Three Essays on Risk Management and Correlation, and Meteorological Hazard

Anaïs Goburdhun

université PARIS-SACLAY

NNT : 2019SACLX008

Thèse de doctorat

Trois chapitres sur la gestion et la corrélation du risque, et le risque météorologique

Thèse de doctorat de l'Université Paris-Saclay
préparée à Ecole Polytechnique

Ecole doctorale n°578 Sciences de l'Homme et de la Société (SHS)
Spécialité de doctorat : Sciences économiques

Thèse présentée et soutenue à Paris, le 28 février 2019, par

ANAÏS GOBURDHUN

Composition du Jury :

Christophe Gouel Chargé de recherche, INRA	Président
Katheline Schubert Professeur, Paris School of Economics	Rapporteur
Andreas Heinen Professeur, Université de Cergy Pontoise	Rapporteur
Eric Strobl Professeur, University of Bern	Directeur de thèse
Geoffrey Barrows Chercheur, Ecole Polytechnique	Co-directeur de thèse

Acknowledgements

Firstly, I would like to thank the scientific team without who this thesis would not have been accomplished. In particular, I am grateful to my advisors Eric Strobl, Pierre Picard and Geoffrey Barrows, who supported and guided my work, and helped me in all the initiatives I took during my PhD. I also thank the members of my thesis committee : Katheline Schubert and Andreas Heinen for their contributions as examiners, and Christophe Gouel as the Chairman of the committee.

I also want to thank the researchers from CREST for their strong support at different stages of my PhD, in particular Francis Kramarz, Pierre Boyer and Guy Meunier, for helping me in my research projects and in finalizing my PhD thesis. I also thank Pierre Cahuc, for hiring me as a teaching assistant for his class at Ecole Polytechnique and for giving me the opportunity to enjoy the job of teaching.

My PhD let me the tremendous opportunity to go to MIT in Cambridge (USA) for eight months for a research project at the Joint Program for Science and Policy for Global Change. I am very grateful to the people who made this project possible : Eric Strobl and Elodie Blanc for their collaboration, the Chaire de Développement Durable EDF - X, CREST - CNRS and Isabelle Méjean for their financing support, and the responsible administrators from the different sides, in particular Lyza Racon. Thank you so much for helping me achieving this project. From both professional and personal points of view, it was an unforgettable experience.
From this experience, I am also very grateful to John Reilly, Kerry Emanuel and Jennifer Morris for their interesting remarks and discussions on my work.

Je remercie toute l'équipe technique et administrative pour leur aide essentielle

dans toutes les étapes de ma thèse. En particulier, je remercie Lyza Racon pour m'avoir particulièrement soutenue dans les moments difficiles. Je tiens aussi à remercier le service informatique, en particulier Sri Srikandan, pour leur disponibilité à toute heure et à distance et pour leur patience malgré mes nombreuses sollicitations.

Au terme de ces huit années d'études intensives jalonnées d'épreuves et de décisions, je tiens à remercier mes parents pour m'avoir écoutée, soutenue et aidée à mener avec succès mes projets. Je remercie aussi infiniment mon frère et allié de toujours, Kevin, pour sa présence et sa complicité inégalables et ce, où que nous soyons dans le monde. Je témoigne une immense reconnaissance envers mes grands-parents qui, au-delà de leur soutien, m'ont transmis leur amour pour la science et la culture dès mon plus jeune âge et ont joué un rôle indéniable dans les choix que j'ai faits. Un grand merci à ma famille, tantes et oncles, cousins et cousines, pour ces moments plein de vie et de joie partagés.

J'adresse mes remerciements à mes collègues doctorants qui ont rendu cette expérience encore plus riche et vivante et toujours avec un esprit d'entraide. Je m'adresse tout particulièrement à Aymeric Guidoux, Alexis Louaas, Reda Aboutajdine, Anasuya Raj et Margot Hovsepian qui ont fait partie de l'aventure du début à la fin, mais aussi à Clémence Tricaud avec qui j'ai partagé mon expérience à Boston et une croisière au clair de lune sur Charles River.

Cet aboutissement n'aurait pas été possible sans le soutien que mes amis proches m'apportent continuellement. Je remercie tout particulièrement Julie et Emilia pour leur sincère et inconditionnelle amitié et présence pendant ces trois ans de thèse. Merci à mes amis qui me suivent depuis la primaire pour certaines, notamment Raphaëlle pour le partage de notre expérience américaine et Clémence pour avoir été très présente malgré la distance. Je tiens également à remercier mes coéquipiers et mon entraîneur d'équitation qui m'ont permis de concilier ma passion et le sport avec ma thèse. Many thanks to my best roommate, Mary Anne, for having shared your apartment and great moments with me for six months in Boston.

Enfin, je remercie Caramelle, Sammy et Lulu pour leurs ronronnements réconfortants.

Cette thèse a été réalisée grâce au support financier du Ministère de l'Enseignement Supérieur et de la Recherche

Résumé long en français

La thèse étudie le risque météorologique et économique sous différents angles principalement dans les pays en développement. Elle se décompose en trois chapitres indépendants analysant dans diverses situations la corrélation des risques liés aux aléas météorologique et climatique ou économique, et étudie le potentiel de la région géographique étudiée pour mettre en place un système d'assurance contre le risque étudié. En effet, cette thèse étudie des risques très susceptibles d'être fortement corrélés : que cela concerne le risque météorologique ou climatique, ou le risque lié à la volatilité des prix, les villes voire pays voisins sont exposés aux mêmes risques et de façon simultanée. Cet aspect essentiel compromet la mutualisation du risque, paramètre primordial du modèle économique de l'assurance. A travers les trois chapitres de la thèse, nous étudierons le bénéfice lié à la mutualisation de ces risques a priori relativement corrélés. Le premier chapitre étudie la corrélation des prix du maïs entre les principales villes en Tanzanie. A l'aide d'un modèle Copula-GARCH, la dépendance entre les cours du maïs des 20 marchés principaux du pays est modélisée et nous pouvons voir si le prix moyen du maïs est lissé en agrégeant les marchés. Cela permet de voir si l'intégration des marchés permet une efficace mutualisation du risque lié à la volatilité des prix. Pour ce faire, nous calculons la Value-at-Risk (VaR), outil couramment utilisé dans le domaine de la finance pour mesurer les percentiles correspondant aux valeurs extrêmes d'une distribution. En l'occurrence, nous modélisons les prix sous la forme de rendements en agrégeant plusieurs régions et regardons l'augmentation ou la baisse de prix maximale arrivant avec une probabilité de 1% sur 20 ans. Si les prix sont effectivement lissés, alors la VaR 1% doit diminuer en valeur absolue. Nous réalisons ce calcul en agrégeant aléatoirement de plus en plus de marchés, ou en agrégeant deux à deux les marchés venant de zones géographiques présentant des climats différents. En moyenne, l'agrégation d'un nombre croissant de marchés a ten-

dance à faire diminuer la VaR, mais on s'aperçoit que c'est surtout en mutualisant deux marchés judicieusement choisis (ie avec des conditions climatiques différentes) que la baisse de risque est optimale. Le second chapitre s'attache au risque cyclonique dans les îles Pacifique sud et son impact sur les infrastructures. Cette étude propose une modélisation des cyclones tropicaux dans la région étudiée et la distribution de probabilité des cyclones associés à leur force, permettant ainsi de tenir compte du climat actuel pour modéliser les coûts. Cette modélisation synthétique des cyclones permet de tenir compte des changements climatiques ayant eu lieux dans les trente dernières années et ainsi de ne pas se référer aux données historiques suffisamment récentes pour être extrapolées dans le futur mais trop peu nombreuses pour une étude statistique. Avec les données liées aux infrastructures (nombre d'étages, matériau principal de construction, etc.) et des courbes de fragilités données, nous établissons un lien entre le vent maximal subi pendant un cyclone et les pertes sur chaque résidence. Ainsi, nous pouvons connaître le coût des cyclones, y compris pour les événements extrêmes de très faible probabilité. Nos résultats montrent que les risques importants concernent des événements arrivant tous les cents voire mille ans. Le troisième chapitre propose une extension d'un émulateur statistique des rendements agricoles selon des variables climatiques. L'étude est réalisée sur huit cultures : le manioc, les arachides, le millet, les légumineuses, le colza, la betterave sucrière, la cane à sucre et le tournesol. Nous modélisons l'impact de l'accroissement marginal de la température, des précipitations ou de la concentration en CO2 en faisant une estimation statistique sur des modèles de culture et non sur des données historiques. Cela permet de prendre en compte des effets extrêmes sur des valeurs météorologiques pas ou peu observées jusqu'à présent. La robustesse du modèle est évaluée, entre autres, à l'aide de copules pour comparer la dépendance spatiale entre le modèle et notre émulateur statistique et vérifier que notre estimation capture bien la dépendance géographique.

Introduction

This thesis is an empirical approach of risk management and meteorological hazard with a focus on developing countries cases. The three chapters consist in empirical studies applied to agriculture and natural disasters. In many respects, uncertainty is a major obstacle to development for low-income countries especially when their economy mostly relies on agriculture. Setting apart the income generated by agriculture, developing countries exposed to natural disasters such as tropical storms are facing an additional major vulnerability : because of meteorological hazard generating sizeable damages on local population and infrastructures, their slow economy is worsened at the pace of natural disasters.

Regardless of the development of the country, uncertainty is always a factor against development. Because the future is unpredictable, it is hard for the economic agents to anticipate the outcomes of their decisions. In particular in developing countries, it has been showed that people tend to be risk-adverse, that is agents who prefer having a lower return for sure than an expected return including risk. Consequently, when economic agents in developing countries face uncertainty for commodity prices, damages caused by tropical storms, agricultural losses due to droughts or floods worsened by climate change, etc., they are willing to find a protection against the risk. This protection consists in having a certain income instead of a higher but riskier income. Often, an option to this issue is possible by the mean of an insurance that offers indemnity in case of a loss, in exchange for premiums.

While insurance is common in developed countries for most types of losses, developing countries are facing a serious lack of financial products against risk. An explanation for this is the institutional aspect : the overall political and economical context often

does not allow for implementing competitive private or public insurance companies. The political failure to establish an efficient insurance scheme can be explained by different factors such as political instability, corruption or unfavorable diplomatic relations with neighbored countries. This only adverse climate is enough to deter policy makers from taking any action to provide financial tools against risk. The economical obstacles are, among others, the lack of information and facilities to establish a business company. Moreover, markets from remote areas or geographically hardly accessible are not well integrated to each others, which jeopardizes transport, interregional purchases and more generally national actions. On a technical aspect, it is often too costly to estimate in the fields losses following a disaster, especially for correlated risks, which makes the insurance unaffordable. This thesis proposes innovative solutions to tackle the issue of affordability and feasibility.

A key parameter in estimating the feasibility of a loss insurance is correlation. The cost of insuring a loss substantially different depending on the risk correlation among the insured : for a highly correlated cost, like for flooding in a restricted area, losses will occur at the same time. In this case, the insurer has to provide all the payouts at the same time, which implies a sizeable reserve. Because of the opportunity cost and reinsurance fees, holding enough reserve to fulfill its obligations is costly. Consequently, it is more expensive to insure a correlated loss rather than losses occurring independently (like health insurance, for instance). This parameter is essential when dealing with natural disasters insurance at a localized area in the globe. For example the CCRIF (Caribbean Catastrophe Risk Insurance Facility), that is providing a certain coverage to multiple countries in the Caribbean zone, is subject to a number of studies estimating the extreme losses occurring with a very low probability. Usually in developing countries, the losses that should be primarily covered are agriculture and commodity products, infrastructures and health.

This thesis tackles the issue of spatial dependence for different risks or predictions in developing countries or at a broader scale. These studies imply the use of statistical tools called copulas to model multivariate dependency allowing variant correlation depending on the values.

In the first chapter, we look at spatial correlation of maize prices among markets in Tanzania. We examine the potential for implementing a risk-sharing mechanism among markets to smooth extreme prices variations. Indeed, potential price chock in the future in a serious threat for farmers who cannot anticipate safely their income in the future and change their production accordingly. Here, negative correlation is an advantage compared to a null or positive correlation because, by integrating the markets of the country, we expect to reduce the variance by averaging the prices. This type of scheme has an interest only if the dependence among the markets is opposed enough to compensate prices fluctuations. We study maize prices over 13 years for the 20 main markets in Tanzania. We model the price volatility and the dependence structure among the 20 regions using a multivariate DCC-GARCH model. From this model, we can generate data for another time period and see if the extreme price variations can be smoothed by pooling regions together. To assess the benefits of pooling regions, we take the price returns from our simulation and compute the Value-at-Risk 1% and the Conditional Value-at-Risk 1%, two values returning respectively the threshold exceeded by 1% of the highest (or lowest) data values, and the average value of the data exceeding this threshold. We run our calculations on both increasing and decreasing prices. We base our results on a similar methodology : we compare the VaR or CVaR from the multivariate model and the average VaR or CVaR from the univariate models, that is without accounting for the correlation. We then use the difference between the multivariate and univariate model to draw our conclusions. Our first main result is that when increasing the number of regions pooled together, the extreme price fluctuations are more and more smoothed. More explicitly, when pooling more and more regions (from two to ten), the difference between the multivariate and univariate models for CVaR 1% increases significantly, meaning that the risk decreases. Our second main result is that it is already efficient to pool only two regions together if they are properly paired up. We show that pooling a market from the North and a market from the South of the country is on average more efficient in reducing the risk, and we provide the ranking of the best pairings.

In the same vein, the second chapter is examining the dependence structure of infrastructures damages in the South Pacific islands with tropical cyclones. The final purpose is similar to the CCRIF : since the South Pacific islands are very exposed to

tropical cyclones, providing insurance coverage for losses generated by these storms is essential. However, as explained above, a key parameter in premiums calculation is the losses for an extreme event happening very rarely (say one-in-100-year event). Omitting this potential loss could lead to the incapacity to pay the owed indemnities. In this chapter, we model losses for extreme events occurring with a different probability. We model losses by using synthetical storms rather than using historical data so as to account for the impact of climate change on hurricane frequency and strength. This model provides the maximum wind speed experienced per storm. From this maximum wind speed, we can derive the losses on each building using fragility curves. Having the probability of occurrence and the damage associated of each storm, we can calculate the expected damage for each island. Our study is restricted to six islands since the wind field simulations provide non null damages for six islands out of the initial group of 14 islands. From our calculations, we get a data set of 1000 storms with their probability of occurrence and damages for each building (localized on each island). We fit the losses distribution to three Archimedeans copulas : the Frank, Gumbel and Clayton copulas. From these estimation, we can generate any number of losses simultaneously for the six islands. From this larger data set, we compute the Value-at-Risk (or return period) and Conditional Value-at-Risk at different confidence levels. We find that losses are drastically increasing when dealing with extremely rare tropical storms : for instance, with the Gumbel and Clayton copulas, a 1000-year storm would generate respectively losses equal to 103MM and 113MM USD. When adding an estimation of storm surge, we find 1,207MM and 332MM for the same copulas and events.

The third chapter provides a statistical estimation of crop yields in the future at the grid-cell level. It uses crop models predictions for the next century and estimates the impact on yields of a marginal increase of weather data (temperature, rainfall and CO_2 concentration). The main motive for providing a statistical emulator rather than simply using the crop models already available is mostly the ease of use and public availability. The crop emulator fits the predictions of the GGCM (Global Gridded Crop Model) with a multiple regression for each crop and each climate model. In this paper, the eight crops studied are cassava, millet, sunflower, sugar cane, sugar beet, groundnut/peanut, rape seed and pulses, and the two climate models are GCM (General Circulation Models) taken for the RCP8.5, ie. the worse case scenario in terms of climate change. We

test the accuracy of the predictions from the statistical emulator by comparing the dependence structure of yields predictions in Africa from the emulator and from the crop model. This accuracy is measured by common tools assessing the goodness-of-fit of an estimation, such as RMSE or out-of-sample goodness-of-fit, but also using tools not commonly used on this purpose. Capturing the spatial dependence in our estimation is of the essence when considering implementing new policies or financial tools to redistribute crops production depending on the stringency of climate change effects of the different zones in Africa. To this end, we also look at the spatial dependence of crop yields using copulas. Here, using Vine copulas, we fit a dependence structure to our statistical emulator on the one hand and to the crop model on the other hand. We generate simulations from our two models and look at the similarity of the two models. We show that the dependence structure across Africa is captured by the statistical emulator, ie. we can still predict the unequal impact of climate change on crop yields in the different areas of Africa, and probably more generally, across the globe.

Chapitre 1

Spatial Correlation among Maize Markets in Tanzania : a Risk Analysis

Introduction

Maize accounts for the major part of calorie intake and national production in Tanzania. Food security and farmers income strongly rely on maize consumption (Wilson, Lewis, 2015)[17], eg. on maize affordability and availability. Importantly, since it is a rain-fed crop produced on a large proportion by small-scale farmers, it is likely to be very exposed to weather variability and climate change consequences (Arndt & al., 2012)[1]. Like in most rural areas affected by poverty, Tanzanian farmers living with very small income has tend to be risk-adverse (Mosley, Verschoor, 2005 ; Yesuf, Bluffstone, 2007)[10, 18]. Protecting farmers against maize price volatility is therefore an essential challenge.

This paper aims at exploring maize prices correlation across the main markets in Tanzania. We want to proof that, if price movements are weakly correlated, there is potential for sharing the risk of extreme price increase or decrease by integrating the markets. We focus on maize prices co-movement by using a Copula-GARCH model

that represents prices volatility by integrating pairwise correlations.

A major obstacle to protect farmers in developing countries against price and/or weather shocks is that the risk is often highly correlated. However, the principle of insurance is risk pooling, as defined by Mehr, Cammack and Rose (1985) : *"Insurance may be defined as a device for reducing risk by combining a sufficient number of exposure units to make their individual losses collectively predictable."*[9] Hence, insuring a strongly correlated risk means for an insurance company to have gathered a reserve large enough since all the payouts will occur simultaneously, and keeping an important reserve is costly (Goussebaïle, Louaas, 2017)[5]. Crop insurance is very costly and leaves little room to private insurance for three main reasons : moral hazard, adverse selection, and high administrative costs (Knights and Coble, 1997 ; Skees et al., 1997 ; Goodwin and Smith, 1995)[8, 15, 4]. With a high demand for weather hazard protection in developing countries, publicly provided crop yield insurances have been implemented with heavy government subsidization and have failed mainly because of managerial issues (Hazell, Pomareda, Valdez, 1988 ; Skees, Hazell, Miranda, 1999)[6, 16]. It has been shown for the case of China that spatial diversification of weather risk could substantially lower the premium price by reducing the buffer load (Okhrin, Odening, Xu, 2013)[12]. In parallel, a study made for the US case shows that efficient risk pooling is possible and effective for a private crop insurance market if farmers insure their production for a minimum amount (Wang, Zhang, 2013)[**holly**].

Commodity prices are particularly affected by volatility for three major market fundamentals (FAO, 2011)[14] :

— agricultural output varies depending on seasonal, weather and disease parameters ;
— the demand for agricultural product is particularly inelastic, allowing prices to vary a lot without affecting the demand ;
— in case of important price shift, supply cannot respond immediately because production takes considerable time in agriculture.

Volatility becomes a serious threat to food security when consumers have a low access to food. On the contrary, extremely low prices becomes an issue to farmers. Since a major part of consumers are also farmers in Tanzania, both price movements

pose a significant threat to food security, as illustrated by for the 2007-2008 food crisis (FAO, 2011)[**faoprice**]. This observation is particularly true for less processed food and is exacerbated in countries where social support remains virtually nonexistent. Another key observation is that for traditional crops (such as millet, sorghum and cassava), domestic prices have been more volatile during the crisis than marketed crops (such as rice, wheat and maize). This is mainly explained by a lower reliability in traded crops : the local food consumption of this variety of maize comes almost only from local production, which is more vulnerable to weather and production fluctuations. However in Africa, maize could almost be considered as a traditional crop since African consumers eat white maize while the variety of maize traded on the world market is yellow maize. Maize prices are consequently very prone to volatility in Tanzania.

