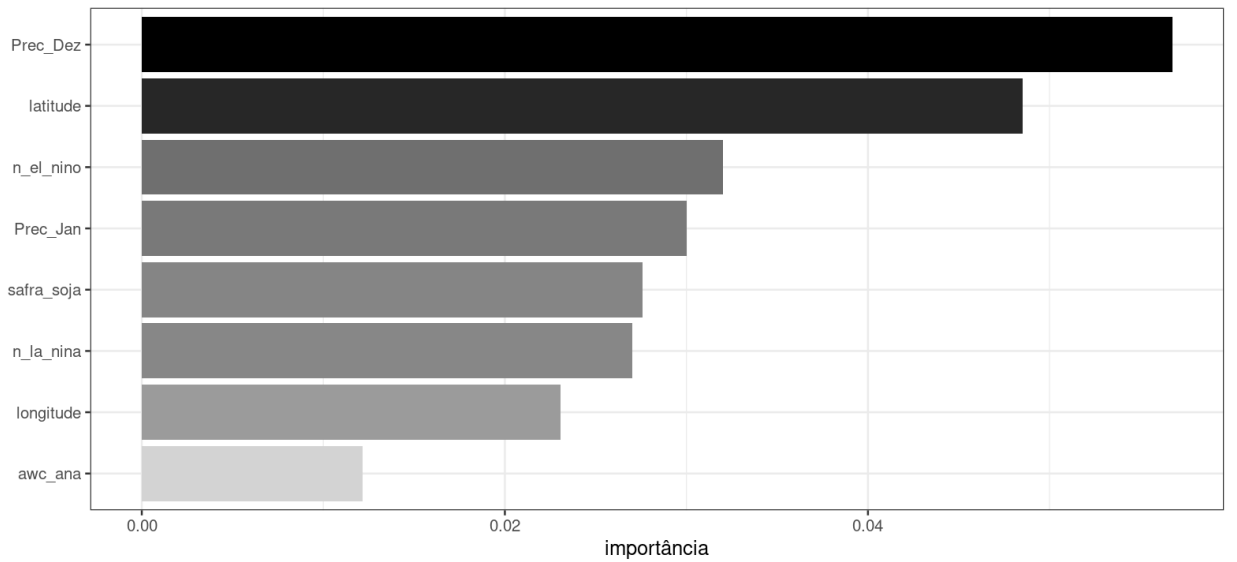


Figure 7: Variable importance of the predictors in the loss model.

Figure 8: Total precipitation distribution in Brazil for November.

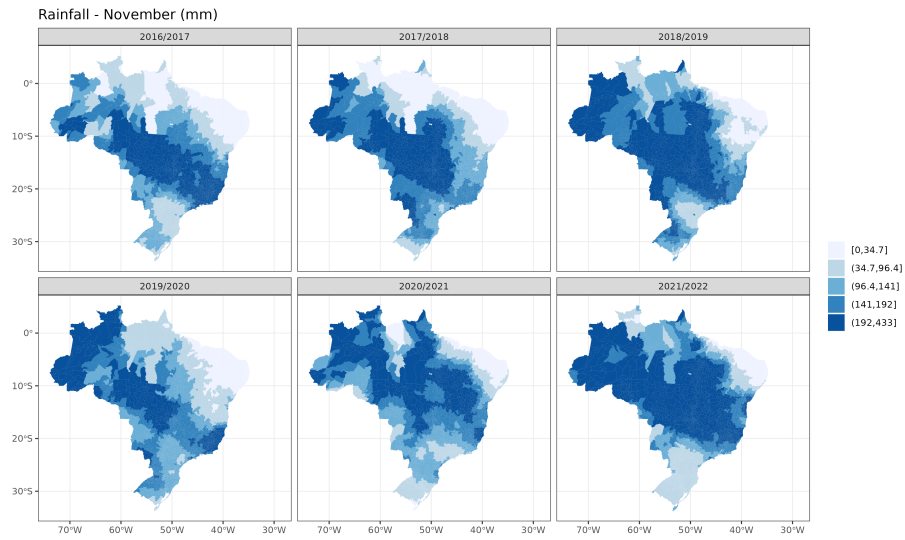


Figure 9: Total precipitation distribution in Brazil for December.

