Supplementary Mateiral

A data-driven framework for assessing climatic impact-drivers in the context of food security

Marcos Roberto Benso[1], Roberto Fray Silva[2], Gabriela Gesualdo Chiquito[1], Antonio Mauro Saraiva[2], Alexandre Cláudio Botazzo Delbem[4], Patricia Angélica Alves Marques[3], and Eduardo Mario Mendiondo[1]

[1]São Carlos School of Engineering, University of Sao Paulo, Sao Carlos-SP, 13566-590, Brazil
[2]Institute of Advanced Studies, University of São Paulo, São Paulo, SP, Brazil
[3]Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, SP, Brazil
[4] Institute of Mathematics and Computer Sciences, University of Sao Paulo, Sao Carlos-SP, 13566-590, Brazil

Correspondence: Marcos R Benso (marcosbenso@gmail.com)

1 Introduction

This supplemental document contains details of data pre-processing and provides additional visual aids for the main manuscripts. Readers might benefit from further details on how crop yield data were obtained and processed and further evaluation of the different crop yield datasets used.

2 Crop yield data

Data from Brazilian Institute of Geography and Statistics

Crop yield data is a crucial component in understanding the impacts of climate on food production. Crop yields are generally made available at the municipality level. In Brazil, the Brazilian Institute of Geography and Statistics (IBGE) and state agencies collect agricultural data using surveys, interviews, and expert elicitation to create an annual database for more than 40 crops at the municipal level (de Geografia e Estatística, 2022). The raw data from IBGE can be assessed in the Multidimensional Statistics Database (BME). One of the main challenges of this dataset is that data represents annual statistics and does not represent different cycles. The double cropping system is widely adopted in the study area, therefore representing a potential bottleneck. However, IBGE has started to collect maize in the first and second cycles since 2003. This matches the period when maize's second cycle is intensified in Brazil.

Data from Parana state statistical yearbooks

The Department of Rural Economy (Deral) of the Paraná state, Brazil, is also responsible for collecting crop data at the municipal level. The method of collecting and processing data is similar to what is done by IBGE; therefore, a high level of redundancy is expected from these two datasets. This redundancy is important to validate data and remove outliers that might reduce the quality of a model. The same number of municipalities selected using IBGE data was used in data from Deral. This dataset is derived from the Gross Value of Production, derived from the price and the quantity of production of 30 crops.

Global dataset of historical yields (GDHY)

The Global Dataset of Historical Yields is a global annual time series of 0.5 º grid-cell estimates for maize, rice, wheat, and soybean from 1981 to 2016. For each grid cell, crop yields are estimated in ton/ha based on Food and Agriculture Organization (FAO) country-level yield statistics and then corrected using the remote-sensed leaf area index (LAI), the fraction of photosynthetically active radiation (FPAR) and crop-specific radiation use efficiency derived from reanalysis. Crop areas and crop calendars were derived from Monfreda et al. (2008) and Sacks et al. (2010). More details on the dataset are described in Iizumi and Sakai (2020) and Iizumi et al. (2014).

The dataset was aggregated to the municipal level using zonal statistics in the terra package (Hijmans, 2023) in R Studio.

In the literature, three major problems have been reported regarding the quality of crop yield data for risk analysis, namely the presence of outliers, technological trends, and heteroskedasticity. Removal of outliers is a complex problem since we are working with extreme events. The definition of an outlier must be carefully taken to eliminate valuable data. According to Ozaki et al. (2008), soybean and maize crop yield data tend to correlate, considering drought years, present correlation within approximately 150 km, and, within this range, the normality assumption can be supported. Therefore, for simplification, we assumed that crop yields within the IBGE's immediate region are highly correlated, that is, R^2 > 0.6 and p-value < 0.05.

Changes in technology in seed production, fertilizers, and land management impact crop yields (Liu and Ker, 2020). This effect is well documented in the agronomic literature and increases the averages and leads to changes in non-constant variance, i.e., heteroskedasticity (Tolhurst and Ker, 2015; Harri et al., 2011). The effect of timely technological adoption and improvements on crop yields was treated in a two-step process as proposed by Zhu et al. (2011), first by removing trends and then testing and adjusting the heteroskedasticity of the residuals.

In the first step, we tested the presence of monotonic trends using the Mann-Kendall test (Mann, 1945) with the Kendall R Package (McLeod, 2022). If the Mann-Kendall indicates a p-value lower than 0.05, the null hypothesis is accepted, and the crop yield series is considered to have a monotonic trend that must be corrected. We choose the Local Polynomial Regression Fitting (LOESS) (Cleveland et al., 2017) to model crop yields y_t at the year t. The residuals ϵ_t are considered to be the detrended crop yields.