Our first result shows that the risk sharing among the regions is decreasing when aggregating more and more regions. When comparing the Conditional Value at Risk of the multivariate model to the univariate model, we decrease the risk by 0.058 when pooling only two regions, and by 0.157 for increasing prices 0.169 for decreasing prices by pooling 10 regions. These benefits reaped by increasing the number of regions pooled show that the risk is not very correlated. This correlation might at least not be a major obstacle to the implementation of a risk sharing mechanism. Consequently, one can infer from these results that we have a low probability of facing extreme price changes at the same time for several regions. All the regions being very unlikely to be affected by the same level of price volatility simultaneously, there is room for sharing the risk of maize price volatility.

The second main result states that pooling only two regions with different climatic conditions is sufficient to reduce the Value at Risk. Indeed, pooling several regions relatively remote from each other in a same program can lead to high administrative costs and/or can require a substantial amount of time to be fully implemented. We compare the first and the last percentiles where we pool two regions taken in the same area where the risk is higher (i.e. the South) and where we pool two regions on opposite areas (North and South). Considering decreasing prices, the average first percentile per region of the first case is equal to -0.20 and for the second -0.16. When taking two regions in the less risky area (the North), the first percentile is reduced to -0.14, but the gain obtained by reducing this percentile in the Southern area seems to offset the loss.

The paper will be presented as followed : the next section describes the context of maize in Tanzania and data used in the study. In the second part, we outline the econometric method used to model prices. Finally, we show and explain the results obtained in the third part.

1.1 Context and data

1.1.1 Maize in Tanzania : an essential crop

Maize is a major staple food mainly produced by small-scale farmers, and is hence at the core of food security and agricultural challenges in Tanzania (Wilson, Lewis, 2015)[17]. Additionally, with climate change increasing meteorological variability, particularly in Sub-Saharan African countries, crops may be strongly affected by weather shocks in the future.

Representing half of the country's total calorie intake, maize is the most important food staple and a key crop ensuring food security in Tanzania. It is cultivated both as a cash and a food crop : Tanzania is an important exporter of maize to neighboring countries, although it is noteworthy that between 65 and 80% of all maize production is consumed within the producing household. Roughly 85% of producers are small-scale farmers holding a land of less than one hectare with poor infrastructures. [1] In particular, productivity is on average very low, with less than one ton produced per hectare, compared to 5 tons per hectare on average globally. [2] Moreover, being a rain-fed crop, maize is very sensitive to weather variation and its producers are therefore highly vulnerable to extreme weather events.

The southern zone of Tanzania, including Iringa, Mbeya, Songea and Sumbawanga, is the main production zone, accounting for one third of the total production in the country (Baffes, Kshirsagar, Mitchell, 2015)[2]. Despite its interesting production surplus for exportation and local food security, it is a relatively isolated area, with a lack of trans-

1. http://www.world-grain.com/articles/news_home/World_Grain_News/2016/03/Tanzania_to_increase_corn_prod.aspx?ID=\%7BE02287E6-D168-45FE-BA49-88C62E02B6CD\%7D&cck=1

2. USDA : http://usda.mannlib.cornell.edu/

portation network toward the other centers. Hence, it is very unlikely that trade could offset a one-off production loss in a remote region. Regarding marketed maize, the production is accumulated by traders who sell on regional and urban markets (Wilson, Lewis, 2015)[17]. Although this represents a sizable advantage to consumers, this market system can represent a disadvantage for producers who could probably sell at a much higher price without this intermediary. Despite the income loss, it ensures the sale of the production, which is a serious advantage for farmers working in remote areas with inefficient transport network and store infrastructures.

To date, no affordable financial tool has been accessible to farmers. Yet, to avoid food shortages and substantial losses, it is of particular importance to ensure farmers a minimum income, even in case of a weather shock. A key feature in the design of an insurance is to have a number of contracts large enough to mutualize the risk, which is valid under the condition that the risk is weakly correlated. On the opposite, a highly correlated risk event, such as natural disasters at the country scale, is much more difficult to ensure because an important part of the insured will require indemnities at the same time.

1.1.2 Objective of the study

Several tools exist to protect farmers against losses caused by weather anomalies, and particularly extreme weather events such as droughts. Among other tools, like microcredits, insurance is an interesting option and weather-based index insurances are being developed to facilitate their implementation in places where the lack of information and affordable financial access prevail. More precisely, our study would be a feasibility assessment to implement a revenue weather index insurance that would integrate a price reference to protect farmers against price volatility. By taking into account the prices in the payout, the level of indemnities in generally lower than that of a basic weather index insurance (Mulangu, 2015)[11].

As a result of repeated drought events in the country, food security is weak and the government has taken steps to improve weather anomaly resilience. For example, in 1999 the Tanzania Meteorological Agency has been created to provide weather fore-

casts and contribute to the establishment of early warning systems. Other programs, often managed by the United Nations Programs for Development, have also emerged so as to limit the impacts of weather variation on yields and food security. The northern and central regions, including Arusha, Manyara, Shinyanga, Simiyiu and Dodoma, are more often affected by droughts (Osima)[13].

As mentioned previously, an important condition for a risk to be insurable is to be weakly correlated. The aim of this study is to assess the extent to which prices co-move together and to what extent a significant increase in price in a given region can spread to other ones. By assessing the risk of facing extreme decreases or increases, the results of this price correlation can be considered as a preliminary condition to implement a revenue or income insurance related to natural weather disasters. A weak correlation would indicate a high potential for a national insurance scheme, whereas high correlation among prices would be very expensive to insure.

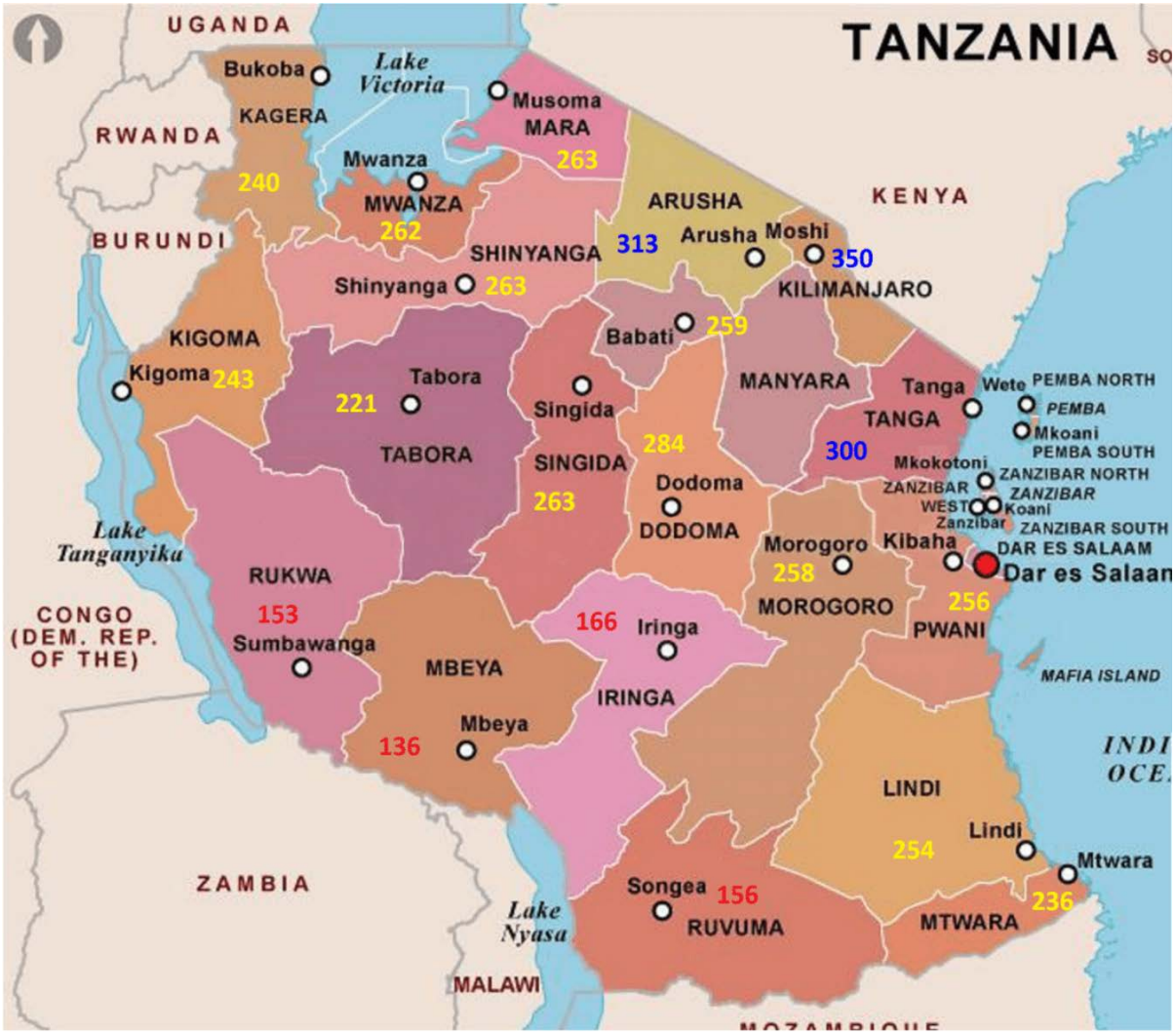
To address this issue, we study price correlation with a Copula-GARCH model. This combines Generalized Autoregressive Conditional Heteroskedasticity model introduced by Bollerslev in 1986 with an improvement in the correlation feature by allowing conditional correlation to be time-varying (Engle, 2001)[**engledcc**]. This model provides the advantage of taking account of the volatility of prices combined with a powerful tool measuring correlation. Indeed, copulas (Joe, 1997)[7] will add to the DCC GARCH model a correlation varying along the values of price change (in particular, we want to include the fact that prices could be more strongly correlated when they are soaring or droping). Copula-GARCH models are commonly used in financial econometrics on stock market applications (Jondeau, Rockinger, 2006) to assess portfolio returns. However, the application of this model to crop insurance is still weak : correlation has mainly been modeled between different crops (Zhu, Ghosh, Goodwin, 2008) without including geographical features, and spatial approaches have been done through linear correlation (Wang, Zhang, 2003). The meteorological aspect will be mainly dealt through linear regressions and empirical copulas.

1.1.3 Data

To estimate the model we rely on monthly maize price time series over 19 years (from January 1995 to December 2013) for the 20 main regions of Tanzania. Prices are deflated with respect to the year 2010 and are given for the main markets of each region. Figure 1.1 gives an example of the distribution of maize prices over the market by showing the prices at a given period (May 2005). Clearly, there is substantial variation in prices across regions, especially between the southern and the northern zones, which would be encouraging for pooling risk in prices. One possible explanation for the variation is that the transportation network is not efficient enough to distribute equally the production among the regions. This price difference is mainly due to the fact that, as mentioned before, the southern zone produces substantially more than the other regions. Consequently, this difference in quantity leads to an equilibrium with higher prices for the less productive regions.

To focus on price volatility rather than an average price (which is likely to move because of seasonal or political reasons), one can convert prices into returns, that is : $log(\frac{P_t}{P_{t-1}})$. Often used in financial econometrics, returns are an interesting tool to highlight prices dynamic. The right side of Figure 1.1 shows the example of the time series of the monthly returns obtained for the markets of Babati and Kigoma. They are respectively in the northern and the central zone and have similar levels of production. However, price dynamics have different trends and these prices seem to be affected by shocks at different times.

FIGURE 1.1: Maize deflated prices in May 2005 for the main cities of the 20 studied regions showing different price levels depending on the geographic areas



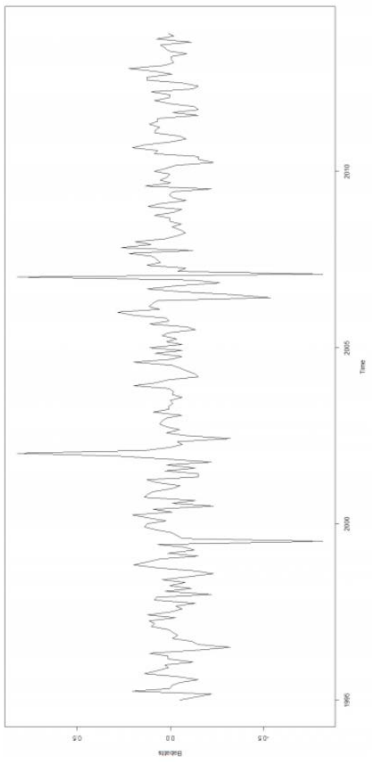


FIGURE 1.2: Monthly maize price returns over 19 years in Babati

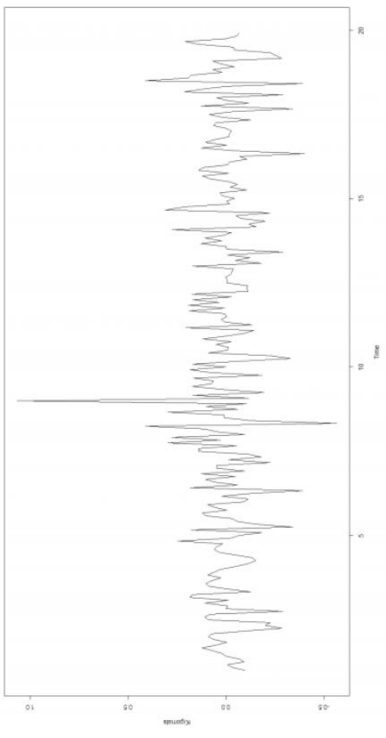


FIGURE 1.3: Monthly maize price returns over 19 years in Kigoma

1.2 Method : modeling correlation through copulas

1.2.1 Modeling price volatility : the GARCH model

General principle of the GARCH model In the general case the univariate GARCH(p, q) model estimates the conditional variance of ϵ_t, σ_t^2, allowing dependence with the squared residuals in the previous p periods and the conditional variance in the previous q periods. The model is built by letting the ϵ_t's be innovations in a linear regression of the form :

$$\begin{aligned} y_t &= x_t' b + \epsilon_t \\ \epsilon_t &= z_t \, \sigma_t \\ \epsilon_t \mid \Omega_{t-1} &\sim \mathcal{N}(0, \, \sigma_t^2) \\ \sigma_t^2 &= \omega + \sum_{j=1}^{p} \alpha_j \, \epsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \, \sigma_{t-j}^2 \\ \omega > 0, \quad \alpha_j \geq 0, \quad \beta_j \geq 0, \quad & \sum \alpha_j + \sum \beta_j < 1 \end{aligned}$$

where y_t is the dependent variable, x_t a vector of explanatory variables, and b a vector of unknown parameters (Bollerslev, 1986). z_t is the surprise term, or the standardized error term, having zero mean and unit variance, often assumed to follow a normal distribution. The case where $q = 0$ corresponds to an ARCH(p) model.

In financial econometrics, the GARCH(1,1) is used and easier to handle ; it is defined, for each time-series i, as :

$$\sigma_{it}^2 = \omega_i + \alpha_{i1} \, \epsilon_{i(t-1)}^2 + \beta_{i1} \, \sigma_{i(t-1)}^2$$

A "Dynamic Conditional Correlation" (DCC) GARCH model The DCC GARCH(1,1) model consists of modeling the 20 time-series using a univariate GARCH(1,1) for each variable and estimate the covariance matrix $\mathbf{H_t}$:

$$\mathbf{H_t} = \mathbf{D_t} \mathbf{R_t} \mathbf{D_t} = \rho_{ij} \sigma_{it} \sigma_{jt}$$

where $\mathbf{D_t}$ is the $k \times k$ diagonal matrix of the time-varying standard deviations from univariate GARCH models with σ_{it} on the i^{th} diagonal ($\mathbf{D_t} = diag(\sigma_{1t}, ..., \sigma_{nt})$), and $\mathbf{R_t}$ is the time-varying correlation matrix.

1.2.2 Multivariate copulas

Presentation of copulas

Definition. Let $(X_1, ..., X_n)$ be a random vector with continuous marginals. The random vector $(U_1, ..., U_n) = (F(X_1), ..., F(X_n))$ has uniformly distributed marginals. The copula of $(X_1, ..., X_n)$ is defined as the joint cumulative distribution function of $(U_1, ..., U_n)$:

$$C(u_1, ..., u_n) = P[U_1 \leq u_1, ..., U_n \leq u_n]$$

Gaussian and Student-t copulas As mentioned before, the use of a copula model can be very interesting in our correlation approach since copulas can provide key information on tail dependence and hence on the probability that prices drop or soar together.

The Gaussian copula (or Normal copula) is constructed from a multivariate normal distribution. Fitting our data to a Normal copula has important consequences on tail dependence : the tail dependence of a Gaussian copula tends to zero for extreme values of one variable.

As a comparison, a fitting estimation to the Student-t copula can be run. One of the characteristics of this copula is to allow the presence of tails.

Tail dependence Before presenting the results, a brief definition of tail dependence is necessary (Aas, 2004)[3] : bivariate tail dependence measures the amount of dependence in the upper and lower quadrant tail of a bivariate distribution. For instance, the upper tail distribution quantifies the probability to observe a large Y , assuming that X is large :

$$\lambda_u(X, Y) = \lim_{\alpha \to 1} P(Y > F_Y^{-1}(\alpha) \mid X > F_X^{-1}(\alpha))$$

and similarly the lower tail is defined by

$$\lambda_l(X, Y) = \lim_{\alpha \to 0} P(Y \leq F_Y^{-1}(\alpha) \mid X \leq F_X^{-1}(\alpha))$$

1.2.3 The Copula DCC GARCH model

The DCC GARCH model is an extension of the GARCH model in that it allows for the correlation to be time-varying thanks to the linear regression model presented previously. This feature represents a strong improvement in forecasting price volatility, however it does not include the possibility of correlation depending on volatility values. Therefore, the copula DCC GARCH will be used so as to bring a new method to calculate the time-varying correlation matrix : here, Gaussian and Student copulas are used to estimate the parameters. These elliptical copulas have a conditional correlation $\mathbf{R_t}$ and constant shape parameter τ.

1.2.4 Assessing the extreme variations of prices with the Value-at-Risk and Conditional Value-at-Risk

Calculating the first and the last percentiles on the returns will be useful to assess the extent to which prices can reach extreme values at the same time. To this end, we first simulate prices returns for the next 19 years for the markets considered. We then generate one time series corresponding to the average of the simulated price returns of these markets. Finally, we compute the distribution function of this new time series and look at the first and last percentiles. For example, by focusing on the 99th percentiles, we have an estimation of the joint probability to face soaring prices of maize in several regions. We then compare the average first and last percentiles per region at the same confidence level by including an increasing number of regions and looking how the threshold for the highest increase in price evolves.

1.3 Results and conclusions

1.3.1 Preliminary results on price correlation

Pearson's correlation coefficient This coefficient measures the linear correlation between two variables and is commonly used in financial economics. Even though it does not allow time-varying correlation and does not take volatility into account, it is interesting to have an overview of the average correlation among the twenty markets over time. Ranging from 0.01 to 0.70 for the extreme highest value, the major part of

the results are mainly ranging between 0.20 and 0.45 (see appendix 1.3.5), which is relatively low and leaves room for pooling the price risk.

Preliminary results on tail dependence with empirical multivariate copulas Tail dependence can be measured empirically, with a non-parametric model giving the conditional probability of observing the same price change between two markets. We can divide this analysis into 3 cases :

- two neighboring markets in the less productive zones, that is the northern and the central zones, which are more likely to be affected by droughts and hence shortages
- two neighboring markets in the most productive zone in the South, where prices can decrease because of unusually high yields
- two markets from a different zone

In all cases, the conditional probability for the two markets is represented by calculating the probability of observing the same price for several levels of price, more or less likely. Hence, the conditional probabilities for values close to 0 represents the likelihood of observing simultaneously important prices decrease, and conversely for high price rises. We are therefore able to see to what extent extreme values can spread to another market depending on price change.

The illustration of the tail dependence of each one of the three categories shows very different trends : the two cases where the markets are close to each other highlight asymmetric tail dependence while the case with two markets far from each other has a smoother joint distribution. Figure 1.4 represents the first case with the example of Arusha and Babati both in the northern zone : even if the conditional probability for low values is higher than the tail of a Normal distribution, a stronger dependence is observed for high price increases. Figure 1.5 illustrates the second case with the example of Songea and Mtwara : it is clear here that important price decreases are highly correlated with a fat lower tail. Lastly, Figure 1.6 represents the third case with a northern market, Arusha, and a southern market, Songea : the distribution is much smoother and does not show fat tails.