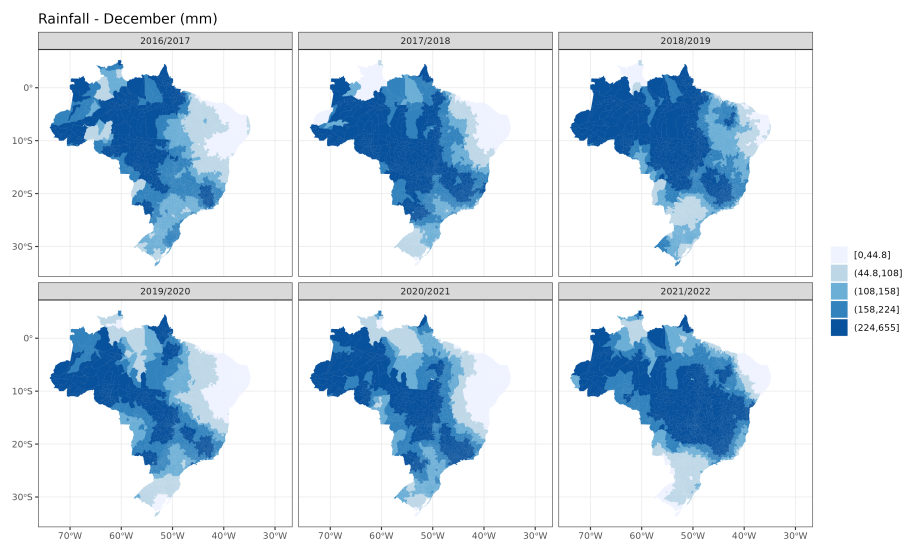


Figure 10: Total precipitation distribution in Brazil for January.

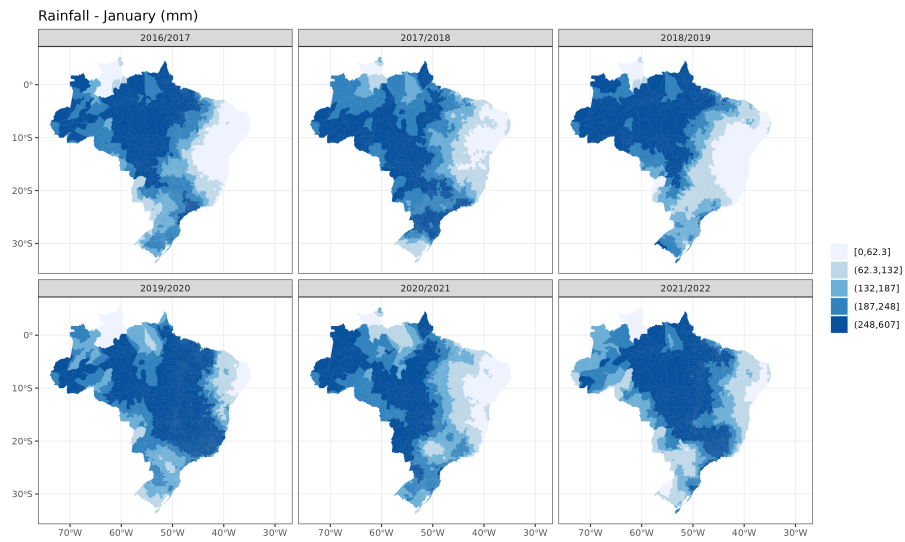


Figure 11: Total precipitation distribution in Brazil for February.

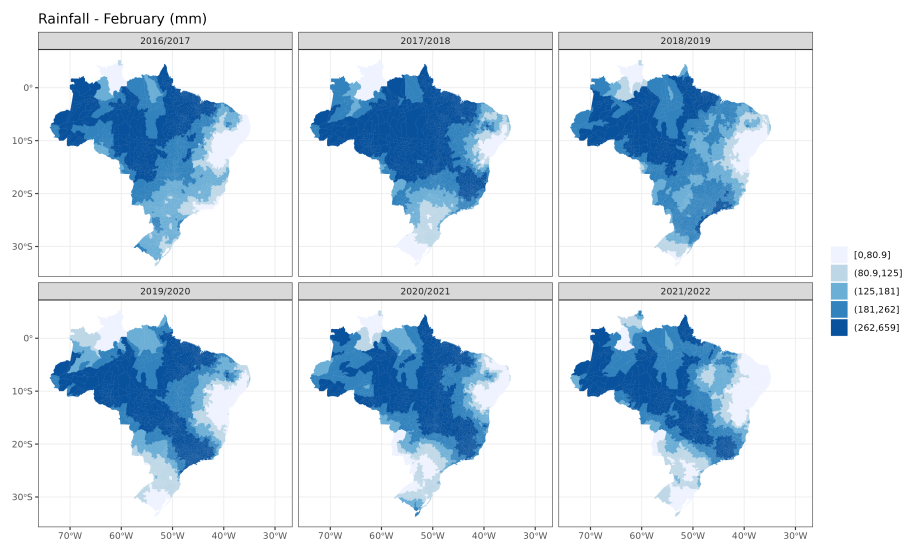


Figure 12: Average temperature distribution in Brazil for November.