We tested heteroskedasticity in the data using the Pagan-Breusch test (Breusch and Pagan, 1979). If the p-value of the test is lower than 0.05, then the null hypothesis is rejected, and the yield series is considered heteroskedastic. We compute the normalized residuals y_t^n for two cases in this case. The first case is when the errors are proportional to the yield level. Then, we obtain the y_t^n considering the proportional error ε_t, calculated by dividing the error term from the LOESS prediction function ϵ_t by the yield predicted by the function. Lastly, the proportional errors are multiplied by the yields observed in 2021. With this procedure, the past yields are expressed in terms of 2010 technology. Otherwise, when the residuals are not proportional to the yield levels, y_t^n is calculated by adding 2010 yields to the residuals of the LOESS prediction function.

After obtaining a consistent time series corrected for outliers, trends, and heteroskedasticity, the next step is adjusting for a distribution function. Statistical modeling of crop yields is relevant for risk management because we do not have a time series long enough to empirically evaluate risk (Liu and Ker, 2020) empirically. Several parametric and nonparametric distribution functions have been proposed to model crop yields (Ozaki et al., 2008). We selected the gamma distribution because it allows one to simulate positive and negative skewness. The flexibility to assume different shapes according to the parameters c(α, β), the shape and rate parameters, respectively. For the gamma distribution, when α is greater than 1, the distribution is skewed to the right, that is, skewed towards higher crop yields. Otherwise, the distribution is skewed towards lower yields.

To eliminate potential outliers, we excluded values considering each year and state. The hypothesis of high correlation of crop yields within studied regions was confirmed for the Soybean (Fig. SS1) and Maize (Fig SS2) data, indicating that this strategy is adequate. Since soybeans have a longer record, the correlations were more stable in the immediate regions and tended to have higher values.

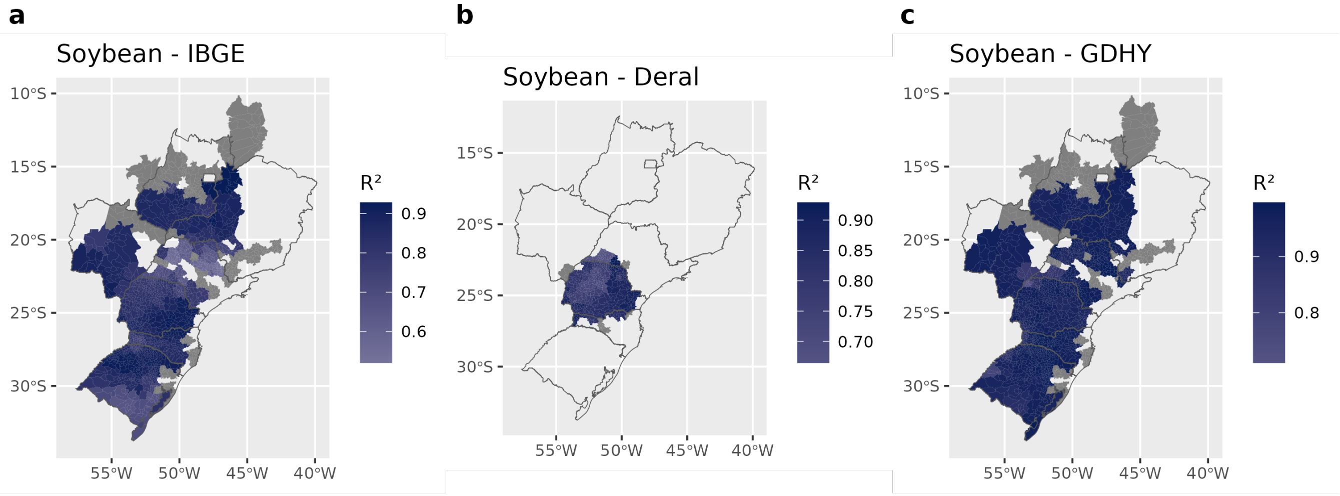

Figure S1. Spatial correlation of municipal soybean crop yields for eath dataset, IBGE, Deral and GDHY