FIGURE 1.4: Tail dependence for maize prices between two markets in the northern zone (Arusha and Babati)

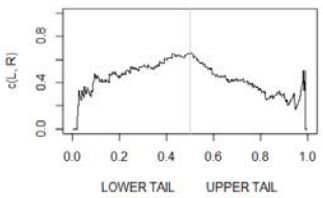


FIGURE 1.5: Tail dependence for maize prices between two markets in the southern zone (Songea and Mtwara)

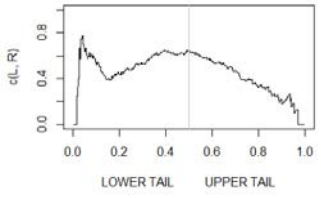
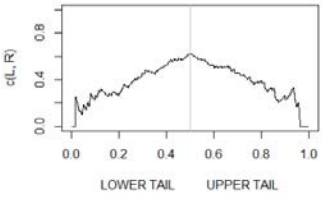


FIGURE 1.6: Tail dependence for maize prices between two markets in different regions (Arusha and Songea)



We can infer from these graphs not only that the correlation is asymmetric but also that the correlation depends on whether the markets are from the same geographic region. Prices seems sensitive to quantities : regions where food shortages are more likely to be observed are more correlated for prices spikes, and regions which have good levels of production have prices changes strongly correlated when they substantially decrease. However, the case of two markets in different regions shows a low correlation with almost no tails. Hence, our historical data show a sizable correlation among markets geographically close, while the different production zones seem to have prices uncorrelated enough to allow for a price related insurance at the national level.

The Copula GARCH fitting (see Appendix at section 1.3.6) To estimate the relevance of the use of a GARCH model in our time series, we first run a test on the autoregression of the terms (we assume here that no variable can explain the autoregressive trend). The Ljung-Box test confirms that there is dependence between the terms at time t and $t-1$, thus using a GARCH model is consistent with our data. For the following models, that is for the GARCH, DCC GARCH and Copula DCC GARCH models, we always run the same Ljung-Box test on the residuals to check that the model has well captured the autoregressive trend : in all our simulations, this trend is captured by the model, with more significant results for the most appropriate models such as the Copula DCC GARCH model.

To make a choice between the different copulas in the Copula DCC GARCH model

(that is, between the Normal, the symmetric and the asymmetric Student-t copulas), we use in addition information criteria often employed for this kind of study : the Akaike, Bayes, Shibata and Hannan-Quinn criteria. We select the model with the lower values for these criteria, and may also use the elapsed time if the values are close to each other. Regarding these features, we have chosen to keep the Copula DCC GARCH model for modeling the distribution of the residuals with an asymmetric Student-t copula. The following results are presented for simulations made with that model.

1.3.2 Main results : reducing the risk of facing extreme price volatility by pooling regions across Tanzania

Defined as the maximum loss not exceeded with a given probability over a given period of time, the Value at Risk will determine in our study the maximum price increase not exceeded for several confidence levels (here we will focus on the VaR 1%). Since the VaR presents the major disadvantage of focusing on the threshold and not accounting for the risk distribution beyond the threshold, we also look at the Conditional Value-at-Risk (CVaR) which is the expected return for returns beyond the VaR. This tool presents the advantage of measuring risk compared to the VaR since it computes the same threshold as the VaR but also computes the expectancy of the values above the VaR. While the VaR omits the distribution of the extreme values, the CVaR differentiates highly aggregated values to spread values and potential extreme values. Hence, we will be able to assess how prices co-move together for an important increase in maize price and see whether the co-movement is alleviated when pooling the markets.

Pooling the risk by randomly adding more and more regions We first show how the risk is reduced for increasing and decreasing prices by including an increasing number of regions. This enables us to see the potential benefit of pooling the risk of facing extreme shifts in maize price at the same time in Tanzania to design an insurance. For this part of the study, we employ the Conditional Value-at-Risk.

Due to computation cost, we are unable to compute the CVaR for all the possible combinations of n markets among 20 : as from 3 regions, that would imply fitting from

1,140 to 184,756 multivariate GARCH models, which is computationally highly demanding. We therefore restrict our comparisons by taking randomly 190[3] different combinations among all the possible and look at the difference in the average CVaR 1% when the n markets are aggregated and when they are isolated. In other words, for the same combination of n markets among 20, we look at the difference between the CVaR 1% for the multivariate GARCH model and the average CVaR 1% of the n univariate GARCH models. We repeat the procedure for 190 possible combinations of n regions and take the average difference between the two CVaR calculated for each combination.

Once this average difference between the aggregate and the isolated cases is obtained for each value of n ranging from 2 to 10 (ie. taking from 2 to 10 markets among 20), we compare the differences and see whether the risk is significantly reduced when increasing the number of regions. If the average difference between the two CVaR is significantly higher with n markets than with $n-1$, hence aggregating n significantly reduces the risk of extreme price volatility compared to aggregating $n-1$ markets.
A sample of 190 combination over thousands different possible combinations is so restricted that the CVaR 1% could vary substantially from a set of n regions to another. Hence, in order to be able to compare the comparable, one should restrict the analysis by comparing the CVaR reduction within the same set of regions.

Table 1.1 and table 1.2 summarize the main results : the CVaR displayed are the average Conditional Value-at-Risk 1% for the price increase and decrease and for the markets taken randomly. Since we are not able to take every possible combination for each n, our sample necessary has a risk trend that does not reflect exactly the average risk profile off every possible combination of n markets among 20. For this reason, we won't analyze the results obtained with the CVaR only. However, when we look at the difference, for every combinations picked up among all the possibilities, between the CVaR for n combinations and the average CVaR with univariate GARCH models, we can infer from the results whether the risk aggregation reap benefits. We calculate $CVaR_{multivariate} - CVaR_{univariate}$ for decreasing prices and $CVaR_{univariate} - CVaR_{multivariate}$ for increasing prices. Hence, a positive difference shows that, for a

3. 190 corresponds to the number of possible combinations of taking 2 elements among 20

19-year time scale, pooling n regions is less risky and that the risk has potential benefits to be exploited. The significance corresponds to the comparison between the average difference with n markets and $n-1$ markets. In the case $n=2$, we compare the CVaR from the multivariate model with the univariate model.

TABLE 1.1: Conditional Value-at-Risk (CVaR, or Expected Shortfall) for increasing prices for an increasing number of regions

Number of markets included	CVaR	Difference with univariate model	Significance
2	0.284	0.058	**
3	0.241	0.084	***
4	0.236	0.107	***
5	0.214	0.109	
6	0.212	0.116	***
7	0.223	0.134	***
8	0.205	0.130	*
9	0.211	0.141	***
10	0.227	0.157	***

Note : Significance evaluated with a $t-test$ between the average difference with n markets and $n-1$ markets. . Significant at 10% level. * Significant at 5% level. ** Significant at 1% level. *** Significant at 0.1% level.

Table 1.1 and table 1.2 display the results for the Conditional Value-at-Risk 1%. Both tables show that the evolution of the CVaR is irregular when adding more and more regions. This is attributable to the fact that we have selected only a few number of combinations among all the possible ones, so the risk profiles are taken randomly. However, the difference between the CVaR of the multivariate and the univariate models is very clear : increasing from 0.058 with 2 regions to 0.157 for increasing prices and 0.169 for decreasing prices with 10 regions, the difference between the aggregated risk and the average individual risk is increasing. The multivariate model have a CVaR decreasing with more regions included compared to the average CVaR of each region considered.

These results show that groups of regions taken randomly with an increasing number of regions have a risk profile less exposed to price volatility risk. Another result that we will use for the next calculations is that pooling two regions is already beneficial in terms of reducing the risk. The significance calculated for the case $n = 2$ measures the significance of the difference between the CVaR with two aggregated regions and the average CVaR with the two same regions taken independently.

TABLE 1.2: Conditional Value-at-Risk (CVaR, or Expected Shortfall) for decreasing prices for an increasing number of regions

Number of markets included	CVaR	Difference with univariate model	Significance
2	-0.280	0.058	***
3	-0.251	0.087	***
4	-0.229	0.099	***
5	-0.226	0.117	***
6	-0.210	0.119	
7	-0.245	0.142	***
8	-0.212	0.137	.
9	-0.218	0.143	**
10	-0.245	0.169	***

Note : Significance evaluated with a $t-test$ between the average difference with n markets and $n-1$ markets. . Significant at 10% level. * Significant at 5% level. ** Significant at 1% level. *** Significant at 1% level

Pooling the risk taking into account the geographic distribution of the regions We now focus on the geographic impact on prices correlation : since the regions are unequal in terms of quantity produced, we know that the mean level of prices follows a similar trend depending on the area. Moreover, since Tanzania has different climatic zones, and assuming climate has a significant impact on prices, one can expect that prices will be differently correlated for regions in different areas. However, we do not know if these effects will be significant enough to overcome a common price movement depending on policy, trade, etc. Therefore, we will proceed with the study on tail dependence in subsection 1.3.1 and see if the correlation is higher for extreme price variations in regions within the same area. If so, aggregating a small number of regions

FIGURE 1.7: Difference between Conditional Value-at-Risk 1% for aggregated regions and independent regions for increasing prices with increasing number of regions included

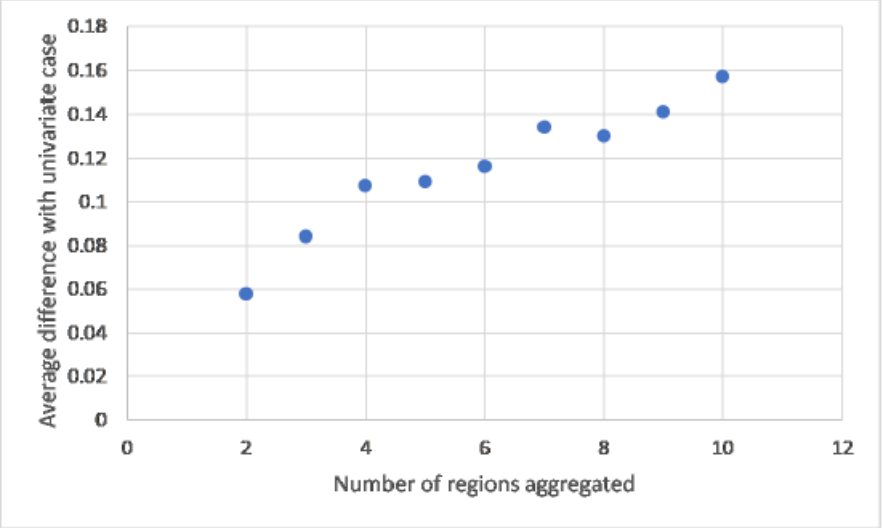
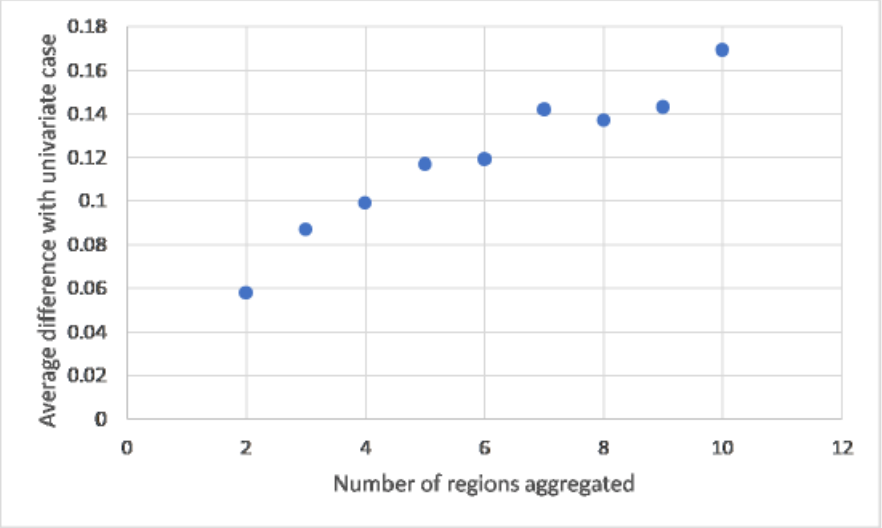


FIGURE 1.8: Difference between Conditional Value-at-Risk 1% for aggregated regions and independent regions for increasing prices with increasing number of regions included



located in different areas would be interesting to reap substantial benefits.

Table 1.5 shows the lower and upper VaR 1% per region by pooling the risk of extreme price co-movement between two regions. We compare the results by taking two regions randomly within the same area : the northern, southern or central area of Tanzania.

TABLE 1.3: Average VaR 1% per region within bimodal area

Number of markets included	Price decrease	Price increase
1	-0.301	0.332
2	-0.171***	0.200*
3	-0.160	0.200
4	-0.111	0.155*
5	-0.128	0.143
6	-0.126	0.130
7	-0.124	0.130
8	-0.111	0.130
9	-0.119	0.126

Note : Significance evaluated with a $t - test$ between the mean with n markets and $n - 1$ markets. * Significant at 10% level. ** Significant at 5% level. *** Significant at 1% level.

Pooling two regions, one from the northern zone and the other one from the southern zone, significantly reduces the average VaR in absolute value compared to the situation where we take two regions within the southern zone (from 20.3% to 16.2%). However, here, it is not sufficient enough to lower the VaR until the rate obtained when the two regions are taken randomly across Tanzania (15%).

Optimization : selecting the combinations of 2 regions that reduces best the Value-at-Risk Having observed that pooling two regions was sufficient to reduce significantly the CVaR and VaR in absolute value, we now look for the ”best” combination. For this part, we look at extreme variation in both decreasing and increasing prices. We can consider the ”best” combination of two regions by looking at the combination where

TABLE 1.4: Average VaR 1% per region within unimodal area

Number of markets included	Price decrease	Price increase
1	-0.445	0.417
2	-0.305**	0.304**
3	-0.248***	0.268
4	-0.255	0.232**
5	-0.238	0.246
6	-0.221	0.241
7	-0.217	0.219*
8	-0.220	0.226
9	-0.217	0.229
10	-0.205	0.224
11	-0.215	0.207*

Note : Significance evaluated with a $t - test$ between the mean with n markets and $n - 1$ markets. * Significant at 10% level. ** Significant at 5% level. *** Significant at 1% level.

the CVaR and VaR are the smallest in absolute value. In table 1.6 and table 1.7, we show the ten pairs of regions having the less risky profile when pooled together. That is, we run a GARCH model for every possible combination of two regions in Tanzania and select the pairs of regions where the CVaR 1% and VaR 1% for increasing or decreasing prices is the lowest in absolute value. The least risky pairs of regions are Arusha and Mwanza for decreasing prices with a CVaR 1% of -0.050 and a VaR 1% of -0.0609, and Arusha and Mwanza for increasing prices with a CVaR 1% of 0.055 and a VaR 1% of 0.0727. Mbeya and Moshi seems also to be a good composition for both decreasing and increasing prices regarding the VaR 1%, respectively equal to -0.074 and 0.073. Mbeya and Moshi are on opposite sides of the country, which makes the lack of correlation between the two prices times series relevant. However, Arusha and Mwanza are relatively close to each other, which means that in a scheme that would involve crop transportation, here transportation costs would not offset the advantages in terms of risk in pooling those two regions.

TABLE 1.5: Average Values-at-Risk 1% taking 2 regions from different zones

	2 regions taken			
	Central zone	Northern zone	Southern zone	randomly
	-0.231	-0.152	-0.334	-0.232
Price decrease		Northern and Southern zones		
		-0.231		-0.232
	0.226	0.151	0.305	0.358
Price increase		Northern and Southern zones		
		0.249		0.358

TABLE 1.6: Least risky combinations

Decreasing prices		Increasing prices	
Pair of regions	CVaR 1%	Pair of regions	CVaR 1%
Arusha Mwanza	-0.050	Arusha Mwanza	0.055
Arusha Morogoro	-0.062	Arusha Morogoro	0.064
Bukoba Mwanza	-0.092	Musoma Mwanza	0.103
Musoma Mwanza	-0.099	Bukoba Mwanza	0.105
Dodoma Mwanza	-0.107	Dodoma Mwanza	0.118
Arusha Bukoba	-0.109	Arusha Bukoba	0.119
Morogoro Mwanza	-0.111	Morogoro Mwanza	0.120
Bukoba Tabora	-0.131	Bukoba Tabora	0.141
Bukoba Dodoma	-0.135	Dar-es-Salaam Mwanza	0.143
Dar-es-Salaam Mwanza	-0.135	Arusha Dar-es-Salaam	0.146

By calculating the minimum of the CVaR and VaR, we are very likely to find several times the same markets in the combinations simply because they are not very exposed to price volatility over time. Hence, we might not capture the benefit from pooling the risk with another market compared to the baseline scenario. Consequently, we can consider the best combination as being the one that improves best the CVaR or VaR reduction. Here, we calculate the difference between the CVaR (VaR) when pooling two regions with the average CVaR (VaR) of each region taken separately. Table 1.8

TABLE 1.7: Least risky combinations

Decreasing prices		Increasing prices	
Pair of regions	VaR 1%	Pair of regions	VaR 1%
Arusha Mwanza	-0.0609	Mbeya Moshi	0.0727
Mbeya Moshi	-0.0735	Arusha Mwanza	0.0733
Arusha Morogoro	-0.0738	Arusha Morogoro	0.0749
Mbeya Morogoro	-0.0850	Mbeya Morogoro	0.0851
Arusha Bukoba	-0.104	Bukoba Morogoro	0.110
Bukoba Morogoro	-0.104	Arusha Bukoba	0.111
Bukoba Mbeya	-0.109	Bukoba Mbeya	0.113
Bukoba Lindi	-0.138	Bukoba Lindi	0.139
Arusha Mbeya	-0.138	Arusha Mbeya	0.140
Dodoma Mwanza	-0.139	Dodoma Mwanza	0.144

and table 1.9 show the combinations that reduces the best the CVaR 1% (VaR 1%) for increasing and decreasing prices. The CVaR (VaR) difference is obtained by computing the CVaR 1% (VaR 1%) for the bivariate GARCH model and substracting it to the average CVaR 1% (VaR 1%) of the univariate GARCH models for the two regions considered. Mbeya and Moshi appear again in the best combinations for decreasing prices, which means that not only are they the less risky duo, but they are also the duo that reap the most benefits when paired together with a VaR reduction of 0.335 points, and a CVaR reduction of 0.645. On the opposite, it seems that the most performant combinations are very different for increasing prices depending on whether we look at the CVaR or the VaR : the pair Mbeya and Shinyanga reap substantial benefits when paired together by reducing the VaR of 0.594 points (even though they did not appear in the least risky regions) and the combination Mbeya Singida performs the best when considering the CVaR with a decrease of 0.676 points. This result can be mainly explained by the different season pattern of the two regions.

TABLE 1.8: Best combinations that decrease the risk (CVaR)

Decreasing prices		Increasing prices	
Pair of regions	CVaR difference	Pair of regions	CVaR difference
Mbeya Singida	0.676	Mbeya Songea	0.619
Mbeya Songea	0.645	Mbeya Mtwara	0.608
Mbeya Moshi	0.645	Songea Singida	0.605
Lindi Mbeya	0.627	Bukoba Mbeya	0.599
Mbeya Tanga	0.620	Mbeya Tanga	0.593
Bukoba Mbeya	0.605	Lindi Mbeya	0.591
Babati Mbeya	0.582	Mbeya Moshi	0.575
Dar-es-Salaam Mbeya	0.579	Mbeya Morogoro	0.568
Mbeya Mtwara	0.577	Babati Mbeya	0.548
Dodoma Mbeya	0.573	Dar-es-Salaam Mbeya	0.545

Note : VaR difference in absolute value obtained by subtracting the VaR 1% obtained for the bivariate GARCH model and the mean of VaR 1% obtained for the two univariate GARCH models for each region.