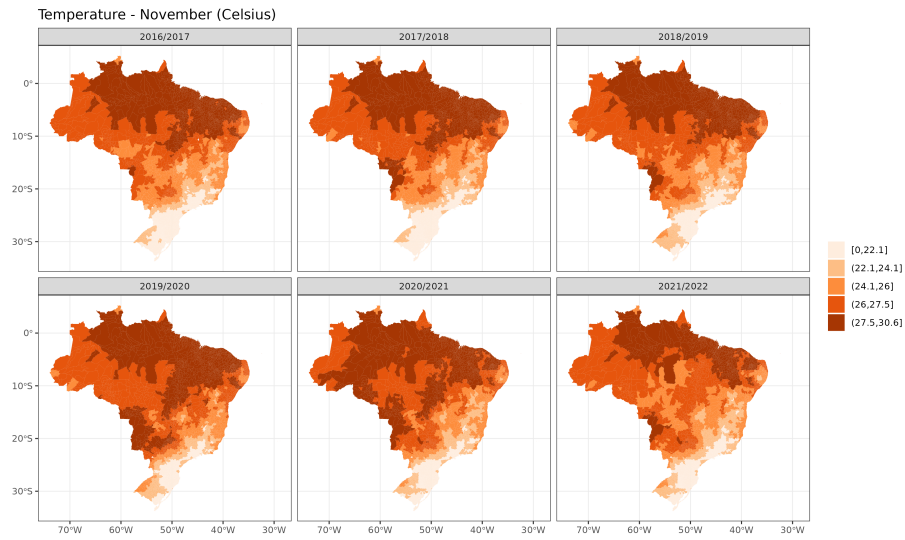


Figure 13: Average temperature distribution in Brazil for December.

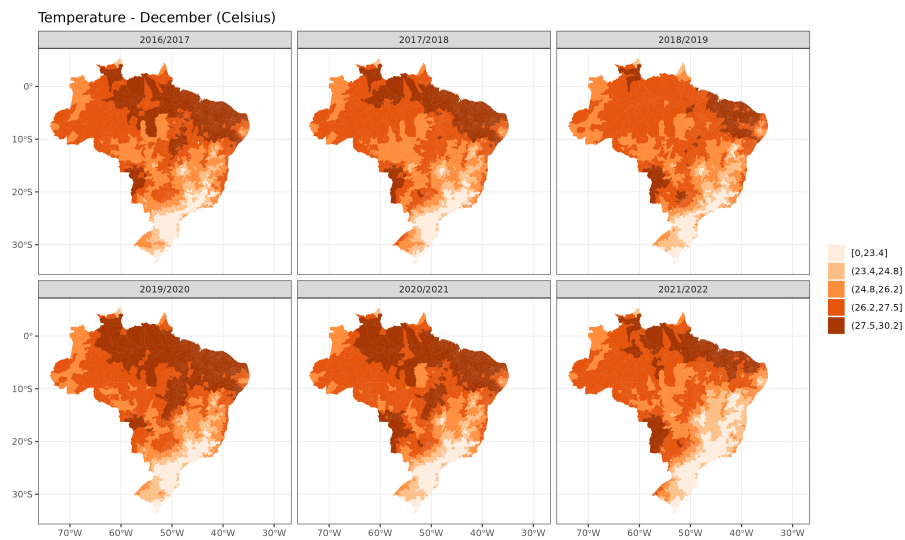


Figure 14: Average temperature distribution in Brazil for January.