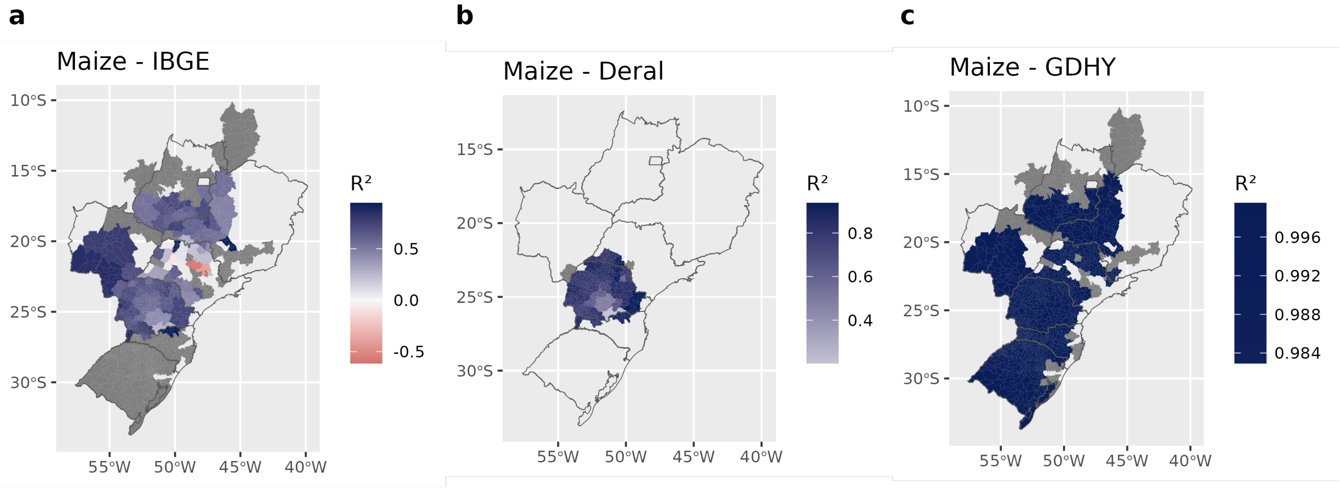

Figure S2. Spatial correlation of municipal maize crop yields for eath dataset, IBGE, Deral and GDHY

The comparison of the datasets used in this study is vital to evaluate the reliability of the data. High-quality crop yield data improves calibration of crop growth models (Rosenzweig et al., 2014). However, they have a broader application in geosciences. Crop yield data is used to parameterize watershed hydrological models, especially in agricultural catchments, and improves the simulation of soil moisture (Sinnathamby et al., 2017). For water resources management, using higher quality crop yield data has improved global knowledge on the water-food-energy nexus (Ai and Hanasaki, 2023; Wang et al., 2023).

We compared crop yields at the municipal level in Brazil. As observed in Figures S3 and S4, IBGE and Parná Deral data for soybeans and maize are highly correlated; however, outliers were detected in both datasets. The outlier removal process improved the agreement between the two datasets, suggesting that eliminating data improved the dataset's quality. Since Deral is only available in Paraná, for the other states of Brazil, only GDHY and IBGE were compared. The global dataset of historical yields aggregated at the municipal level has a weak association with the other datasets. This result confirms what was reported by Iizumi et al. (2014). The GDHY is based on satellite data collected from a fixed cropland map. In many regions of Brazil there is a noticeable increase in croplands, which can influence the estimation of GHDY. Also, the exact location of the planted area within each municipality can vary from year to year.