Conclusion

We have been able to highlight a correlation weak enough among the twenty major markets in Tanzania to consider the possibility of mutualizing the risk of extreme maize price changes. The feasibility of implementing such a scheme and the cost of the potential insurance would be very high if the risk was extremely correlated, since the insurance company would be forced to gather important reserves. Our Copula-GARCH model, applied on our twenty time-series, takes into account the pairwise correlations and assesses a relatively low probability of facing extreme price changes in several regions simultaneously. Including more and more regions helps reducing significantly the risk but we have also shown that pooling two markets from different areas can be sufficient to mutualize the risk. Given the divergence in terms of climatic and production conditions, the southern area is less prone to climatic variability, diversifying the price risk between the northern and the southern areas. This conclusion is interesting to find an alternative to a national scheme that could have non negligible implementing and administrative costs.

TABLE 1.9: Best combinations that decrease the risk (VaR)

Decreasing prices		Increasing prices	
Pair of regions	VaR difference	Pair of regions	VaR difference
Mbeya Moshi	0.335	Mbeya Shinyanga	0.594
Bukoba Mbeya	0.271	Mbeya Sumbawanga	0.477
Mbeya Morogoro	0.271	Songea Sumbawanga	0.0958
Bukoba Lindi	0.240	Kigoma Mbeya	0.0431
Mbeya Singida	0.201	Arusha Mtwara	0.0387
Arusha Mbeya	0.199	Arusha Lindi	0.0373
Singida Songea	0.199	Mtwara Mwanza	0.0338
Bukoba Morogoro	0.180	Morogoro Mtwara	0.0329
Dodoma Mbeya	0.180	Arusha Sumbawanga	0.0264
Arusha Morogoro	0.169	Lindi Morogoro	0.0263

Note : VaR difference in absolute value obtained by subtracting the VaR 1% obtained for the bivariate GARCH model and the mean of VaR 1% obtained for the two univariate GARCH models for each region.

Appendix

1.3.3 The GARCH model

The general GARCH model The model is built by letting the ϵ_t's be innovations in a linear regression of the form :

$$\begin{aligned} y_t &= x_t' b + \epsilon_t && (1.1)\\ \epsilon_t &= z_t \, \sigma_t && (1.2)\\ \epsilon_t \mid \Omega_{t-1} &\sim \mathcal{N}(0, \, \sigma_t^2) && (1.3)\\ \sigma_t^2 &= \omega + \sum_{j=1}^{p} \alpha_j \, \epsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \, \sigma_{t-j}^2 && (1.4) \end{aligned}$$

$$\omega > 0, \quad \alpha_j \geq 0, \quad \beta_j \geq 0, \quad \sum \alpha_j + \sum \beta_j < 1$$

where y_t is the dependent variable, x_t a vector of explanatory variables, and b a vector of unknown parameters (Bollerslev, 1986). z_t is the surprise term, or the standardized error term, having zero mean and unit variance, often assumed to follow a normal distribution. The case where $q = 0$ corresponds to an ARCH(p) model.

The first two unconditional moments are constant, their expression does not include a time component :

$$\begin{aligned} E(\epsilon_t) &= 0 && (1.5)\\ E((\epsilon_t - E(\epsilon_t))^2) &= \frac{\omega}{1 - \sum \alpha_j - \sum \beta_j} && (1.6) \end{aligned}$$

The first conditional moment is, by definition, equal to zero and the second conditional moment is time dependent :

$$\begin{aligned} E(\epsilon_t \mid \Omega_{t-1}) &= 0 && (1.7)\\ E((\epsilon_t - E(\epsilon_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) &= \omega + \sum_{j=1}^{p} \alpha_j \, \epsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \, \sigma_{t-j}^2 && (1.8) \end{aligned}$$

In financial econometrics, the GARCH(1,1) is usually used and easier to handle ; it is defined, for each time-series i, as :

$$\sigma_{it}^2 = \omega_i + \alpha_{i1} \, \epsilon_{i(t-1)}^2 + \beta_{i1} \, \sigma_{i(t-1)}^2$$

GARCH properties proofs

$$E((\epsilon_t - E(\epsilon_t))^2) = \frac{\omega}{1 - \sum \alpha_j - \sum \beta_j}$$

Démonstration. Let us define :

$$\epsilon_t^2 = E(\epsilon_t^2 \mid \Omega_{t-1}) + v_t \tag{1.9}$$

with v_t being the surprise term such that : $E(v_t \mid \Omega_{t-1}) = 0$ and $E(v_t) = 0$. By definition, $E(\epsilon_t^2 \mid \Omega_{t-1}) = \sigma_t^2$, hence we can write :

$$\epsilon_t^2 = \sigma_t^2 + v_t$$

hence :

$$E(\epsilon_t^2) = E(\sigma_t^2) \tag{1.10}$$

Therefore, one can re-write the variance expression :

$$\begin{aligned} \sigma_t^2 &= \omega + \sum \alpha_j \, \epsilon_{t-j}^2 + \sum \beta_j \, \sigma_{t-j}^2 \\ \iff \epsilon_t^2 - v_t &= \omega + \sum \alpha_j \, \epsilon_{t-j}^2 + \sum \beta_j \, \sigma_{t-j}^2 \\ \implies E(\epsilon_t^2) &= \omega + \sum \alpha_j \, E(\epsilon_t^2) + \sum \beta_j \, E(\sigma_t^2) \end{aligned}$$

since $E(v_t) = 0$. And thanks to equation 1.10 :

$$\begin{aligned} \iff E(\epsilon_t^2) &= \omega + \sum \alpha_j \, E(\epsilon_t^2) + \sum \beta_j \, E(\epsilon_t^2) \\ \iff E(\epsilon_t^2) &= \omega + E(\epsilon_t^2) \sum \alpha_j + E(\epsilon_t^2) \sum \beta_j \\ \iff E(\epsilon_t^2) &= \frac{\omega}{1 - \sum \alpha_j - \sum \beta_j} \\ \iff E((\epsilon_t - E(\epsilon_t))^2) &= \frac{\omega}{1 - \sum \alpha_j - \sum \beta_j} \end{aligned} \tag{1.11}$$

□

$$E((\epsilon_t - E(\epsilon_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = \omega + \sum_{j=1}^{p} \alpha_j \, \epsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \, \sigma_{t-j}^2$$

Démonstration. By definition, $E(\epsilon_t \mid \Omega_{t-1}) = 0$, hence :

$$\begin{aligned} E((\epsilon_t - E(\epsilon_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) &= E(\epsilon_t^2) \mid \Omega_{t-1}) \\ &= E(\sigma_t^2 \, z_t^2 \mid \Omega_{t-1}) \\ &= \sigma_t^2 \, E(z_t^2 \mid \Omega_{t-1}) \end{aligned}$$

since σ_t^2 can be expressed as a function of elements depending only on the period $t-1$, this term is constant. And since the moments of z_t do not change over time, we have $E(z_t^2 \mid \Omega_{t-1}) = E(z_t^2)$ with $z_t \sim \mathcal{N}(0,1)$, then we get :

$$\begin{aligned} \sigma_t^2 E(z_t^2 \mid \Omega_{t-1}) &= \sigma_t^2 \\ &= \omega + \sum_{j=1}^{p} \alpha_j \, \epsilon_{t-j}^2 + \sum_{j=1}^{q} \beta_j \, \sigma_{t-j}^2 \end{aligned} \tag{1.12}$$

□

Coefficient of the DCC GARCH model In the DCC GARCH model, the dynamic conditional correlation coefficient between two markets (1 and 2) at time t is expressed as follows :

$$\rho_{12,t} = \frac{E_{t-1}(\epsilon_{1,t} \, \epsilon_{2,t})}{\sqrt{E_{t-1}(\epsilon_{1,t}^2) \, E_{t-1}(\epsilon_{2,t}^2)}}$$

1.3.4 Copulas properties

Sklar's theorem. *Let F* $\in \mathcal{F}(F_1, ..., F_n)$ *be an n-dimensional distribution function with marginals* $F(x_1), ..., F(x_n)$. *Then there exist a copula C :* $[0;1] \to [0;1]^n$ *such that, for all* $\boldsymbol{x} = (x_1, ..., x_n) \in R^n$:

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$$

If $F_1, ..., F_n$ *are continuous, then C is unique.*

The Gaussian copula (or Normal copula) is constructed from a multivariate normal distribution. Each parameter ρ is estimated thanks to the following bivariate formula :

$$C_\rho(u,v) = \int_{-\infty}^{\phi^{-1}(u)} \int_{-\infty}^{\phi^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \, exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dx \, dy$$

ρ being the parameter of the copula and $\phi^{-1}(.)$ the inverse of the standard univariate Gaussian distribution function.

Fitting our data to a Normal copula has important consequences on tail dependence : as shown in the next paragraph, the tail dependence of a Gaussian copula tends to zero for extreme values of one variable.

As a comparison, a fitting estimation to the Student-t copula can be run. The parameters are estimated according to the following formula :

$$C_{\rho\nu}(u,v) = \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \left(1 + \frac{x^2 - 2\rho xy + y^2}{\nu(1-\rho^2)}\right)^{-\frac{(\nu+2)}{2}} dx\, dy$$

One of the characteristics of this copula is to allow the presence of tails (see next paragraph).

Tail dependence Tail dependence is zero for a Normal copula :

$$\lambda_l(X,Y) = \lambda_u(X,Y) = 2 \lim_{x \to -\infty} \phi\left(x \frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}\right) = 0$$

hence it is very likely that if prices soar or increase in a given market because of a localized event, other markets prices will not be significantly impacted. Prices are very unlikely to co-move together for extreme shifts.

For the Student-t copula, the presence of tail is allowed and is calculated as follows :

$$\lambda_l(X,Y) = \lambda_u(X,Y) = 2\, t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$$

where $t_{\nu+1}$ corresponds to the distribution function of a univariate Student's t-distribution with $\nu + 1$ degrees of freedom. In this case, the estimate for the degree of freedom (ν) provides the main information : the lower degree of freedom, the higher the tail dependence.

Matrix correlation between raw returns

	Arusha	Babati	Bukoba	Dar	Dodoma	Iringa	Kigoma	Lindi	Mbeya	Morogoro	Moshi	Mtwara	Musoma	Mwanza	Shinyanga	Singida	Songea	Sumbawanga	Tabora	Tanga
Arusha	1.00	0.33	0.18	0.53	0.56	0.39	0.23	0.12	0.31	0.51	0.48	0.21	0.34	0.20	0.42	0.48	0.25	0.23	0.20	0.56
Babati	0.33	1.00	0.16	0.38	0.45	0.25	0.15	0.35	0.29	0.50	0.17	0.46	0.30	0.26	0.30	0.26	0.43	0.39	0.37	0.45
Bukoba	0.18	0.16	1.00	0.18	0.02	0.15	0.12	0.12	0.12	0.08	0.18	0.10	0.42	0.24	0.34	0.14	0.13	0.30	0.16	0.08
Dar	0.53	0.38	0.18	1.00	0.60	0.52	0.34	0.21	0.39	0.61	0.46	0.32	0.24	0.32	0.37	0.41	0.45	0.43	0.40	0.44
Dodoma	0.56	0.45	0.02	0.60	1.00	0.56	0.20	0.32	0.36	0.70	0.43	0.36	0.20	0.34	0.39	0.53	0.41	0.29	0.44	0.55
Iringa	0.39	0.25	0.15	0.52	0.56	1.00	0.13	0.20	0.45	0.52	0.33	0.28	0.09	0.21	0.38	0.44	0.39	0.35	0.35	0.41
Kigoma	0.23	0.15	0.12	0.34	0.20	0.13	1.00	0.37	0.23	0.27	0.25	0.30	0.24	0.21	0.41	0.37	0.26	0.39	0.52	0.01
Lindi	0.12	0.35	0.12	0.21	0.32	0.20	0.37	1.00	0.28	0.35	0.21	0.64	0.18	0.38	0.40	0.28	0.41	0.37	0.53	0.15
Mbeya	0.31	0.29	0.12	0.39	0.36	0.45	0.23	0.28	1.00	0.43	0.17	0.35	0.03	0.23	0.38	0.35	0.41	0.44	0.38	0.19
Morogoro	0.51	0.50	0.08	0.61	0.70	0.52	0.27	0.35	0.43	1.00	0.30	0.43	0.19	0.35	0.39	0.38	0.49	0.41	0.40	0.56
Moshi	0.48	0.17	0.18	0.46	0.43	0.33	0.25	0.21	0.17	0.30	1.00	0.01	0.32	0.17	0.29	0.53	0.12	0.19	0.37	0.23
Mtwara	0.21	0.46	0.10	0.32	0.36	0.28	0.30	0.64	0.35	0.43	0.01	1.00	0.18	0.29	0.34	0.10	0.58	0.44	0.36	0.36
Musoma	0.34	0.30	0.42	0.24	0.20	0.09	0.24	0.18	0.03	0.19	0.32	0.18	1.00	0.40	0.37	0.27	0.10	0.20	0.20	0.28
Mwanza	0.20	0.26	0.24	0.32	0.34	0.21	0.21	0.38	0.23	0.35	0.17	0.29	0.40	1.00	0.45	0.30	0.19	0.19	0.33	0.12
Shinyanga	0.42	0.30	0.34	0.37	0.39	0.38	0.41	0.40	0.38	0.39	0.29	0.34	0.37	0.45	1.00	0.45	0.38	0.38	0.49	0.28
Singida	0.48	0.26	0.14	0.41	0.53	0.44	0.37	0.28	0.35	0.38	0.53	0.10	0.27	0.30	0.45	1.00	0.31	0.28	0.50	0.29
Songea	0.25	0.43	0.13	0.45	0.41	0.39	0.26	0.41	0.41	0.49	0.12	0.58	0.10	0.19	0.38	0.31	1.00	0.43	0.31	0.32
Sumbawanga	0.23	0.39	0.30	0.43	0.29	0.35	0.39	0.37	0.44	0.41	0.19	0.44	0.20	0.19	0.38	0.28	0.43	1.00	0.47	0.32
Tabora	0.20	0.37	0.16	0.40	0.44	0.35	0.52	0.53	0.38	0.40	0.37	0.36	0.20	0.33	0.49	0.50	0.31	0.47	1.00	0.11
Tanga	0.56	0.45	0.08	0.44	0.55	0.41	0.01	0.15	0.19	0.56	0.23	0.36	0.28	0.12	0.28	0.29	0.32	0.32	0.11	1.00

TABLE 1.10: Correlation among markets prices (Pearson's correlation coefficient)

1.3.6 GARCH goodness-of-fit

Stationarity To use an autoregressive model, the time-series must be stationary, i.e with statistical properties such as mean, variance, autocorrelation, etc. all constant over time. To check for stationarity, we use an Augmented Dickey-Fuller test and obtain the results below :

Market	Statistics	p.value
Arusha	-6.97	0.01
Babati	-7.83	0.01
Bukoba	-6.87	0.01
Dar	-6.87	0.01
Dodoma	-6.69	0.01
Iringa	-6.57	0.01
Kigoma	-5.60	0.01
Lindi	-8.28	0.01
Mbeya	-6.95	0.01
Morogoro	-7.39	0.01
Moshi	-6.44	0.01
Mtwara	-8.49	0.01
Musoma	-6.61	0.01
Mwanza	-7.24	0.01
Shinyanga	-6.65	0.01
Singida	-7.55	0.01
Songea	-8.07	0.01
Sumbawanga	-7.78	0.01
Tabora	-6.83	0.01
Tanga	-6.96	0.01

TABLE 1.11: Augmented Dickey-Fuller test results for the 20 markets

The p-values are all lower than 0.01 so the 20 time-series are stationary. With this condition, we can run a GARCH model.

Autocorrelation To justify the use of an autoregressive model, we need to identify an autocorrelation trend on the variance of returns (ie. the squared residuals). To this end, we run a Portmanteau test (Ljung, Box, 1978) on each times series :

Market	Statistics	p-value
Arusha	0.16	0.688
Babati	14.06	0.000
Bukoba	4.14	0.0419
Dar	3.27	0.070
Dodoma	12.57	0.000
Iringa	0.19	0.666
Kigoma	0.07	0.787
Lindi	7.80	0.005
Mbeya	16.43	5.05e-05
Morogoro	5.15	0.023
Moshi	10.03	0.002
Mtwara	0.06	0.813
Musoma	0.61	0.434
Mwanza	7.51	0.006
Shinyanga	3.41	0.065
Singida	24.58	7.12e-07
Songea	0.44	0.506
Sumbawanga	4.01	0.045
Tabora	2.53	0.112
Tanga	10.98	0.001

TABLE 1.12: Portmanteau test results on the maize price returns

13 times series over 20 have a low p-value (lower than at least 0.05), so we observe an autoregressive trend with one lag for 13 markets in Tanzania. Hence, the use of a GARCH model is justified.

The goodness-of-fit of the selected GARCH model will determine whether the model

has a stochastic trend. Therefore, if the GARCH model fits well to our data, we should not observe any autoregression in the residuals. We check this condition by running the same Portmanteau test as previously on the squared residuals of the DCC GARCH model and of the Copula DCC GARCH model (choosing the Normal copula) :

Market	Statistics	p-value
Arusha	3.57	0.059
Babati	0.00	0.978
Bukoba	2.05	0.152
Dar	0.01	0.925
Dodoma	0.01	0.913
Iringa	0.01	0.937
Kigoma	1.01	0.315
Lindi	0.27	0.601
Mbeya	2.38	0.123
Morogoro	3.58	0.058
Moshi	3.24	0.072
Mtwara	3.90	0.048
Musoma	0.13	0.718
Mwanza	0.38	0.539
Shinyanga	1.69	0.193
Singida	3.41	0.065
Songea	5.03	0.025
Sumbawanga	1.56	0.212
Tabora	0.34	0.561
Tanga	0.07	0.788

TABLE 1.13: Portmanteau test results on the DCC GARCH model

All the p-values have substantially increased, indicating no autoregressive trend remaining in the residuals and that both models are correct. However, for the second, all the p-values are higher than 0.05 and generally higher than those of the first model (the linear DCC GARCH model). Consequently, the Copula DCC GARCH model selected fits well to our data.

Market	Statistics	p-value
Arusha	2.66	0.103
Babati	1.81	0.179
Bukoba	1.11	0.292
Dar	0.08	0.774
Dodoma	0.05	0.820
Iringa	0.20	0.658
Kigoma	0.04	0.840
Lindi	0.06	0.812
Mbeya	0.12	0.724
Morogoro	0.07	0.785
Moshi	3.18	0.074
Mtwara	0.80	0.370
Musoma	0.50	0.478
Mwanza	0.03	0.868
Shinyanga	0.05	0.823
Singida	0.31	0.578
Songea	0.82	0.365
Sumbawanga	0.44	0.509
Tabora	0.69	0.406
Tanga	0.48	0.490

TABLE 1.14: Portmanteau test results on the Copula DCC GARCH model

Symmetry and tails We can strengthen this robustness looking at the symmetry and the tails in the distribution of the returns and of the model residuals. To this end, we look at the Skewness (for the symmetry) and Kurtosis (for the tails) values, which are respectively equal to 0 and 3 for a Normal distribution.

1.3.7 Values-at-Risk

Significance in the difference in the Values-at-Risk For 10 markets and more, the difference in the Value-at-Risk is not significant.

TABLE 1.15: p-values of the t-tests comparing the average Value-at-Risk pooling n and $n+2$ regions - Case of increasing prices

2 and 4 markets	4 and 6 markets	6 and 8 markets	8 and 10 markets
0.0003178	0.01117	0.008773	0.5286

Bibliographie

[1] C. ARNDT et AL. “Agriculture and Food Security in Tanzania”. In : *Review of Development Economics* 16.3 (2012), p. 378–393.

[2] J. BAFFES, V. KSHIRSAGAR et D. MITCHELL. “Domestic and External Drivers of Maize Prices in Tanzania”. 2015.

[3] Modelling the DEPENDENCE STRUCTURE OF FINANCIAL ASSETS; A SURVEY OF FOUR COPULAS. In : (2004).

[4] B. K. GOODWIN et V. H. SMITH. *The Economics of Crop Insurance and Disaster Aid*. Washington D.C. : The AIE Press, 1995.

[5] A. GOUSSEBAÏLE et A. LOUAAS. “Insurance Market Equilibrium for Correlated Risks”. In : (2017).

[6] P. HAZELL, C. POMAREDA et A. VALDEZ. “Crop Insurance for Agricultural Development: Issues and Experience”. In : *Journal of Development Economics* 28.3 (1988), p. 401–406.