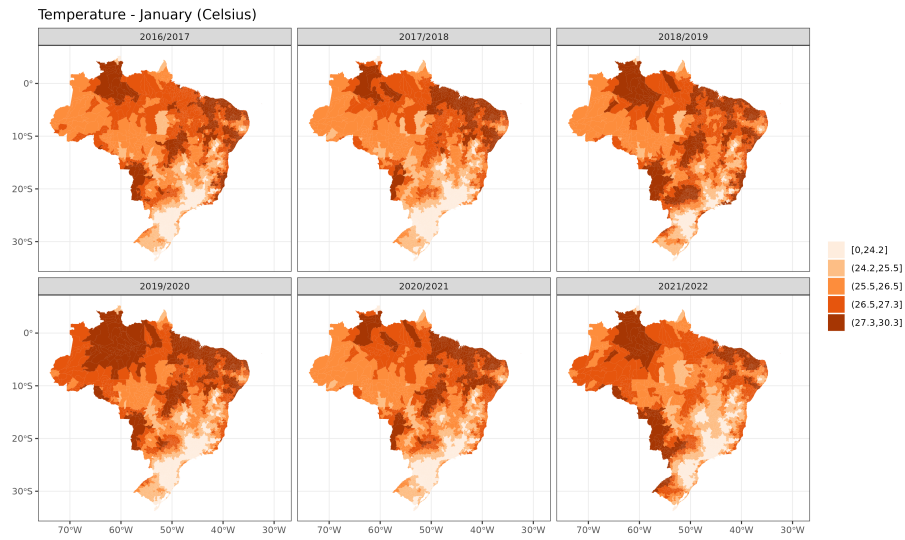


Figure 15: Average temperature distribution in Brazil for February.

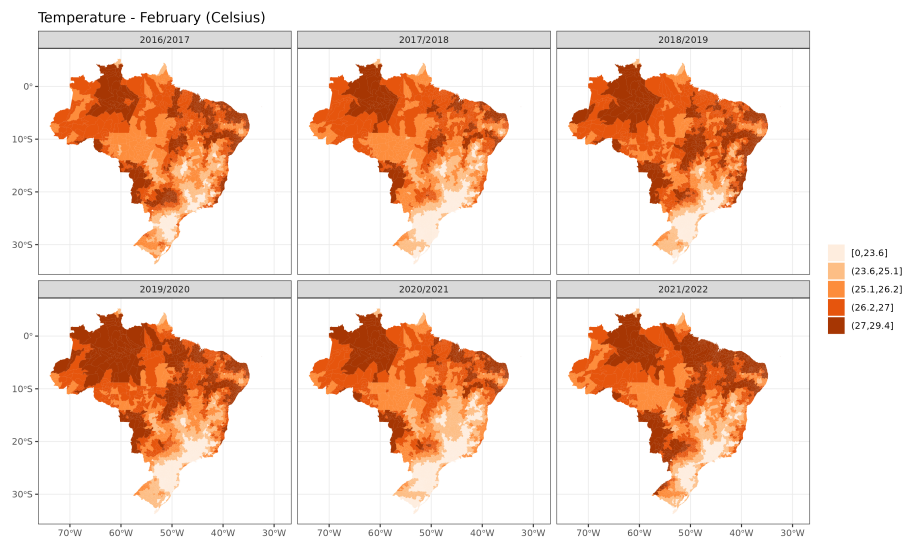


Figure 16: Available water capacity (AWC) for municipalities in Brazil.