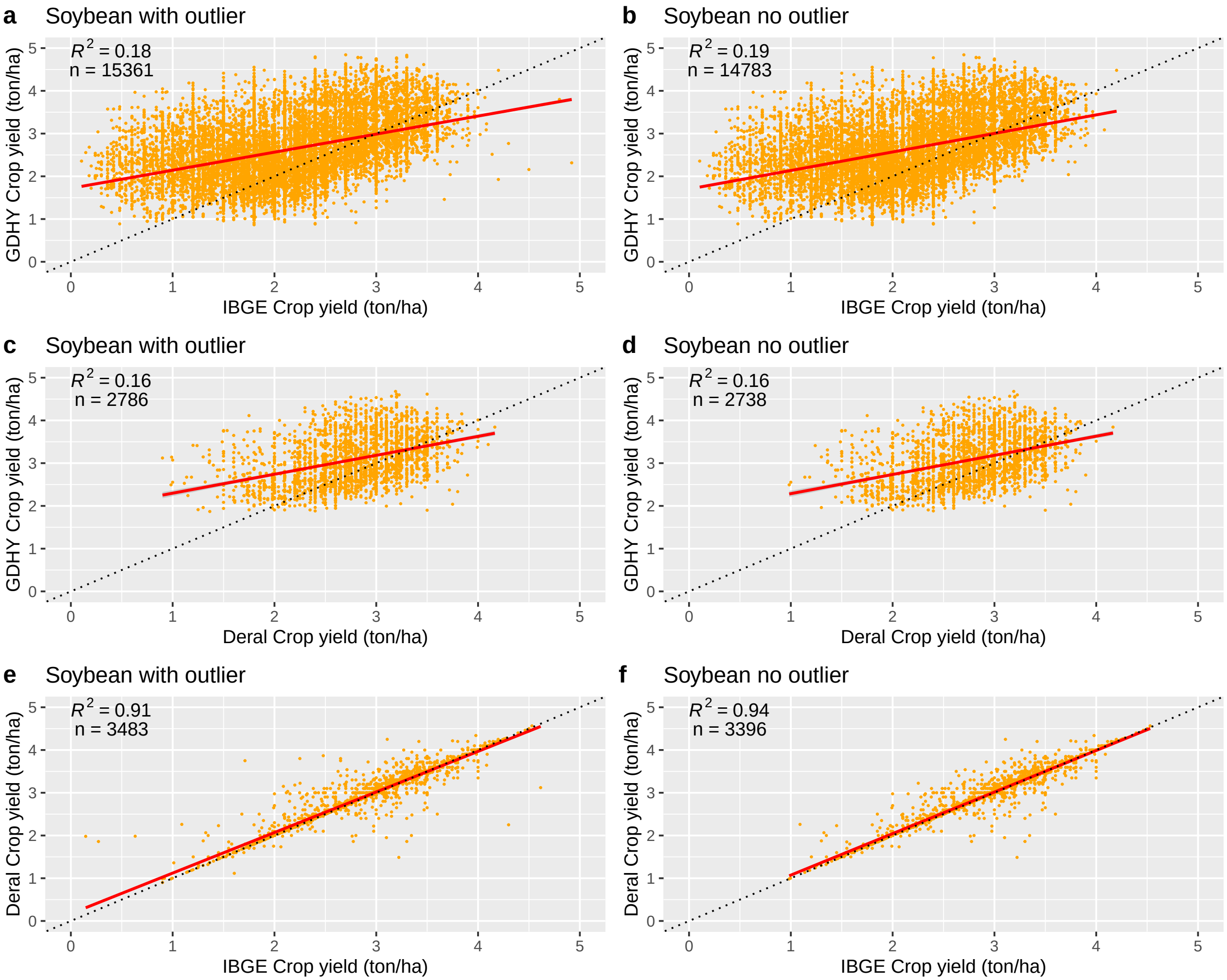

Figure S3. Correlation analysis of three different datasets for soybean crop yields: (Global dataset of historical yields ($yield_{gdhy}$), Brazilian Institute for Geography and Statistics ($yield_{IBGE}$), and Paraná Department of Rural Economy ($yield_{PR}$)

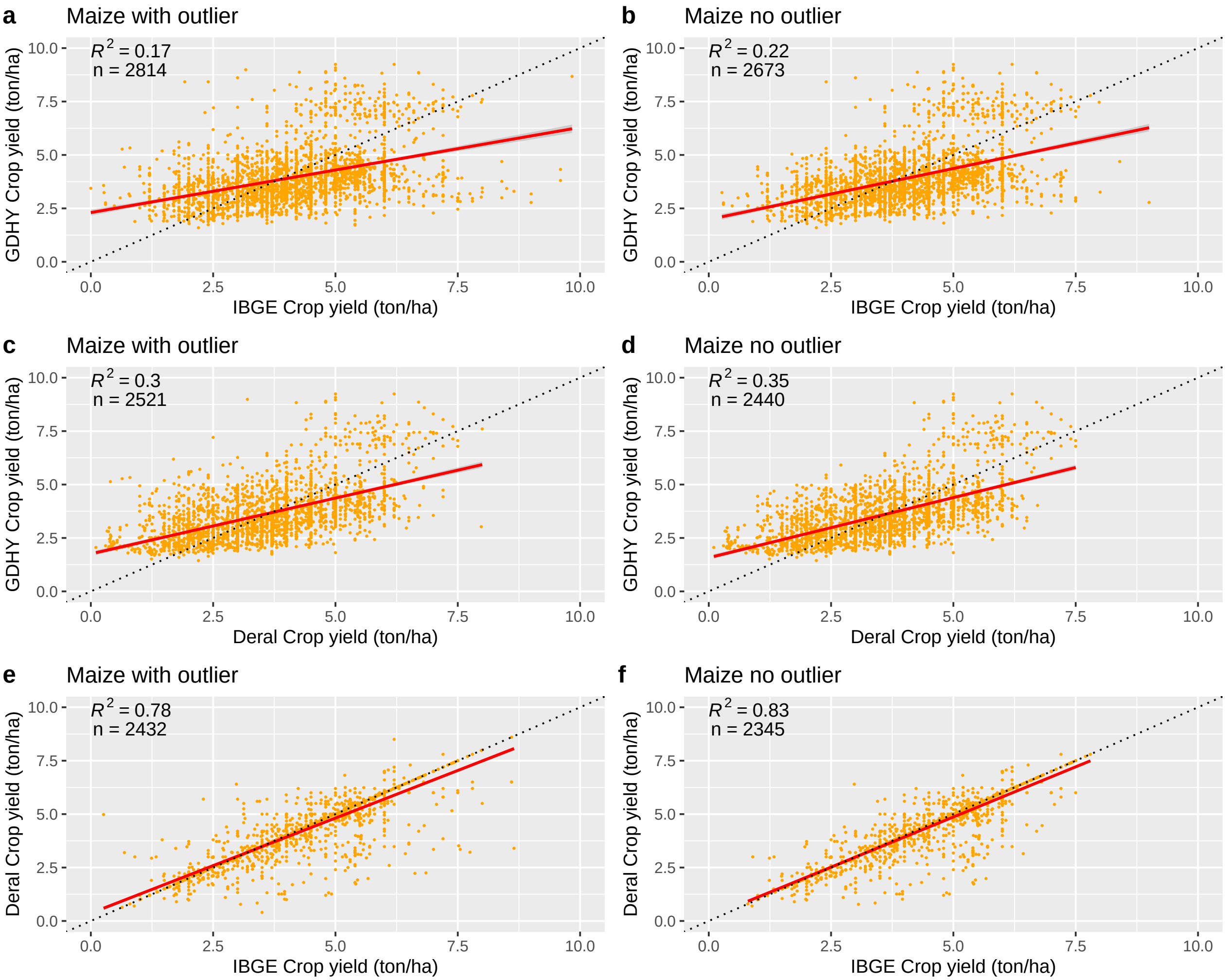

Figure S4. Correlation analysis of three different datasets for maize second cycle crop yields: (Global dataset of historical yields ($yield_{gdhy}$), Brazilian Institute for Geography and Statistics ($yield_{IBGE}$), and Paraná Department of Rural Economy ($yield_{PR}$)