[7] H. JOE. *Multivariate Models and Dependence Concepts*. London : Chapman & Hall, 1997.

[8] T. O. KNIGHT et K. H. COBLE. “Survey of U.S Multiple Peril Crop Insurance Literature since 1980”. In : *Review of Agricultural Economics* 19 (1980), p. 128–156.

[9] R. I. MEHR, E. CAMMACK et T. ROSE. *Principles of insurance*. Illinois : Richard D. Irwin, Inc., 1985.

[10] P. MOSLEY et A. VERSEHOOR. “Risk Attitudes and the ’Vicious Circle of Poverty’”. In : *The European Journal of Development Research* 17.1 (2005), p. 59–88.

[11] F. MULANGU. “How to Use Weather Index Insurances to Address Agricultural Price Volatility”. In : United Nations Conference on Trade et Development. Geneva, Switzerland, 4 2015.

[12] O. OKHRIN, M. ODENING et W. XU. “Systemic Weather Risk and Crop Insurance: the Case of China”. In : *The Journal of Risk and Insurance* 80.2 (2013), p. 351–372.

[13] S. E. OSIMA. *Drought Conditions and Management Strategies in Tanzania*. Rapp. tech. Tanzania Meteorological Agency,

[14] *Price Volatility in Food and Agricultural Markets: Policy Responses*. Rapp. tech. FAO et OECD, June 2011.

[15] J. R. SKEES, J. R. BLACK et B. J. BARNETT. “Designing and Rating an Area Yield Crop Insurance Contract”. In : *American Journal of Agricultural Economics* 79.2 (1997), p. 430–438.

[16] J. R. SKEES, P. HAZELL et M. MIRANDA. “New Approaches to Crop Yield Insurance in Developing Countries”. In : *International Food Policy Research Institute* 55 (1999).

[17] R. TREVOR WILSON et J. LEWIS. “The Maize Value Chain in Tanzania”. In : *FAO* (2015).

[18] M. YESUF et R. BLUFFSTONE. “Risk Aversion in Low-Income Countries: Experimental Evidence from Ethiopia”. In : *American Journal of Agricultural Economics* 91.4 (2009), p. 1022–1037.

W :/agoburdhun/Documents/Dissertation/bibliocorrelation.bib

Chapitre 2

Cyclone Risk Correlation Among the South Pacific Islands

co-written with Ilan Noy and Eric Strobl

2.1 Introduction

Being part of the most affected and damaged countries in the world due to natural disasters (World Bank, 2010, PCRAFI, 2013)[12, 5], the Pacific Island Countries (PICs) struggle to develop and are typically among the most in need for disaster risk coverage. In the same vein as the CCRIF (Caribbean Catastrophe Risk Insurance Facility) formed in 2007 and including 16 Caribbean countries (CCRIF, 2017)[6], the PICs are now joining together to form a multi-country insurance through the PCRAFI (Pacific Catastrophe Risk Assessment & Financing Initiative)(PCRAFI, 2015)[1]. A major obstacle to this type of insurance is the occurrence of extremely rare and devastative event with particularly high payout. These events happen with a very low probability and might be seriously underestimated when calculating the reserve for example. Unlike health or accident insurances for instance, the risk of facing a natural disaster for countries in

the same area is very correlated, hence the insurance scheme cannot rely on risk mutualization to offset the payout when a disaster occurs. Insuring such a risk for several countries very exposed to tropical storm risk results in holding high reserve, which is costly for the insurance scheme (and hence for the member governments).

To date, a strong literature is being developed to model the impacts of natural disasters. Using the same wind field model as Strobl (2012)[22], studies have been carried on to measure the impact of tropical storms in developing countries on the agriculture sector (Blanc and Strobl, 2016 ; Mohan and Strobl, 2017)[3, 18], on properties (Sealy and Strobl)[21] and on economic growth using nightlight imagery (Bertinelli, Mohan, Strobl, 2016)[2]. Our paper brings a contribution to this literature by providing a risk assessment for the PICs.

Our paper consists in modeling the impact of synthetical storms on buildings in the South Pacific islands. Having a series of storms, we can model the risk correlation of facing a devastating tropical storm over the studied islands. While climate change is inducing key changes in meteorological parameters, it is necessary to generate a probability distribution of storm losses similar to the damages faced in the future. Indeed, according to Emanuel (2011), tropical storms are bound to become less frequent and with a higher intensity. In this context, using historical data to generate a density of probability function is very likely to be inaccurate because observed storms in the past may not be similar to the future ones. To this end, we use Boose et al.'s (2004)[4] version of the well-known Holland (1980)[15] wind field model to generate 3000 synthetic storms over the PICs. For each of these storms, we have the maximum wind speed experienced on each building and the probability of occurrence of the storm.

To model the damages on the buildings, we have a data set with a number of variables on each building : exact location, value (in USD), number of stories, frame material and main use (residential or commercial). We use damage functions depending on the maximum wind speed set for different types of building on the Hazus software so as to determine V_{half}, which is the wind speed at which 50% of the building is damaged (V_{thresh}, the maximum wind speed at which the building has no damage, is assumed to be constant). Using Emanuel's (2011) damage function, we are able from our values

of V_{thresh} and V_{half} to determine the fraction of each property damaged by any storm modeled. Since most of the losses caused by a natural disaster are explained by storm surge and fresh water, we also create another data set where we add an estimation of damages due to storm surge (with flood level). We use the estimation made for Hong Kong by Chan and Walker (1979)[7] to infer flood level from maximum wind speed for each building and for each storm.

We fit our data on expected losses to the Gumbel copula, an Extreme Value copula, modeling the higher correlation among the PICs when facing extremely rare and devastating tropical storms. Having null losses over a group of islands, we finally have a data set of five islands : Cook, Niue, Samoa, Tonga and Vanuatu. Having fitted the losses to their joint probability of occurrence over these five PICs, we are able to generate future damages and compute the loss summed over the five islands.

After generating 1000 storm damages on the six islands studied, we compute the losses for different return-periods (equivalent to a Value at Risk) and Conditional Value at Risk for different confidence levels. For the wind damages alone, damages are respectively equal to 3,154M, 3,157M and 682M USD for a 100-year event for the Gumbel, Clayton and Frank copulas, and equal to 103MM, 113MM and 5MM USD for a 1000-year event. For the 0.1% most destructive storms the CVaR is respectively equal to 370MM, 250MM and 210MM USD. By adding an approximation of storm surge damages, the losses for a 1000-year return period event it is equal to 832MM and 1,177MM USD for the Clayton and the Frank copulas. The CVaR 0.1% is respectively equal on average to 970MM and 1,300MM USD (1,000MM USD represents roughly 42% of the GDP [1]GDP 2016 of the six aggregated islands).

2.2 Context and data

2.2.1 Context

Over the last few years, the Pacific Island Countries have been severely affected by extreme natural events, including tsunamis, flooding and earthquakes. Among others, during 2012-2014, the PIC experienced extreme floods and a tsunami in the Solomon

Islands and TC Ian in Tonga following a magnitude 8.0 earthquake. Storm surges and floods affected more recently respectively the Marshall Islands and the Solomon Islands [1]. Tropical cyclones have been the major cause of losses and damage over the last sixty years with damage in excess of US$3.2 billion and more that 9.2 million people affected in the PIC (World Bank, 2012)[24].

Several obstacles have prevented so far the PICs in raising post-disaster liquidity. Being small size countries, they have very few capacity to spread risk within one country and have very small economies. Furthermore, being net importers and having limited access to international insurance market, the PICs have limited options available for post-disaster finance. To this end, a regional scheme similar to the CCRIF is being implemented for all the PICs. The Pacific Disaster Risk Financing and Insurance (DRFI) Program under the Pacific Catastrophe Risk Assessment and Financing Initiative (PCRAFI) aims at discussing solutions for financing disaster resilience and response. Ideally, a self-financed insurance scheme would be help the countries being less dependent to foreign aid and investing in resilient infrastructures.

2.2.2 Data

We are considering in this study 14 islands (or groups of islands) : Cook, Federal States of Micronesia (FSM), Fiji, Kiribati, Marshall, Nauru, Niue, Palau, Samoa, Solomon, Tokelau, Tonga, Tuvalu, Vanuatu.

For each island, our data set includes the main residential and commercial buildings, ranging from 1,106 to 65,535 buildings per island, with in total 343,053 buildings for all the dataset. We can identify the geographical position of each building by its latitude and longitude. A number of structural information on each building are described in table **??** and allow for assessing the exposition to extreme wind speed.

2.3 Wind field model

We use a data set of 3000 synthetic storms over the Pacific islands generated following the method described by Emanuel, Sundararajan, Williams in 2008[11] with, for

TABLE 2.1: Data on buildings

Variable	Description
Latitude, longitude	Values in degrees
Main Occupation	Residential, commercial...
Occupation	General commercial, education, government...
Construction	Frame material
Number of stories	Number
Floor area	Square meters
Att type	Modeled
Value	in USD

TABLE 2.2: Main occupations of the buildings

Value	Frequency
Commercial	89,201
Industrial	5,162
Infrastructure	7,946
Public	36,760
Residential	1,150,860
Others	7,820

each building coordinates, the maximum wind speed and the probability of occurrence of the storm considered. Despite a good knowledge of past hurricanes as far back as 1855, the number of observations is insufficient to generate a probability distribution function and the climate context is obviously very different from the now. To have a more accurate view of the probability of occurrence of storms and their intensity, hypothetical tropical storms are generated using recent historical data within a coupled ocean-atmosphere tropical storm model. The tracks model takes into account a number of global meteorological parameters such as temperature, humidity, wind and sea surface temperature. It uses the a version of the Holland (1980)[15] wind field model strengthened by Boose et al.'s (2004)[4] described below. Several proto-storms are generated at random time and location, and move according to large scale winds of a given climate state. Under certain conditions, some of them develop into full scale hur-

ricanes (ie., after passing the threshold of maximum wind speed equal to 119 km/h). This methodology was here implemented so as to generate 3000 hurricane-strength storms traversing the South Pacific islands using meteorological data of years 1980 - 2010. Each year has a given expected frequency of storms and each synthetic storm is assigned to one of the 30 years climatology ; hence, we can calculate out the annual probability of each possible storm. This model assumes that the climate in the future will be similar to the climate observed during 1980 - 2010 with no more changes in the future. Consequently, the wind field model might underestimate losses based on the assumption that climate change would increase losses in the future.

Since the wind field model (of n storms) does not affect every island, we restrict our study on the islands being affected by at least one storm in our data set. Consequently, we are studying 6 islands for which we non-zero losses over our data set of storms.

The level of wind a field will experience during a passing typhoon depends crucially on that field's position relative to the storm and the storm's movement and features. It thus requires explicit wind field modeling. To calculate the wind speed experienced because of typhoons within each pixel, we use Boose et al.'s (2004)[4] version of the well-known Holland (1980)[15] wind field model. More specifically, the wind experienced at time t because of typhoon j at any point $P = i$, that is, W_{ij}, is given by :

$$W_{ijt} = GF\left\{V_{m,jt} - S\left[1 - \sin(T_{aijt})\right]\frac{V_{h,jt}}{2}\right\} \left\{\left(\frac{R_{m,j,t}}{R_{it}}\right)^{B_{jt}} \times \exp\left[1 - \left(\frac{R_{m,j,t}}{R_{it}}\right)^{B_{jt}}\right]\right\}^{1/2} \tag{2.1}$$

where V_m is the maximum sustained wind velocity anywhere in the typhoon, T is the clockwise angle between the forward path of the typhoon and a radial line from the typhoon center to the pixel of interest, $P = i$, V_h is the forward velocity of the hurricane, R_m is the radius of maximum winds, and R is the radial distance from the center of the hurricane to point P. The remaining ingredients in (2.1) consist of the gust factor G and the scaling parameters F, S, and B, for surface friction, asymmetry due to the forward motion of the storm, and the shape of the wind profile curve, respectively.

Regarding the parameters calculated in equation (2.1), V_m is given by the storm-track data available in our data set, V_h can be directly calculated by following the storm's movements between locations, and R and T are calculated depending on the location of interest $P = i$. All other parameters are estimated or assumed. For example, we have no information on the gust wind factor G and the surface friction F. Nevertheless, a number of studies provide results that we use for these parameters : G can be estimated to 1.5 (as Paulsen and Schroeder 2005[19] show) and Vickery et al. (2009)[25] show that in open water the reduction factor is about 0.7 and reduces by 14% on the coast and by 28% 50 km inland. Hence, we use a reduction factor that linearly decreases between 0 and 50 km inland. B is determined using Holland's (2008)[15] approximation method and R_max using the parametric model estimated by Xiao et al. (2009)[26].

2.4 Calculation of the expected loss per island

2.4.1 Estimation of each building damages from wind data

Wind damage function Emanuel (2011)[10] proposes a damage function providing the fraction of property b_i affected (denoted $frac_{b_i s}$) by a storm s given the minimum wind speed at which damages start having non-zero value (V_{thresh}) and at which half the property is destroyed (V_{half}). This function is made on the assumption that the hurricanes climate changes linearly with time and taking into account the three climate scenarios published by the IPCC with increasing damages.

$$frac_{b_i s}^{wind} = \frac{v_{b_i s}^3}{1 + v_{b_i s}^3}$$

where

$$v_{b_i s} = \frac{max[V_{b_i s} - V_{thresh}, 0]}{V_{half} - V_{thresh}}$$

where $V_{b_i s}$ is the maximum wind experienced at property b_i with storm s. i is the island considered and does not impact our results so far.

Note that V_{half} will vary among our different building categories depending on their building features.

TABLE 2.3: Estimations of V half

		Number of stories						
		1	2	3	4	5	6	8
Timber frame		259	229	a	a	a	a	a
Masonry/Concrete frame		296	399	303	280	259	a	a
Steel frame	Residential	263	266	269	273	277	274	268
	Commercial	219	224	229	235	240	245	256
		a = average wind speed $= 261km.h^{-1}$						
		$V_{thresh} = 92km.h^{-1}$						

Estimation of V_{half} A number of building features are provided in our dataset and can be linked to features used in damage assessments depending on maximum wind experienced. The software HAZUS®[8] models hurricane and provides a manual with fragility curves depending on some key building features, such as walls material, roof material, number of stories, building height, etc. For each case, fragility curves provide a percentage of building loss as a function of wind speed. Hence, by matching our data to a part of building features provided in the manual, we can determine V_{half} for each building category. The values showed on table 2.3 obtained are estimated by hand on the fragility curves drawn on the manual of the HAZUS software. Hence, because we don't have the same number of building features as the ones used for drawing the fragility curves, and because of the error we make in assessing the value by hand, we include a lecture error.

Estimating storm surge impact from our data Tropical cyclones have an impact on buildings through wind field and especially fresh water and storm surge. To date, fresh water and storm surge can only be modeled through specified software with non openly accessible data. Since their impact depends on many different variables and their interactions, it can be hardly estimated with the data used for our wind field model. However, some research has been conducted to estimate a link between variables such as the maximum wind speed experienced or the central pressure, and storm surge (Chan,

Walker, 1979 ; Irish, Resio, Ratcliff, 2008 ; Pradhan, Mitra, De, 2012)[7, 17, 20]. However, since storm surge height is also explained by variables such as the bathymetry, these studies are specific to one region, and results when applied to other regions must be interpreted with caution.

To have an insight of the order of magnitude of the losses due to storm surge, we decide to apply the storm surge height model designed by Chan and Walker (1979)[7] for Hong Kong. Using our results for maximum wind speed experienced by building, we can have a rough estimate of the storm surge height :

$$S_{b_i s} = \begin{cases} 0.3048 \times (0.088 \times \frac{V_{b_i s}}{1.85} - 0.75) & \text{for} \quad V_{b_i s} < 185\ km/h \\ 0.3048 \times (0.00217 \times (\frac{V_{b_i s}}{1.85})^2 + 0.43) & \text{for} \quad V_{b_i s} > 185\ km/h \end{cases}$$

$S_{b_i s}$ being the storm surge height in meters for building b_i in island i with storm s, and $V_{b_i s}$ the maximum wind experienced at property b_i with storm s.

We can then calculate, for each building b_i and storm s, the fraction damaged due to storm surge :

$$frac_{b_i s}^{surge} = \begin{cases} \frac{1}{n_{b_i}} \frac{S_{b_i s}}{3} & \text{if} \quad \frac{S_{b_i s}}{3} < n \\ 1 & \text{if} \quad \frac{S_{b_i s}}{3} \geqslant n \end{cases}$$

with n the number of stories of the building b_i and we assume the average height of a story being equal to 3 meters.

2.4.2 Estimation of the islands expected losses

Wind field damages only We estimate the annual expected loss for each island and for each storm. It is important for our case study not to aggregate the data over the different storms to calculate one unique annual expected loss accounting for every storm probability of occurrence since we need to focus on the loss correlation among the islands. Hence, we will need to have one vector per island with the expected loss associated with each synthesized storm. Having this data matrix will enable us to assess the correlation evolution depending on the level of damages among the islands with copulas (see section 2.5).

We keep the probability of occurrence of each storm to weight the impact of a given storm given the likelihood that it occurs. Nevertheless, this probability of occurrence is the same for every island, the specific impact on each building is specified through the wind field model that design the maximum wind speed for each geographic coordinate.

$$E(L_{is}) = \sum_{b_i} E(L_{b_i s}) = \sum_{b_i} frac_{b_i s}^{wind} \times value_{b_i} \times prob_s$$

i the island considered

s the synthetic storm

b_i the buildings in island i

L_{is} the losses over island i during storm s

$frac_{b_i s}$ fraction of building b_i affected given its properties and wind of storm s

$value_{b_i}$ total value of building b_i

$prob_s$ probability of occurrence for storm s

As mentioned before, since our wind field model outputs provide null damages over the storms simulated for 8 islands (Federal States of Micronesia (FSM), Fiji, Marshall, Nauru, Palau, Solomon, Tokelau and Tuvalu), we are dropping them for the remaining study on correlation. We keep the 6 remaining islands : Cook, Kiribati, Niue, Samoa, Tonga and Vanuatu. One can infer from this observation that risk is relatively weakly correlated among the islands : indeed, we have generated a large number of tropical storms that accounts for extreme event with low probability and we still have null vectors for the 8 islands cited above. We can make the assumption that, whatever the storm size and intensity, the islands are sufficiently spread so that they are not all impacted at the same time.

Wind field damages and estimation of storm surge damages By adding the estimated effect of storm surge, we can compute the new expectancy of losses as follow :

$$E(L_{is}) = \begin{cases} \sum_{b_i} (frac_{b_i s}^{wind} + frac_{b_i s}^{surge}) \times value_{b_i} \times prob_s & \text{if} \quad frac_{b_i s}^{wind} + frac_{b_i s}^{surge} < 1 \\ \sum_{b_i} value_{b_i} \times prob_s & \text{if} \quad frac_{b_i s}^{wind} + frac_{b_i s}^{surge} \geqslant 1 \end{cases}$$

2.5 Risk profiles and correlation between each pair of islands

2.5.1 Presentation of copulas

General concepts

Definition 2.5.1. copula Let $(X_1, ..., X_n)$ be a random vector with continuous marginals. The random vector $(U_1, ..., U_n) = (F(X_1), ...F(X_n))$ has uniformly distributed marginals. The copula of $(X_1, ..., X_n)$ is defined as the joint cumulative distribution function of $(U_1, ..., U_n)$:

$$C(u_1, ..., u_n) = P[U_1 \leq u_1, ..., U_n \leq u_n]$$

Tail dependence Bivariate tail dependence measures the amount of dependence in the upper and lower quadrant tail of a bivariate distribution (Aas, 2004). For instance, the upper tail distribution quantifies the probability to observe a large Y , assuming that X is large :

$$\lambda_u(X, Y) = \lim_{\alpha \to 1} P(Y > F_Y^{-1}(\alpha) \mid X > F_X^{-1}(\alpha))$$

and similarly the lower tail is defined by

$$\lambda_l(X, Y) = \lim_{\alpha \to 0} P(Y \leq F_Y^{-1}(\alpha) \mid X \leq F_X^{-1}(\alpha))$$

Sklar's theorem 1. Let F $\in \mathcal{F}(F_1, ..., F_n)$ be an n-dimensional distribution function with marginals $F(x_1), ..., F(x_n)$. Then there exist a copula C : $[0; 1] \to [0; 1]^n$ such that, for all **x** $= (x_1, ..., x_n) \in R^n$:

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$$

If $F_1, ..., F_n$ are continuous, then C is unique.