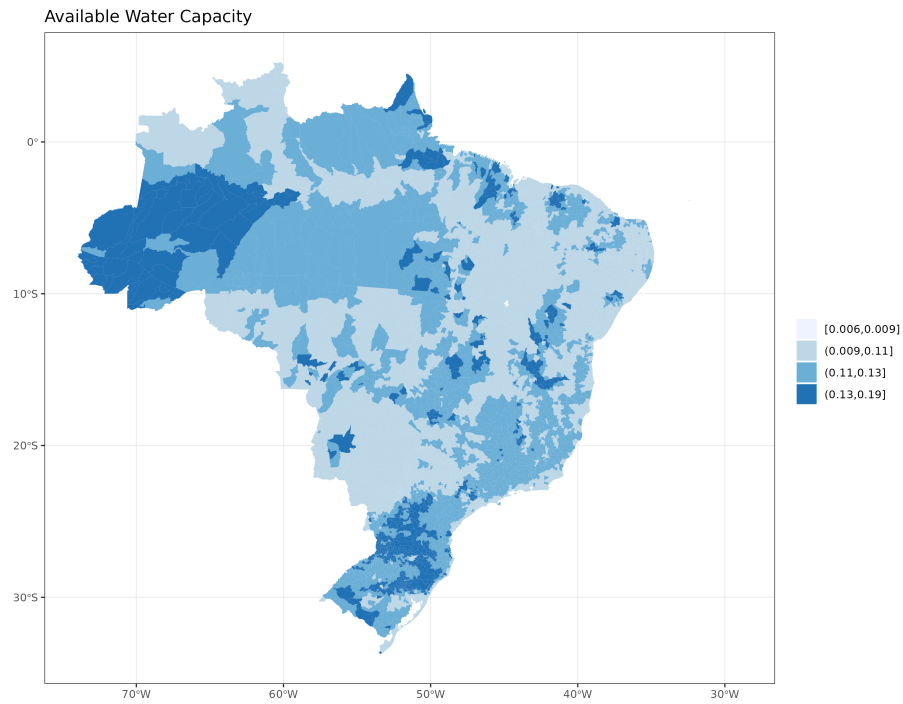
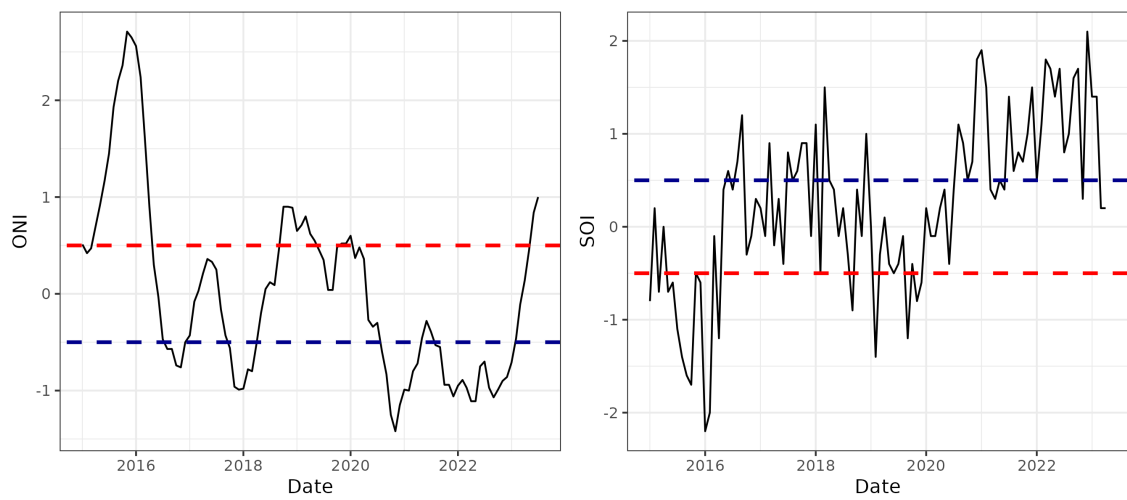


Figure 17: Time series for the Oceanic Niño Index (ONI) and Southern Oscillation Index (SOI) spanning January 2015 to April 2023. The red lines mark the threshold for El Niño events, while the blue lines denote the La Niña threshold.



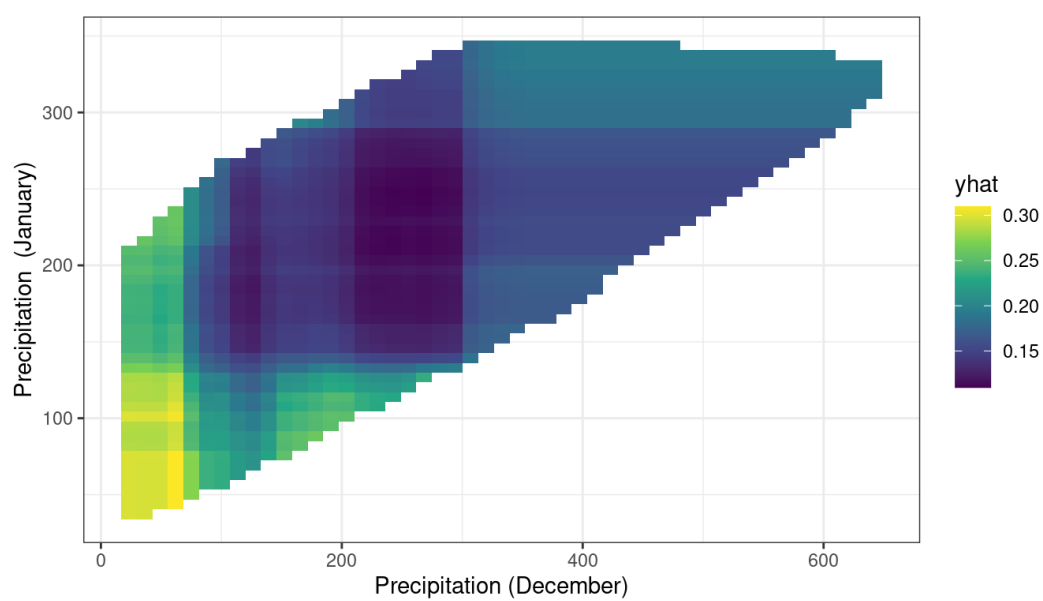


Figure 18: Partial dependence plots showing level curves for the average predictions of the relative losses based on the pair of December and January total precipitation values.