In order to evaluate risk, testing and removing trends is a fundamental step to remove the effects of technology advances on the data (Harri et al., 2011). The Mann-Kendall trend analysis of soybean and maize yields for Brazilian municipalities considering three datasets unveiled a consistent pattern of trends. Data for all municipalities in our study presented significant positive monotonic trends considering p values less than 0.05. For maize, on the other hand, significant positive monotonic trends were observed in the majority of municipalities. However, they were not universally present in the states of MS, GO, SP, and MG in the IBGE dataset. The lack of trends can be attributed to a limitation of the dataset, particularly related to not having long-term data for the second maize cycle.

Since most data presented positive trends, we applied a LOESS model for all the municipalities. The residuals of the LOESS models were then tested for heteroscedasticity. Other studies evaluated the presence of heteroscedasticity in crop yield data, Vicente (2004) was tested using the Brazilian agricultural census 1995/1996, Ozaki et al. (2008) for soybean, maize, and wheat in Paraná, and RODRIGUES et al. (2013) at farm-level studies in São Paulo.

The presence of heteroscedasticity represents systematic changes in crop yield data Yang et al. (1992) and to the best of our knowledge, this is the first study that considers and evaluates spatial characteristics of heteroscedasticity of municipal level in the long term.

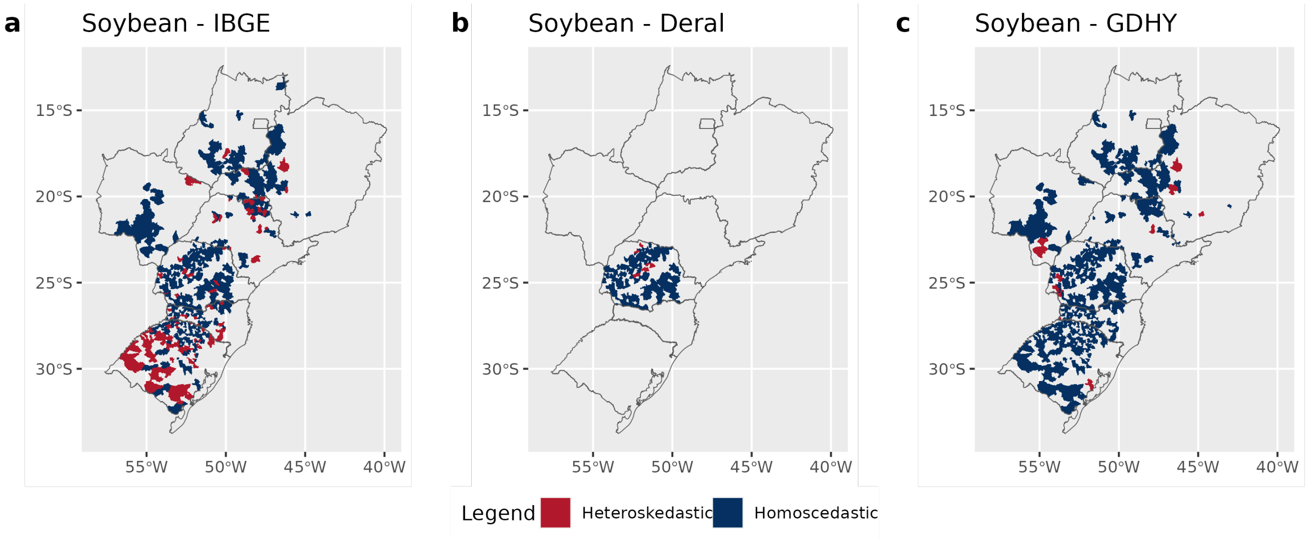

Figure S5. Heteroskedasticity Test Results for soybean crop yields in Brazilian Municipalities