Expression of the copulas in the bivariate case

Fitting the copulas to our data Several multivariate copulas exist to model the dependence across the islands. We chose to fit our data to three Archimedean copulas with different dependence patterns, in particular different tail dependence.

Linear correlation The multivariate copulas use the linear correlation parameters to be estimated, hence it is indispensable to estimate the linear matrix correlation as a first step.

Table 2.5 displays the Kendall correlation coefficients for each pair of islands. We can observe that the pair Kiribati / Niue has a negative correlation equal to -0.003 which, even if relatively low in absolute value, will prevent the estimation calculation to work properly with an Extreme Value Copula. Because Extreme Value Copula only handle variables with positive correlation, we can't fit the data to this category of copulas. Consequently, we fit our data to Archimedean copulas, which allow negative correlation. However, we show our results obtained by fitting to an Extreme Value Copula the data set minus the Kiribati data.

Fitting result and robustness We fit our data to three Archimedean of copulas : the Gumbel copula, the Clayton copula and the Frank copula. Table 2.6 summarizes the goodness-of-fit results with the parameters and p-value. As we can see, the Gumbel copula does not fit to our data set as well as the Clayton and Frank copula. Given the linear correlations computed for the six islands displayed in table 2.5, the average correlation over the losses reached for each pair of islands is lower than 0.35 for most of the pairs. The Gumbel copula is an asymmetric copula with greater dependence in the positive tail than in the negative. High correlation for positive values should fit to the extreme values since natural disasters cause highly correlated losses geographically, however we assume that this copula doesn't fit because the rest of the dependence pattern doesn't represent the rest of the values. The Clayton and the Frank copula both fit very well to our data, with a p-value lower than 0.05%. They both represent weaker correlation, the Clayton copula representing greater correlation in the negative tail, and

Copula name	$C(u,v)$	$\phi(t)$	Range of θ
Gumbel	$exp(-log(u)^{-\theta} + log(v) - \theta)^{1/\theta}$	$-log(t)^{\theta}$	$[1;\infty)$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta^{-1}(t^{-\theta} - 1)$	$(0;\infty)$
Frank	$-\frac{1}{\theta} log\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right)$	$log\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$	$(-\infty,\infty)$

TABLE 2.4: Copulas distributions in the bivariate case

TABLE 2.5: Kendall correlation coefficients for the 6 islands studied

	Cook	Kiribati	Niue	Samoa	Tonga	Vanuatu
Cook	1.00	0.04	0.07	0.09	0.34	0.06
Kiribati	0.04	1.00	*-0.003*	0.16	0.08	0.10
Niue	0.07	*-0.003*	1.00	0.11	0.13	0.11
Samoa	0.09	0.16	0.11	1.00	0.15	0.77
Tonga	0.34	0.08	0.13	0.15	1.00	0.12
Vanuatu	0.06	0.10	0.11	0.77	0.12	1.00

the Frank copula representing symmetric dependence. We will comment our results mostly based on these two copulas results.

TABLE 2.6: Goodness-of-fit of the three copulas to the two data sets

	Wind field		Wind field and storm surge	
	Parameter	p-value	Parameter	p-value
Gumbel copula	1.3211	0.066	1.3307	0.2822
Clayton copula	0.6417	0.0005	0.6809	0.0005
Frank copula	1.9254	0.0005	2.0025	0.0005

2.5.2 Return period losses (or Values-at-Risk) and Conditional Values-at-Risk

Return period losses and Value-at-Risk

Definition 2.5.2. A return period, also known as a recurrence interval, is an estimate of the likelihood of an event. The return period R can be expressed as follows :

$$R = \frac{n+1}{m}$$

n number of years on record ;

m number of occurrence of the considered event.

Definition 2.5.3. Value at Risk (VaR) is the maximum loss not exceeded with a given probability defined as the exceedance probability, over a given period of time.

$$VaR_\alpha(L) = inf\{l \in \mathbb{R} : P(L > l) \leq 1 - \alpha\}$$

Having fitted our data to the three Archimedean copulas, we are now able to randomly simulate n values of losses accounting for the dependence among our six variables[9]. From these new vectors generated, we obtain for each simulation an aggregated loss associated for the six islands obtained by adding the simultaneous losses of each island. From this time series giving the aggregated loss for the six islands, we can calculate the Value-at-Risk or losses for several return periods.

TABLE 2.7: Values-at-Risk for different exceedance probabilities : 5%, 2.5%, 1% and 0.01% with 1000 random simulations

VaR	95%	97.5%	99%	99.9%
Return period	20-year	40-year	100-year	1000-year
Gumbel copula	9,455.86	127,641.8	3,154,349	1.03e08
Clayton copula	42,027.58	543,744.9	3,156,899	1.13e08
Frank copula	62,866.98	253,650.7	682,406.5	5,628,445

Table 2.14 highlights that the risk profile of the islands regarding storm damages is very oriented towards low probability of extreme events. Indeed we can observe that the 95th percentile of loss distribution in our simulation is very low : 9,455, 42M and 62M USD respectively for the Gumbel, Clayton and Frank copulas for a 20-year event. Damages are increasing exponentially when dealing with very low probability of large events : damages are respectively equal to 3,154M, 3,157M and 682M USD for a 100-year event, and equal to 103MM, 113MM and 5MM USD for a 1000-year event.

Importantly, one has to note that the fact that our calculations are made with only 6 islands over the 14 initially studied does not mean that we are omitting potential additional loss. Indeed, the reasons why we have removed the other islands were either null damages over the n storms simulated. Consequently, our results can be extended to the 14 islands with very unlikely potential changes.

Conditional Value-at-Risk

Definition 2.5.4. The upper Conditional Value at Risk (CVaR), also called Mean Ex-

cess Loss and Expected Shortfall, is the expected losses strictly exceeding VaR :

$$CVaR_\alpha(L) = E[L_i|L \geq VaR_\alpha]$$

This measure brings two information that the VaR alone does not provide : it accounts for the value of extreme losses and it weights the potential losses by their probability of occurrence. In other words, it takes into account the shape of the losses density function after the threshold given by the VaR.

Table **??** shows the results we obtain for the CVaR for the same exceedance probability as the VaR previously computed. Unsurprisingly, since the extreme losses simulated occur with a very low probability (non-zero losses arising with a probability of less than 5% and increasing significantly with a probability of less than 1%), the CVaR significantly increases with the exceedance probability : the average losses for the 5% most destructive storms is equal to 13MM, 11MM and 4.8MM USD respectively for the Gumbel, Clayton and Frank copulas, and the for the 0.1% most destructive storms is respectively equal to 370MM, 250MM and 210MM USD. 210MM USD represent roughly 9% of the GDP [2] of the six aggregated islands.

TABLE 2.8: Conditional Values-at-Risk for different exceedance probabilities : 5%, 2.5%, 1% and 0.01% with 1000 random simulations

CVaR	95%	97.5%	99%	99.9%
Gumbel copula	13e06	27e06	66e06	370e06
Clayton copula	11e06	21e06	53e06	250e06
Frank copula	4.8e06	9.4e06	23e06	210e06

Measures by adding the estimated effects of storm surge Our estimation presented above allows us to include the significant effect of storm surge in the losses. Here, we keep the six islands taken for the wind field model so as to be able to compare the results (Cook, Kiribati, Niue, Samoa, Tonga and Vanuatu). Since it is calculated with a strong link with the maximum wind experienced for each building, it increases the losses already experienced for each storm and each building with the wind. Tables 2.16 and 2.17 show the estimated losses when accounting for the wind and the storm

2. GDP 2016

surge impacts. Here when adding the effect of storm surge, the losses are much higher even for the average 5% worst storms. Indeed, the losses associated to a 20-year return period event is exceeding 300MM USD for the Clayton and Frank copulas, and for a 1000-year return period event it is equal to 832MM and 1,177MM USD. When looking at the CVaR, the figures are a bit higher with losses occurring with a probability lower than 0.1% respectively equal on average to 970MM and 1,300MM USD. By comparison, 1,000MM USD represents roughly 42% of the GDP [3]GDP 2016 of the six aggregated islands.

TABLE 2.9: Values-at-Risk with wind field and storm surge impacts

	95%	97.5%	99%	99.9%
Return period	20-year	40-year	100-year	1000-year
Gumbel copula	6.7 e06	59 e06	228 e06	1,207 e06
Clayton copula	26 e06	111 e06	338 e06	832 e06
Frank copula	27 e06	144 e06	330 e06	1,177 e06

TABLE 2.10: Conditional Values-at-Risk with wind field and storm surge impacts

	95%	97.5%	99%	99.9%
Gumbel copula	200 e06	370 e06	750 e06	1,800 e06
Clayton copula	190 e06	320 e06	500 e06	970 e06
Frank copula	250 e06	400 e06	750 e06	1,300 e06

2.6 Conclusion

Using a wind field model, we have been able to generate a series of synthetical tropical storms over the PICs with the maximum wind speed for each building and the probability of occurrence of the storms. So as to have an order of magnitude of the combined effect of storm surge, we generate another dataset with the same storms by adding to the losses due to the wind an estimation of storm surge damages. After inferring the expected losses from the damage function for each island, we fit our data to three Archimedean copula, the Gumbel, Clayton and Frank copulas. Copulas let us accounting for a different correlation among the time series when a rare and extreme

disaster occurs. Having modeled the correlation among the six islands (Cook, Kiribati, Niue, Samoa, Tonga and Vanuatu) we kept for our study, we have generated a significant number of observations that allowed us to calculate the losses for several return periods : for the Gumbel, Clayton and Frank copulas, damages are respectively equal to 3,154M, 3,157M and 682M USD for a 100-year event, and equal to 103MM, 113MM and 5MM USD for a 1000-year event. For the 0.1% most destructive storms the CVaR is respectively equal to 370MM, 250MM and 210MM USD. By adding an approximation of storm surge damages, the losses for a 1000-year return period event it is equal to 832MM and 1,177MM USD for the Clayton and the Frank copulas. The CVaR 0.1% is respectively equal on average to 970MM and 1,300MM USD (1,000MM USD represents roughly 42% of the GDP [4]GDP 2016 of the six aggregated islands).

These figures show that a significant part of the GDP of the five PICs studied will be lost due a very extreme tropical cyclone occurring with a very low probability. This loss tends to be under-estimated because it has not been experienced so far and because these events are more complicated to value in an insurance scheme, mostly because of their very low probability of occurrence and the uncertainty associated with the estimation of the loss associated.

2.7 Appendix

FIGURE 2.1: Example of fragility curve drawn from HAZUS and used to determine V_{half}

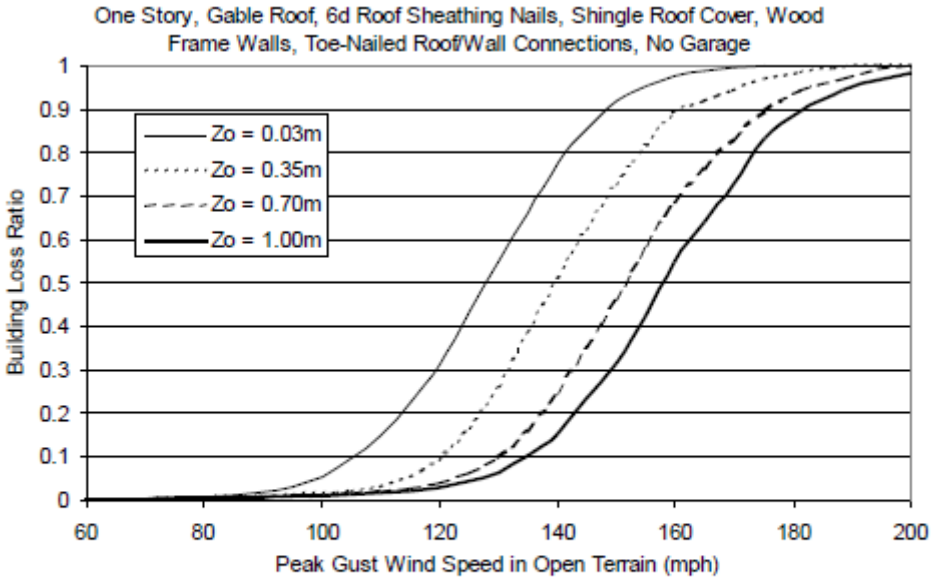


Figure H.1. Building Loss Function for Single Family Residential Building (One Story, 6d Roof Sheathing Nails, Gable Roof, No Garage, Toe-Nailed Roof Wall Connections, Wood Frame).

TABLE 2.11: GDP of the six islands kept for the risk assessment (Source : World Bank)

Country	GDP (MM USD 2016)
Cook	290
Kiribati	181.6
Niue	32.8
Samoa	685.9
Tonga	395.2
Vanuatu	773.5
Total	2,359

2.7.1 Results with an Extreme Value Copula

Extreme value copulas

In the same vein as classic copulas, extreme value copulas have been designed so as to model positive dependence between two or more variables during rare events. Extreme value copulas arise in the domain of extreme value theory and can be mainly applied to finance, insurance and environmental science. We expose here the main theory of extreme value copulas (Gudendorf, Segers, 2010, and Hougaard, 1986)[13, 16].

Let $X_i = (X_{i1}, ..., X_{id})$, $i \in \{1, ..., n\}$ be a sample of independent and identically distributed (iid) random vectors with common distribution function F, margins $F_1, ..., F_d$, and copula C_F.We assume F is continuous and consider the vector of componentwise maxima :

$$M_n = (M_{n,1}, ..., M_{n,d}) \qquad \text{where } M_{n,j} = \max_{1 \le i \le n} X_{ij}$$

and $j \in \{1, ..., d\}$. F^n is the joint distribution function of M_n and $F_1^n, ..., F_d^n$ the marginal distribution functions of M_n. The copula C_n of M_n is given by

$$C_n(u_1, ..., u_d) = C_F(u_1^{1/n}, ..., u_d^{1/n})^n, \qquad (u_1, ..., u_d) \in [0, 1]^d$$

When the sample size n tends to infinity, the limits of copulas C_n belong to the family of extreme value copulas.

Definition 2.7.1. A copula C is called an *extreme value copula* if there exists a copula C_F such that

$$C_F(u_1^{1/n}, ..., u_d^{1/n})^n \to C(u_1, ..., u_d) \qquad (n \to \infty)$$

for all $(u_1, ..., u_d) \in [0, 1]^d$. The copula C_F is said to be in the *domain of attraction* of C.

The Gumbel-Hougaard copula

Theory Consider the Archimedean copula

$$C_\Phi(u_1, ..., u_d) = \Phi^{\leftarrow}(\Phi(u_1) + ... + \Phi(u_d)), \qquad (u_1, ..., u_d) \in [0, 1]^d$$

with function $\Phi : [0, 1] \to [0, \infty[$ and inverse $\Phi^{\leftarrow}(t) = inf\{u \in [0, 1] : \Phi(u) \le t\}$; the function Φ should be strictly decreasing and convex and satisfy $\Phi(1) = 0$, and $\Phi^{\leftarrow}$

should be d-monotone on $[0, \infty[$.

If the following limit exists,

$$\theta = -\lim_{s \to 0} \frac{s\Phi'(1-s)}{\Phi(1-s)} \in [1, \infty[$$

then the domain-of-attraction condition is verified for C_F equal to C_Φ, the tail dependence function being

$$\ell(x_1, ..., x_d) = \begin{cases} (x_1^\theta + ... + x_d^\theta)^{1/\theta} & \text{if } 1 \leq \theta \leq \infty \\ x_1 \vee ... \vee x_d & \text{if } \theta = \infty \end{cases}$$

for $(x_1, ..., x_d) \in [0, \infty)^d$. The parameter θ measures the degree of dependence ranging from independence ($\theta = 1$) to complete dependence ($\theta = \infty$).

The extreme value *Gumbel-Hougaard* copula associated to ℓ is

$$C(u_1, ..., u_d) = exp\{-((-\log u_1)^\theta + ... + (-\log u_d)^\theta)^{1/\theta}\}$$

It is noteworthy that the Gumbel-Hougaard copula is both an Archimedean copula and an extreme value copula at the same time. Extreme Value Copula don't fit data with negative correlation ; we need therefore to drop one of them from the data we will aggregate in our multivariate Gumbel copula. To determine whether we drop Kiribati or Niue, we calculate the correlation matrix with the Pearson coefficient. We see in table 2.12 that, with this other correlation coefficient, Kiribati and Niue are still negatively correlated and in addition, Kiribati has 2 other negative correlations (with Samoa and Vanuatu). Based on this supplementary observation, we make the decision to drop Kiribati.

TABLE 2.12: Pearson correlation coefficients for the 6 islands studied

	Cook	Kiribati	Niue	Samoa	Tonga	Vanuatu
Cook	1.00	0.00	0.00	0.00	0.01	0.00
Kiribati	0.00	1.00	*-0.0008*	*-0.0008*	0.00	*-0.0007*
Niue	0.00	*-0.0008*	1.00	0.43	0.00	0.50
Samoa	0.00	*-0.0008*	0.43	1.00	*-0.0018*	0.95
Tonga	0.01	0.00	0.00	*-0.0018*	1.00	*-0.0018*
Vanuatu	0.00	*-0.0007*	0.50	0.95	*-0.0018*	1.00

Result for fitting the data with wind field to the Gumbel copula We use R software[23] to fit our data to a Gumbel copula with dimension 5. Following the observation of the previous paragraph, we have a dataset of 5 variables corresponding to Cook, Niue, Samoa, Tonga and Vanuatu. We first generate a Gumbel copula of dimension 5 and then use the function `fitCopula` from package `copula`[14] to fit the copula parameter to our data (see appendix).

TABLE 2.13: Results of fitting data to Gumbel copula

	Parameter	Maximized loglikelihood
Estimates	7.88	13108

The robustness is assessed with the maximized loglikelihood, which is equal to 13108. However, due to our limited choice to that copula, this robustness parameter will not be compared to another potential model of dependence.

```
   Call: fitCopula(copula, data = data, method = "mpl")
Fit based on "maximum pseudo-likelihood" and 3106 5-dimensional observations.
Copula: gumbelCopula
alpha
7.88
The maximized loglikelihood is 13108
Optimization converged
```

Result for fitting the data with wind field and storm surge losses to the Gumbel copula

```
   Call: fitCopula(copula, data = data, method = "mpl")
Fit based on "maximum pseudo-likelihood" and 3105 5-dimensional observations.
Copula: gumbelCopula
param
6.655
The maximized loglikelihood is 11333
```

```
Optimization converged
```

TABLE 2.14: Values-at-Risk for different exceedance probabilities : 5%, 2.5%, 1% and 0.01% with 1000 random simulations

	95%	97.5%	99%	99.9%
Return period	20-year	40-year	100-year	1000-year
VaR	2,091.17	51,171.23	1,036,699	84,540,742

TABLE 2.15: Conditional Values-at-Risk for different exceedance probabilities : 5%, 2.5%, 1% and 0.01% with 1000 random simulations

	95%	97.5%	99%	99.9%
CVaR	1.2e07	2.4e07	5.9e07	3.5e08

Wind field effect

TABLE 2.16: Values-at-Risk with wind field and storm surge impacts

	95%	97.5%	99%	99.9%
Return period	20-year	40-year	100-year	1000-year
VaR	3,008,749	25,139,836	2.62e08	9.78e08

TABLE 2.17: Conditional Values-at-Risk with wind field and storm surge impacts

	95%	97.5%	99%	99.9%
CVaR	1.6e08	3.1e08	6.5e08	1e09

Bibliographie

[1] *Advancing Disaster Risk Financing & Insurance In The Pacific*. Rapp. tech. Washington DC : World Bank, fév. 2015.

[2] L. BERTINELLI, P. MOHAN et E. STROBL. “Hurricane Damage Risk Assessment in the Caribbean: An Analysis Using Synthetic Hurricane Events and Nightlight Imagery”. In : *Ecological Economics* 124 (2016), p. 135–144. URL : https://doi.org/10.1016/j.ecolecon.2016.02.004.