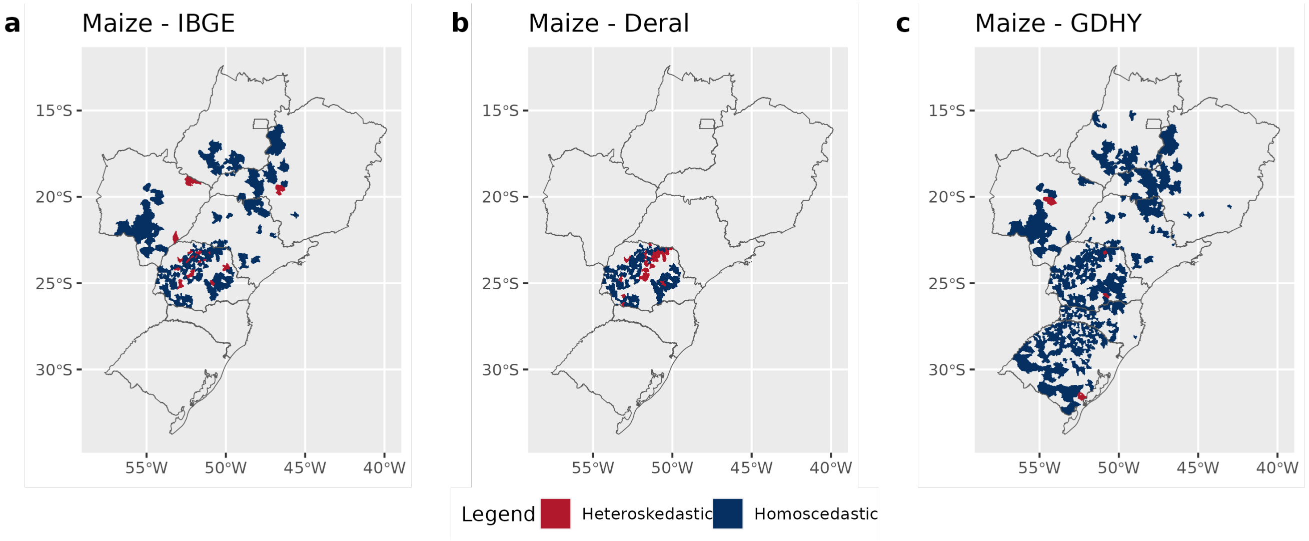

Figure S6. Heteroskedasticity Test Results for maize crop yields in Brazilian Municipalities

3 Crop yield risk monitoring

To illustrate the relationship between extreme climate impacts on food production and climate indices, we highlight the main loss events from 1997 to 2021 in the study area in Figure S7. Several distinct periods of crop yield losses emerged during the study period, which required further analysis. The notable events among these were 2010 for soybeans, with an average of 16% of crop yield losses, and 2021 for maize, with an average of 40%, which impacted agricultural productivity in the region.

The patterns of crop yield losses observed in the region raise two main concerns. The first is that the severe crop yield losses presented in the previous examples have happened only once in the entire time series, representing an imbalance in the values of the data set. One implication of this situation is that models might not have sufficient cases of severe failure to be trained adequately and might underestimate losses. The second concern is related to the decision of what to do with these anomalous events. Possible solutions are to use it for training, testing, or removing it from the dataset. We opted to maintain these events in the analysis with the warning that this might interfere with model performance. However, we wanted to evaluate the ability of the model to predict unprecedented loss events.