[3] E. BLANC et E. STROBL. “Assessing the Impact of Typhoons on Rice Production in the Philippines”. In : *Journal of Applied Meteorological and Climatology* 55.4 (2016), p. 993–1007. DOI : 10.1175/JAMC-D-15-0214.1.

[4] E. BOOSE, M. SERRANO et D. FOSTER. “Landscape and regional impacts of hurricanes in Puerto Rico”. In : *Ecological Monographs* 74 (2004), p. 335–352. DOI : 10.1890/02-4057.

[5] *Catastrophe Risk Assessment Methodology*. Rapp. tech. Washington DC : World Bank, 2013.

[6] *CCRIF SPC Annual Report 2016-2017*. Rapp. tech. CCRIF SPC, 2017.

[7] H. F. CHAN et G. O. WALKER. “Empirical Studies of the Peak Surge due to Tropical Cyclones at Hong Kong”. In : *Journal of the Oceanographical Society of Japan* 35 (1979), p. 110–117.

[8] DEPARTMENT OF HOMELAND SECURITY. *Multi-hazard Loss Estimation Methodology, Hurricane Model, Hazus-MH 2.1, Technical Manual*. Federal Emergency Management Agency. Washington, D.C, USA, URL : https://www.fema.gov/plan/%20prevent/hazus.

[9] C. DUTANG. *gumbel: package for Gumbel copula*. R package version 1.10-1. 2015.

[10] K. EMANUEL. "Global Warming Effects on U.S. Hurricane Damage". In : *American Meteorological Society* 3 (2011), p. 261–268. URL : https://doi.org/10.1175/WCAS-D-11-00007.1.

[11] K. EMANUEL, R. SUNDARARAJAN et J. WILLIAMS. "Hurricanes and Global Warming: Results from Downscaling IPCC AR4 Simulations". In : *American Meteorological Society* 3.89 (2008), p. 347–367. URL : https://doi.org/10.1175/BAMS-89-3-347.

[12] *Financial Protection of the State against Natural Disasters: a Primer*. Rapp. tech. Washington DC : World Bank, 2010.

[13] G. GUDENDORF et J. SEGERS. "Extreme-Value Copula". In : *Copula Theory and Its Applications. Lecture Notes in Statistics*. Springer, Berlin, Heidelberg, 2010, p. 127–145. URL : https://doi.org/10.1007/978-3-642-12465-5_6.

[14] M. HOFERT et al. *copula: Multivariate Dependence with Copulas*. R package version 0.999-17. 2017. URL : https://CRAN.R-project.org/package=copula.

[15] G. J. HOLLAND. "An Analytic Model of the Wind and Pressure Profiles in Hurricanes". In : *Mon. Wea. Rev.* 108 (1980), p. 1212–1218. DOI : 10.1175/1520-0493(1980)108,1212:AAMOTW.2.0.CO;2.

[16] P. HOUGAARD. "A Class of Multivariate Failure Time Distribution". In : *Biometrika* 73.3 (1986), p. 671–678. URL : http://www.jstor.org/stable/2336531.

[17] J. L. IRISH, D. T. RESIO et Ratcliff J. J. "The Influence of Storm Size on Hurricane Surge". In : *American Meteorological Society* 38 (2008), p. 2003–2013. DOI : 10.1175/2008JPO3727.1.

[18] P. MOHAN et E. STROBL. "A Hurricane Wind Risk and Loss Assessment of Caribbean Agriculture". In : *Environment and Development Economics* 22.1 (2017), p. 84–106. DOI : 10.1017/S1355770X16000176.

[19] B. M. PAULSEN et J. L. SCHROEDER. "An Examination of Tropical and Extratropical Gust Factors and the Associated Wind Speed Histograms". In : *Journal of Applied Meteorology* 44 (2005), p. 270–280. DOI : 10.1175/JAM2199.1.

[20] D. PRADHAN, A. MITRA et U. K. DE. “Estimation of pressure drop and storm surge height associated to tropical cyclone using Doppler velocity”. In : *Indian Journal of Radio & Space Physics* 41 (2012), p. 348–358.

[21] K. SEALY et E. STROBL. “A Hurricane Loss Risk Assessment of Coastal Properties in the Carribean: Evidence from the Bahamas”. In : *s* ().

[22] E. STROBL. “The Macroeconomic Impact of Natural Disasters in Developing Countries; Evidence from Hurricane Strikes in the Central American and Caribbean Region”. In : *Journal of Development Economics* 97.1 (2012), p. 130–141. DOI : 10.1016/j.jdeveco.2010.12.002.

[23] R Core TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL : http://www.R-project.org/.

[24] *The Sendai Report: Managing Disaster Risks for Resilient Future*. Rapp. tech. Wshington D.C., U.S.A. : GFDRR, 2012.

[25] P. J. VICKERY, F. J. MASTERS et M. D. POWELL. “Hurricane Hazard Modeling: The Past, Present, and Future”. In : *Journal of Wind Engineering and Industrial Aerodynamics* 97.7-8 (2009), p. 392–405. DOI : 10.1016/j.jweia.2009.05.005.

[26] Y.-F. XIAO, Y.-Q. XIAO et Z.-D. DUAN. “The Typhoon Wind Hazard Analysis in Hong Kong of China with the New Formula for Holland B parameter and the CE wind field model”. In : *Proc. Seventh Asia-Pacific Conf. on Wind Engineering* (2009). URL : http://www.iawe.org/Proceedings/7APCWE/M2B_4.pdf..

W :/agoburdhun/Documents/Dissertation/bibliopcrafi.bib

Chapitre 3

Statistical Emulator of Eight Crops Yields from Global Gridded Crop Models

co-written with Elodie Blanc

3.1 Introduction

Predicting crop yields accounting for climate change is a challenge that researchers and policymakers are taking up with more and more interest. An important number of studies on this topics (Roudier & al., 2011 ; Kang & al., 2009)[9, 4] is leading to rising concern about the distribution of positive and negative impacts of climate change on agriculture across the globe.

So far, two types of crop models can be found in the literature :

— process-based models, which are designed at the plant scale using a bench of biologic and physical parameters to predict the plant growth in a certain context. These models present the advantage of accounting for the physiology of the plant and provide in general very accurate results. However, they are often pri-

vately owned and their utilisation requires specific skills.

— statistics models, which are estimated using historical observation and making a link between the weather and crop yields with time series. While this type of model is easier to handle, it has the major inconvenient of being limited to observed data. Hence, predicting the effect of out-of-sample weather values on crop yields is inaccurate.

The purpose of our study is to provide a crop emulator that benefits from the easiness to handle and accessibility of statistics models and the accuracy of process-based models. Following Blanc & Sultan (2015)[2] and Blanc (2017)[1], we run a statistical model on the projections made by the process-based model available for the crops studied, the Global Gridded Crop Model (GGCM) LPJmL. On the one hand, by generating our data with a model and not using historical observations, we take into account future changes with extreme effects of weather on crop yields not observed so far. On the other hand, since our model provides accurate results with a limited number of features and without requiring any restricted access to a data base, running the model is very facilitated and accessible.

Facilitating the availability to accurate data on crop yields predictions is of particular importance in developing countries where the economy is strongly relying on agriculture and where financial means are limited. Furthermore, predicting crop yields at this level of accuracy is of a crucial interest for smallholders who are more risk adverse (Yesuf & Bluffstone, 2007)[13]. Here we are focusing on subsistence crops in developing countries, hence paying attention to future yields and anticipate means face climate change impacts is in the core of the problem.

We take a couple of studies on statistical models predicting crop yields in the future as examples (Schlenker & Roberts, 2009 ; Lobell & Burke, 2010)[10, 5] to measure the goodness-of-fit of our results, such as RMSE, NRMSE and the log-ratio between the crop emulator and the crop model. We also take into account the non linear effect of weather variables in our estimation (Schlenker & Roberts, 2009)[10] and the potential particular effect of extreme weather values (Moriondo & al, 2011)[6]. Indeed, crop yield response might increase or decrease exponentially for extreme weather values that have not been observed so far.

Our results show that the crop emulator statistically designed provides accurate results very close to the GGCM. Following Blanc (2017)[1], we keep only the monthly mean temperature and precipitation, and the mid-year CO_2 concentration as explanatory variables. Using a fractional polynomial model to take into account the non-linear effect of weather variables, we fit very accurately our model, with a general average tendency to underestimate the effects of climate change assessed by the GGCM.

3.2 Context and data

3.2.1 Context

Accounting for more than a quarter of the global total production and 36% of the production in Africa in 2016 (FAOSTAT)[8], the eight groups of crops we are studying here play a major role in agriculture production at the global scale. They also represent a key role in food production in developing countries, after the crops studied in the previous paper of Blanc(2017)[1], maize, sorghum, wheat and rice. The importance of agricultural production in developing countries is mostly due to their higher vulnerability and exposure to the effects of climate change : beyond the natural climatic conditions (desert, mangrove, fragile ecosystems, etc.) that are more sensitive to climate changes, poor countries won't be able to invest within durable infrastructure to cope with the consequences of climate change.

Predicting yields is of particular importance to be able to anticipate crisis and implement policies that could redistribute food crops where climate change has opposite impacts. But even if models exist so as to predict accurately yields at a very localised scale, making it available to the most is of the essence. Indeed, assuming that crop models are the most precise model existing, the lack of accessibility and easiness to handle might be a serious obstacle to knowledge dissemination, including to countries with poor technology resources. By fitting a statistical model to the predictions of the GGCM with fewer variables and easy to handle model, we intend to provide the accuracy of the crop models to the most.

TABLE 3.1: Crop production (in tons) in the world in 2016 (source : FAO)

Crop	Production (t)	Percentage of total production
Cassava	281,896,830	2.63%
Groundnuts/nuts	61,755,728	0.58%
Millet	30,353,830	0.28%
Pulses	4,093,581	0.04%
Rapeseed	84,137,080	0.79%
Sugar beet	285,326,548	2.64%
Sugar cane	2,013,721,491	18.80%
Sunflower	49,932,460	0.47%
Total	2,811,217,548	26.24%

3.2.2 Climate models and data

Data are used are the grid cell level at the $0.5 \times 0.5^{\circ}$ resolution [1].

Daily weather data are used to make our estimations and predictions. To this end, we use two climate models from the CMIP5 climate models, or General Circulation Models (GCMs) : HadGEM2-ES (designed by the Met Office Hadley Centre and Instituo Nacional de Pesquisas Espaciais)) and GFDL-ESM2 M (provided by the NOAA Geophysical Fluid Dynamics Laboratory). These two models represent respectively high and low levels of global warming (Warszawski & al., 2007)[12]. Both models provide two data periods from 1975 to 2099 : the 'historical' data from 1975 to 2005 and 'future' data from 2006 to 2099 and for different climate scenarios. We base our estimations on the RCP8.5, which is the Representation Concentration Pathway (RCP) with the highest level of global warming. We include the concentration of CO_2 from this data base.

These climate models are used first as an input for the GGCM, mainly using the daily precipitation, minimum and maximum temperatures so as to obtain the yields predictions according to the GGCM. These climate data are used a second time, with this time the monthly average of precipitation (Pr), temperature ($Tmean$) and the annual

1. Equal to a 30 arc-minute resolution

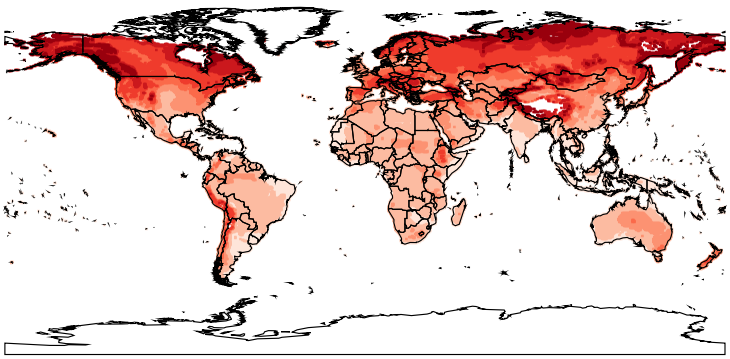
mid-year CO_2 concentration, in our regressions to make a link between these variables on crop yields generated by the GGCM.

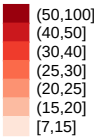
TABLE 3.2: Statistic summary information for the weather data by GCM

		GFDL				Had GEM 2			
Crop	Unit	Mean	Variance	Min	Max	Mean	Variance	Min	Max
Pr_1	mm/day	2.89	3.82	0	147.08	2.81	3.62	0	152.08
Pr_2	mm/day	3.23	4.14	0	175.98	3.25	4.24	0	174.54
Pr_3	mm/day	3.26	4.10	0	127.33	3.25	4.06	0	174.54
Tmean_1	°C	20.20	9.78	-6.48	45.09	21.56	9.51	-7.19	46.82
Tmean_2	°C	22.05	8.49	-3.57	45.25	23.36	8.42	-3.51	47.52
Tmean_3	°C	21.28	8.92	-6.11	45.89	22.59	8.81	-7.34	46.68
CO_2	ppm	527.48	176.11	325.86	926.67	527.35	176.05	325.86	926.67

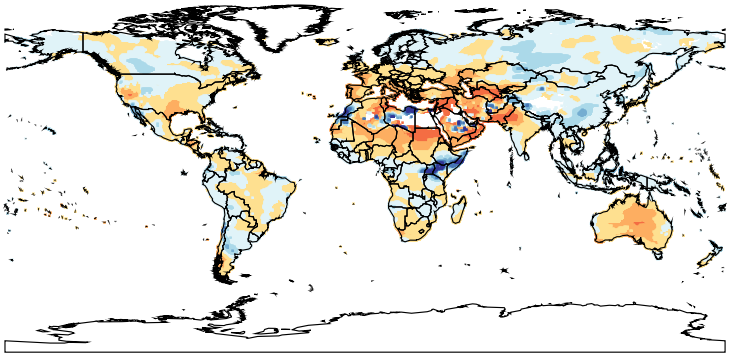
FIGURE 3.1: Change in temperature, precipitations and CO_2 concentration between years 2001 and 2100 with model GFDL

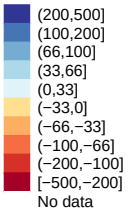
(a) Temperature



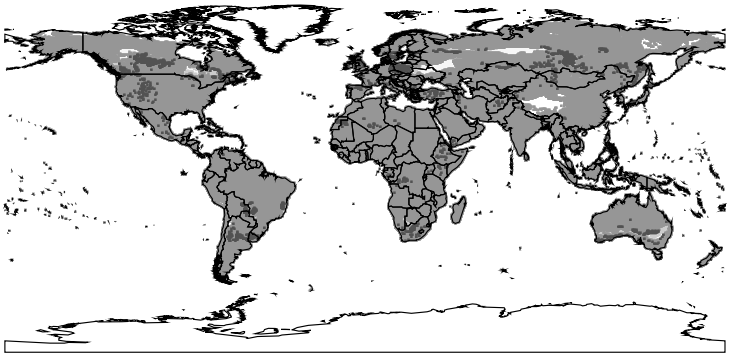


(b) Precipitation





(c) CO_2



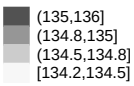
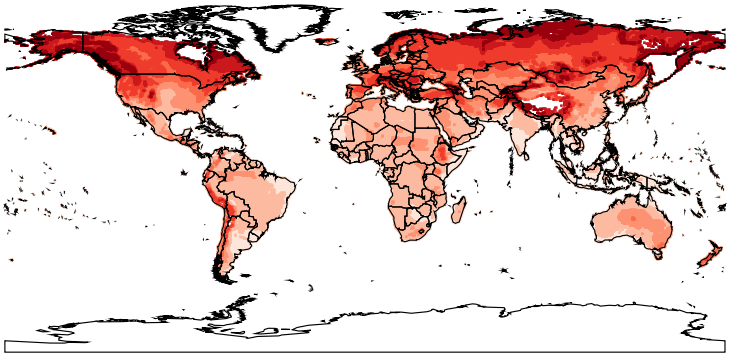
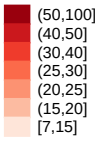


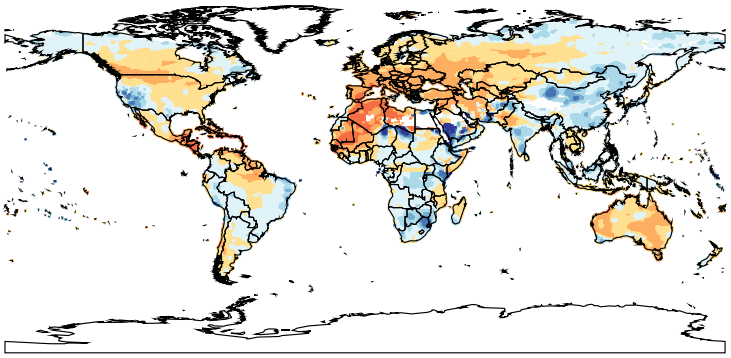
FIGURE 3.2: Change in temperature, precipitations and CO_2 concentration between years 2001 and 2100 with model HadGEM 2

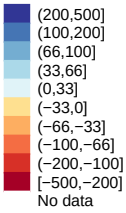
(a) Temperature



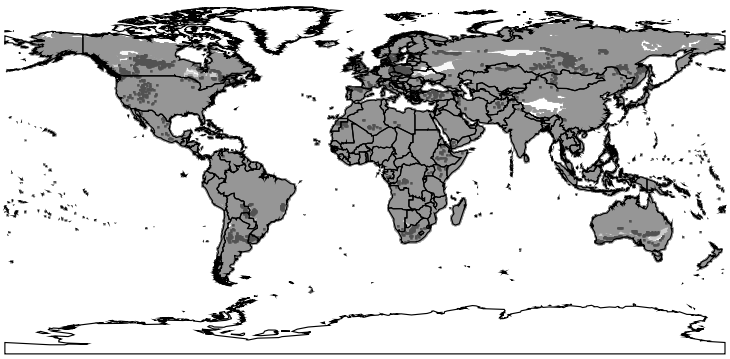


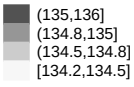
(b) Precipitation





(c) CO_2





As shown on figure 3.1 and figure 3.2, temperature and CO_2 variations in the two climate models are very similar and the main difference between the two models lie on rainfall predictions. Because plant growth is strongly sensitive to water availability, including irrigation in our model would impact very significantly the magnitude of yields changes. However, irrigation remains very uncertain in the future because we don't know the extent to which governments will invest and foster it and how much technology progress and availability will be until 2100. Hence, accounting for irrigation would imply strong assumptions or omissions that will shift sizably the results. Since we are dealing with long term effects with a lot of uncertainty on this parameter, we don't consider irrigation in our study.

3.2.3 Soil classification

Combining simultaneously solid, liquid and gaseous states of different elements, soils have an extremely broad range of characteristics across the world. Twelve soil orders have been defined by the FAO-UNESCO so as to group soils with common characteristics : since significant differences in climate change responses have been observed (Blanc and Sultan, 2015), we use these data for our regressions, at the grid cell level. Indeed, crop yields will have significantly different responses to climate change due to the different soil composition and reactions to changes in the atmosphere.

Hence we distribute our data among 12 orders of soil taxonomy of FAO-UNESCO[11].

3.2.4 Crop model and data

Data base We carry on this statistical estimation for eight crops : cassava, millet, sunflower, sugar cane, sugar beet, rape seed/canola, groundnuts/peanuts and pulses. Data for these crops are available in one of the Global Gridded Crop Models (GGCMs) as part of the ISI-MIP Fast Track. In this study, we will therefore use the predictions provided by the Lund Potsdam-Jena managed Land (LPJmL) dynamic global vegetation and water balance model. Like other GGCM, this model simulates processes such as photosynthesis, plant growth, maintenance and regeneration losses, soil moisture, runoff, evapotranspiration and vegetation structure [3].

In the GGCM, as well as in the statistical emulator that will be generated from our estimations, we have a data set of crop yields across the globe through 125 years. This data set represents the crop yields estimated in the 'historical' and 'future' periods for a certain number of grid cells (presented in the next paragraph) if the crop into question was harvested in this grid cell. In the end, for a large majority of grid cells on the globe, we have data and run our estimation as if the crop considered was harvested in these area. This present the advantage of having an overview, for a given crop, of how yields evolve in the future across the globe and see the differences among the different regions. In our results, we present the evolution of both the crop yields if it was harvested everywhere in the world and the average of crop yields weighted by area harvested. Using the MIRCA2000 dataset (Portmann & al., 2010)[7], we have data for each grid cell, we can therefore weight our result by area harvested at the grid cell level.