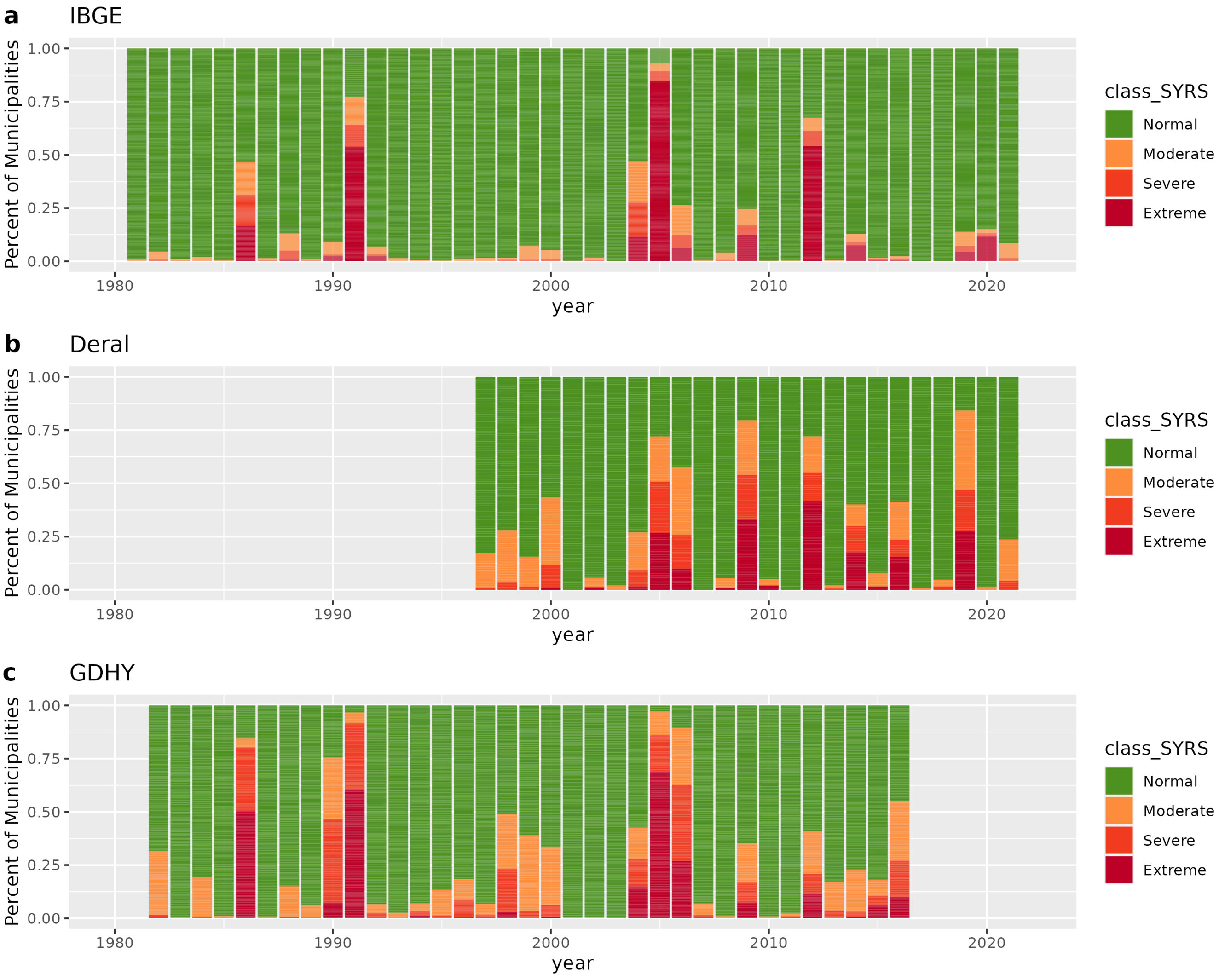

Figure S7. Temporal Variation of Soybean Crop Yields Across Risk Classes for IBGE, Deral and GDHY Datasets. The year-to-year distribution of crop yields is categorized into four risk classes: ’Normal,’ ’Moderate,’ ’Severe,’ and ’Extreme.’

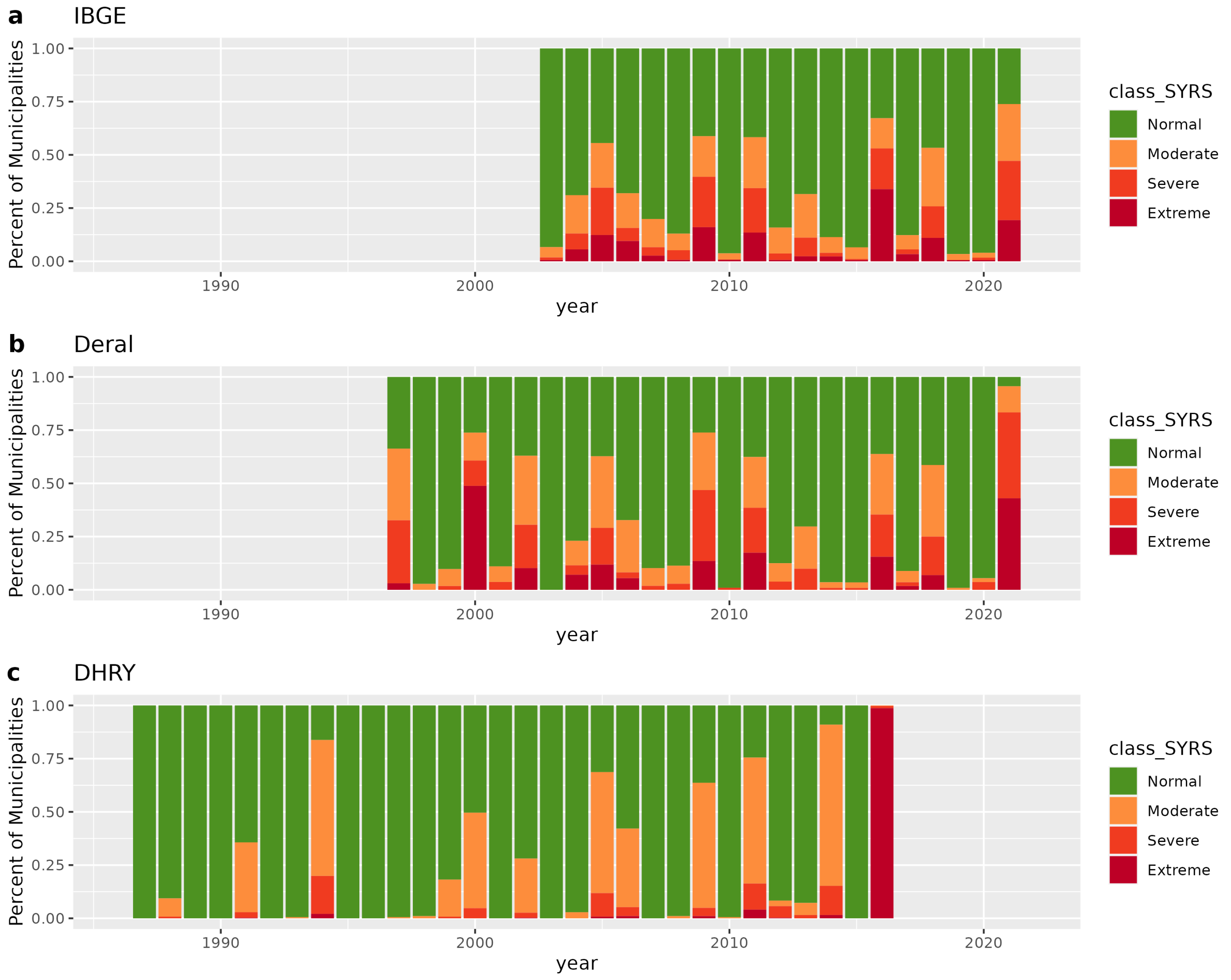

Figure S8. Temporal Variation of Maize Crop Yields Across Risk Classes for IBGE, Deral and GDHY Datasets. The year-to-year distribution of crop yields is categorized into four risk classes: ’Normal,’ ’Moderate,’ ’Severe,’ and ’Extreme.’

References

Ai, Z. and Hanasaki, N.: Simulation of crop yield using the global hydrological model H08 (crp. v1), Geoscientific Model Development, 16, 3275–3290, 2023.

Breusch, T. S. and Pagan, A. R.: A simple test for heteroscedasticity and random coefficient variation, Econometrica: Journal of the econometric society, pp. 1287–1294, 1979.

Cleveland, W. S., Grosse, E., and Shyu, W. M.: Local regression models, in: Statistical models in S, pp. 309–376, Routledge, 2017.

de Geografia e Estatística, I. B.: Produção Agrícola Municipal 2022, http://www.sidra.ibge.gov.br/bda/pesquisas/pam, 2022.

Harri, A., Coble, K. H., Ker, A. P., and Goodwin, B. J.: Relaxing heteroscedasticity assumptions in area-yield crop insurance rating, American Journal of Agricultural Economics, 93, 707–717, 2011.

Hijmans, R. J.: terra: Spatial Data Analysis, https://CRAN.R-project.org/package=terra, r package version 1.7-29, 2023.

Iizumi, T. and Sakai, T.: The global dataset of historical yields for major crops 1981–2016, Scientific Data, 7, 97, 2020.

Iizumi, T., Yokozawa, M., Sakurai, G., Travasso, M. I., Romanenkov, V., Oettli, P., Newby, T., Ishigooka, Y., and Furuya, J.: Historical changes in global yields: major cereal and legume crops from 1982 to 2006, Global ecology and biogeography, 23, 346–357, 2014.

Liu, Y. and Ker, A. P.: When less is more: on the use of historical yield data with application to rating area crop insurance contracts, Journal of Agricultural and Applied Economics, 52, 194–203, 2020.

Mann, H. B.: Nonparametric tests against trend, Econometrica: Journal of the econometric society, pp. 245–259, 1945.

McLeod, A.: Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test, https://CRAN.R-project.org/package=Kendall, r package version 2.2.1, 2022.

Monfreda, C., Ramankutty, N., and Foley, J. A.: Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, Global biogeochemical cycles, 22, 2008.

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

RODRIGUES, M., Corá, J. E., Castrignanò, A., Mueller, T. G., and Rienzi, E.: A Spatial and Temporal Prediction Model of Corn Grain Yield as a Function of Soil Attributes., Agronomy Journal, pp. 1878–1887, 2013.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., et al.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proceedings of the national academy of sciences, 111, 3268–3273, 2014.

Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N.: Crop planting dates: an analysis of global patterns, Global ecology and biogeography, 19, 607–620, 2010.

Sinnathamby, S., Douglas-Mankin, K. R., and Craige, C.: Field-scale calibration of crop-yield parameters in the Soil and Water Assessment Tool (SWAT), Agricultural water management, 180, 61–69, 2017.

Tolhurst, T. N. and Ker, A. P.: On technological change in crop yields, American Journal of Agricultural Economics, 97, 137–158, 2015.

Vicente, J. R.: Economic efficiency of agricultural production in Brazil, Revista de Economia e Sociologia Rural, 42, 201–222, 2004.

Wang, J., Wei, J., Shan, W., and Zhao, J.: Modeling the water-energy-food-environment nexus and transboundary cooperation opportunity in the Brahmaputra River Basin, Journal of Hydrology: Regional Studies, 49, 101 497, 2023.

Yang, S.-R., Koo, W. W., and Wilson, W. W.: Heteroskedasticity in crop yield models, Journal of Agricultural and Resource Economics, pp. 103–109, 1992.

Zhu, Y., Goodwin, B. K., and Ghosh, S. K.: Modeling yield risk under technological change: Dynamic yield distributions and the US crop insurance program, Journal of Agricultural and Resource Economics, pp. 192–210, 2011.