Also, an interesting feature in this GGCM is that it takes into account soil characteristics from the Harmonized Solid Database, hence we will be able to separate our estimations depending on the soil type of each grid cell. With more than 15,000 soil mapping units combining national and regional soil profiles with the FAO-UNESCO Soil Map of the World, the LPJmL model provides an extensive integration of soil interaction with plant growth.

Statistical description of the data Table 3.3 and 3.4 describe the data from the GGCM (LPJmL). Table 3.3 with the number of observations recorded and the number of grid cells. For each grid cell and for each crop, we have several observations since we are running our model on several years and with several climate models. We have the most extended data across the globe for sugar beet and pulses with more than 64,000 grid cells and the most observations for sugar beet, rape seed/canola and pulses, with more than 16,000,000 observations.

Table 3.4 shows the summary of crop yields obtained with the GGCM over the data set for both models GFDL and Had GEM 2. In general, the average yields are higher for the GCM Had GEM 2, where for example the average yields of sugar cane are equal to 5.48 t/ha while it is equal to 5.01 t/ha with the GFDL model.

TABLE 3.3: Statistic summary information on the crop data from the LPJmL model with the two climate models

Crop	Observations	Grid cells
Cassava	11,383,804	50,876
Millet	15,042,276	58,953
Sunflower	15,726,235	62,040
Sugar cane	14,665,355	58,932
Sugar beet	16,451,553	64,517
Rape seed/canola	16,593,806	64,930
Groundnuts/peanuts	13,858,337	54,279
Pulses	16,404,485	64,488

TABLE 3.4: Statistic summary information for the crop yields (t/ha) by GCM

	GFDL				Had GEM 2			
Crop	Mean	Variance	Min	Max	Mean	Variance	Min	Max
Cassava	1.03	2.24	0	35	1.48	2.36	0	34.82
Millet	0.39	0.56	0	8.05	0.44	0.57	0	8.05
Sunflower	0.73	1.01	0	11.35	0.80	0.97	0	13.09
Sugar cane	5.01	7.78	0	35	5.48	8.22	0	35
Sugar beet	2.68	3.82	0	35	2.85	3.79	0	35
Rape seed/canola	1.08	1.13	0	22.04	1.09	1.09	0	23.19
Groundnuts/peanuts	0.61	0.93	0	20.61	0.65	0.92	0	24.25
Pulses	1.01	1.28	0	32.56	1.08	1.25	0	31.86

3.3 Method

For each crop considered, and since we have only one GGCM, we run our estimations with all the data set, from 'historical' to 'future' data, and with both GCM described before. We use the GGCM (LPJmL) to run our regression and obtain estimates for each variable considered. Once our estimation is done, we can project the yields for each crop with the statistical crop emulator so as to compare the projections from our model and the GGCM. Finally, we will have in our statistical projection, a different estimation for the 8 crops studied (cassava, millet, sunflower, sugar cane, sugar beet, rape seed/canola, groundnuts/peanuts and pulses). The yields projected will be available for every grid cell where the GGCM provides estimation, regardless of whether the crop is actually grown in the grid cell considered. To account for the area where the crop is grown, we provide afterwards results by taking the weighted average on area harvested using the MIRCA2000 dataset (Portmann & al., 2010)[7].

Following the method presented by Blanc & Sultan (2015)[2] and Blanc (2016)[1], we favor the estimation that takes into account soil orders instead of running one estimation for the globe. Therefore, we run for each crop a different estimation for each soil. While the study by Blanc & Sultan (2015)[2] makes an estimation with an extensive number of variables (such as the number of days of rain, etc), we chose a parsimonious set of weather variables with the mean monthly temperature and the mean monthly precipitation during the three growing months, and the annual midyear CO_2 concentration. Following these studies, and so as to be able to make a uniform data base of crop yields projections for all the crops, we keep the same polynomial model (that we will call *S1fpint*) that takes a fifth order polynomial specification for each meteorological variable. The fifth order polynomial specification is described Appendix at section 3.8.1.

$$
\begin{aligned}
Yields_{lat,lon,gcm,y} = \alpha &+ \sum_{i=1}^{3} \beta_i Pr\_i_{lat,lon,gcm,y} + \sum_{i=1}^{3} \theta_i Tmean\_i_{lat,lon,gcm,y} \\
&+ \nu CO2_{gcm,y} + \sum_{i=1}^{3} \gamma_i Pr\_i_{lat,lon,gcm,y} * Tmean\_i_{lat,lon,gcm,y} \\
&+ \sum_{i=1}^{3} \omega_i Pr\_i_{lat,lon,gcm,y} * CO2_{gcm,y} \\
&+ \sum_{i=1}^{3} \kappa Tmean\_i_{lat,lon,gcm,y} * CO2_{gcm,y} \\
&+ \delta_{lat,lon} + \epsilon_{lat,lon,gcm,y}
\end{aligned}
\tag{3.1}
$$

where for each year y, $Yields$ corresponds to crop yields simulated by process-based crop models for each grid cell (defined by its longitude lon and latitude lat) under each climate model, gcm ; Pr and $Tmean$ variables correspond to monthly mean precipitation and temperature variables. The indices $\_i$ correspond to the month of growing season (composed of three months depending on the hemisphere). CO_2 is the annual midyear CO_2 concentration level in the atmosphere ; δ is a grid cell fixed effect ; and ϵ an error term.

As mentioned before, a polynomial specification is made for each one of the three weather variables considered. Indeed, as shown in the previous studies (Blanc & Sultan, 2015 ; Blanc, 2016)[2, 1], the weather variables studied here have a non-linear effect and including a quadratic effect allows a better fitting to the data. Indeed, they show that they have a better fit to the crop model when employing a quadratic function to model crop yields response to each weather variable. The quadratic term in its simplest form presents the advantage of allowing non-linear effects of weather on yields but is constrained to symmetric effects. To allow for greater flexibility and asymmetry in the response of yields to weather variation, we use a fifth order polynomial specification ($S1fpint$). However, since this specification only doesn't take into account the effect of extreme weather events, we relax the symmetry constraint and allow a non-parametric flexibility by using a fractional polynomial specification.

As mentioned in the data description, the GGCM used to make our estimations accounts for soil features, so we add another specification to our model by running a

different estimation per soil type (and per crop). Since we run a new estimation for each soil type, we necessary have to reduce the number of soil types compared to the data available in the GGCM. Therefore we simplify this interaction by considering the 12 general soil orders of the USDA soil taxonomy [11]. Since the soil type is time invariant, its effects will be captured in the grid cell fixed effect δ, enabling us to isolate the effect of weather variation on crop yields. In addition to this fixed effect that we isolate, we run one estimation per soil type since the weather variation may have significant different impacts on crop yields depending on the soil order.

Our model runs the estimation taking into account the 3 months most important in the growing season, so we also define the growing season period for each crop. Figure 3.3 to figure 3.10 show, for each hemisphere and for each crop, the planting and maturity frequency over months. We observe a trend in most of the crops : in the North the growing season is in June, July and August and in the South, the growing season is in December, January and February. For sugar cane, we observe an overlap between the planting and the maturity seasons. This is mainly explained by the fact that sugar cane is not sowed every year, so the planting season is not always linked to the maturity season. Cassava also has a different trend in the North : it has two growing seasons at different moments : a smaller growing season in April and a second one in October and November. We take this difference into account in our estimations.

FIGURE 3.3: Cassava seasonality

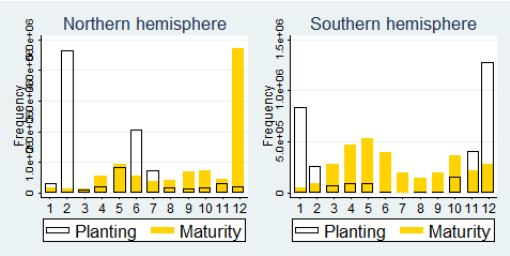


FIGURE 3.4: Millet seasonality

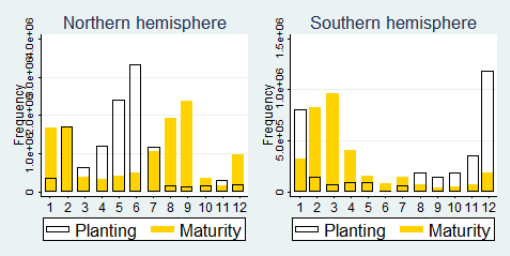


FIGURE 3.5: Sunflower seasonality

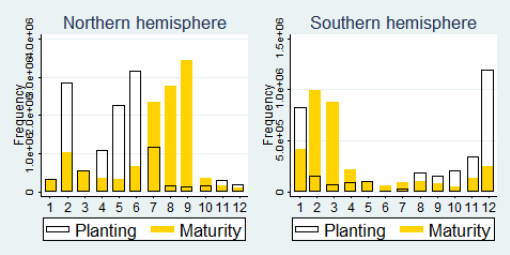


FIGURE 3.6: Sugar cane seasonality

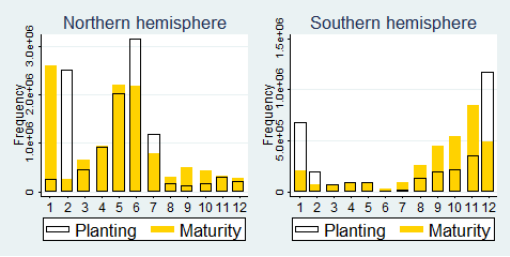


FIGURE 3.7: Sugar beet seasonality

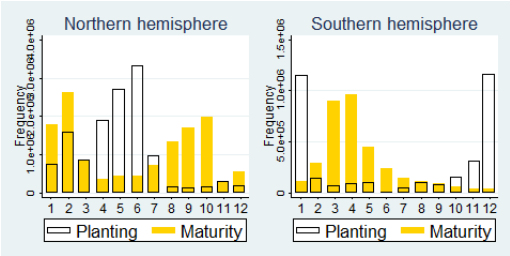


FIGURE 3.8: Rape seed/canola seasonality

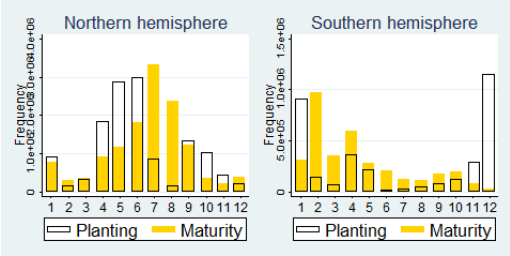


FIGURE 3.9: Groundnuts/peanuts seasonality

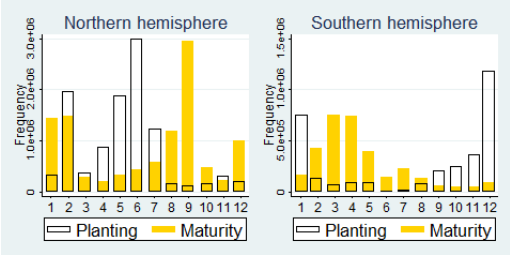
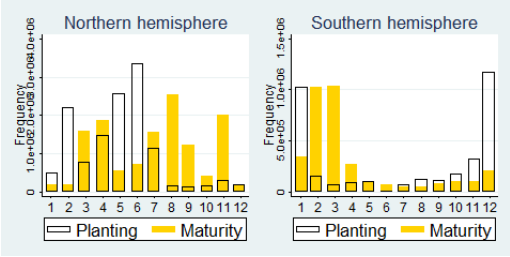


FIGURE 3.10: Pulses seasonality



3.4 Results

3.4.1 Global average evolution of yields

FIGURE 3.11: Comparison between GGCM and emulator of yields for cassava, millet, groundnuts, pulses, rape seeds and sunflowers

Averaged for the world

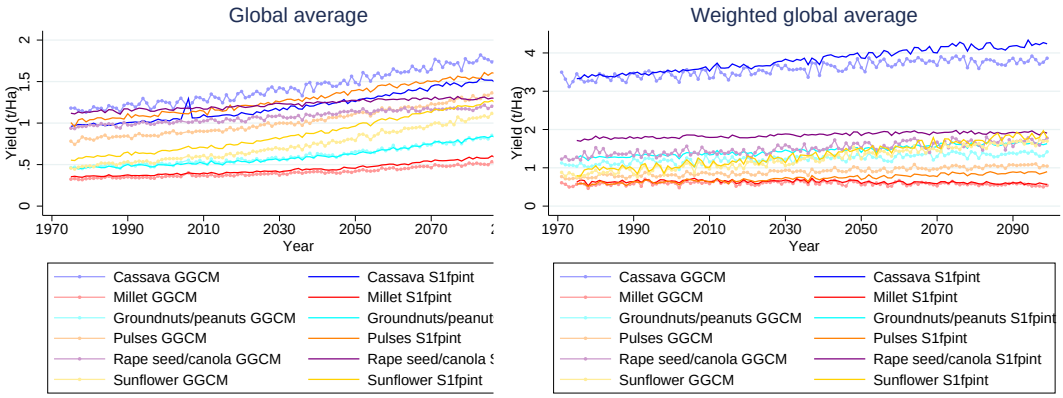


FIGURE 3.12: Comparison between GGCM and emulator of yields for sugar beet and sugar cane

Averaged for the world

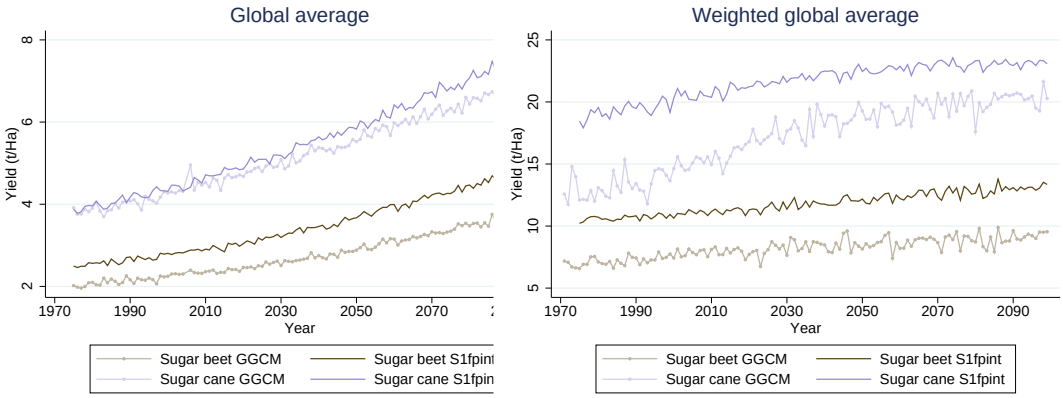


Figure **??** shows the prediction of the global average of yields for all crops in the next century, with both the statistical emulator and the GGCM. Figure **??** is the same

representation weighted by area harvested using the data base MIRCA 2000 [7]. This data base provides the area harvested and the total area of each grid cell, allowing to compute this weighted average. The global average without weighting by area harvested gives a good overview of how crop yields evolve in general, without accounting for the most harvested crops. Hence, we can infer from figure **??** that crop yields increase on average across the globe. On both graphs, we see that sugar cane and sugar beet are the crops that have higher yields compared to the others. For average yields regardless of the area harvested, we observe a general trend of increasing yields for all the crops studied. That trend is strengthened when averaging with the area harvested : keeping sugar cane apart, all crops have on average much higher yields and increasing yields at an even higher rate than the world average. This results show that the distribution of crops harvested areas is relatively efficient since we have on average better yields and better increase in the future. Sugar cane is the exception since it shows a decreasing yield over the next century when averaging crop yields by the area harvested. However, with yields between 20 and 30 t/ha over the 21st century, it still clearly has the highest yields among the 8 crops studied.

3.4.2 Change in yields at the grid cell level and comparison between the statistical crop emulator and GGCM

Figures **??** to **??** show, for each crop, the change in yields at the grid cell level expressed as a percentage of the difference between the years 2010s and 2090s. The results provided here correspond to the average between both GCM. Orange colors correspond to a decrease in yields and blue colors represent an increase. Yellow color corresponds to almost no change. White areas can correspond either to no data or to non significant yields, ie. equal to less than 1 t/ha in the GGCM. We have also removed areas where the crop emulator and the GGCM provided projections of opposite signs. Changes are displayed both for the statistical emulator (S1fpint) and the GGCM (LPJmL) with the same legend. Here, we don't account for the area harvested.

To see precisely where the statistical emulator over- or underestimate the strength of the impacts of climate change on yields compared to the GGCM, we compute the log ratio of yields of the statistical emulator over the GGCM. Using the log-ratio rather

than a simple ratio presents the advantage of highlighting places where the climate change effects are unerestimated or overestimated. For example, if yields increase by 20% with the GGCM against 10% with the crop emulator, the ratio would be equal to 0.5. On the opposite, if the GGCM predicts yields increase by 10% against 20% with the crop emulator, the ratio equals 2. In these symmetric cases, the log-ratio would be equal to, respectively, -0.3% and 0.3%. Green colors correspond to an underestimation of the statistical emulator and brown colors correspond to an underestimation. If yields are increasing, an overestimation of the extend of climate change impact corresponds to higher yields in the statistical emulator than in the GGCM. On the opposite, in cells where yields are decreasing, the overestimation of the statistical emulator corresponds to lower yields than in the GGCM. We consider that a log ratio higher than 1 or lower than -1 corresponds to a notable difference between the crop emulator and the GGCM.

Cassava Significant yields are observed for cassava mostly in tropical regions : South America, Southern Africa and South Eastern Asia. Yields are expected to increase in most of these regions, except in some parts of the Amazonian forest and of Central Africa. Under- and overestimation made by the statistical emulator are evenly distributed across the globe. The average log-ratio is pretty low and equal to -0.64 with a variance of 2.14, which means that the statistical emulator estimates relatively well the effects of climate change compared to the GGCM.

Millet Millet has a lot of blank observations since its yields are very low across the globe. Indeed, table 3.4 on the data from the GGCM show that millet has the lowest mean among the 8 crops with a mean equal to 0.39t/ha with the GFDL ESM2 M model and to 0.44t/ha with the Had GEM 2 model. We observe significant data for the South-East of Asia, the East of North America and some parts of Eurasia. Yields are increasing in general, except on the west coast of America and the South East of Asia where yields are decreasing. For both increasing and decreasing yields, the crop emulator tends to underestimate the impacts of climate change compared to the GGCM. On average, the logarithm ratio is equal to -1.47.

Sunflowers Significant data are observed almost everywhere in the globe, with no significant or no data in Central Asia, Australia and the Sahara. Yields are predominantly

increasing in every place where data is available. Most of the grid cells considered have predictions were yields are supposed to increase by more than 50% over the next century. With an average of the logarithm ratio of -0.14, the statistical emulator estimate almost equally the impacts of climate change as the GGCM.

Sugar cane Like for sunflowers, we have non significant data observed in the Sahara but a good coverage of data across the globe. Yields are also mostly increasing on the available data. We observe decreasing yields on the Amazonian forest, in a part of the Sub-Saharan Africa and in the South East of Asia. In these regions, the statistical emulator tend to overestimate the decreasing yields due to climate change compared to the GGCM. Concerning the increasing yield regions, the statistical emulator under- or overestimated the effects predicted by the GGCM evenly across the globe. With a standard deviation of the logarithm ratio of 2.15, it has the highest variance in term of difference between the statistical emulator and the GGCM among all the crops. The average logarithm ratio is equal to -1.13.

Sugar beet For sugar beet, we have a very good coverage of significant and non null data in the globe. Again, we observe mostly increasing yields at the global scale and decreasing yields in the Amazonian forest and in part of the Sub-Saharan Africa, and here in Australia. The crop emulator strongly overestimates the effects of climate change in Africa and in Russia. On the American continent, it tends to underestimate the effects compared to the GGCM. With a mean of -0.14 for the log-ratio, the statistical emulator makes on average very similar predictions to the GGCM.

Rape seeds / canola Here we observe significant data in most of the American continent, Europe and in a part of the South East of Asia. We have no significant data in all Africa. Climate change has different impacts across the America : it leads to decreasing yields over the Amazonian forest and in the eastern part of the North of America. In the remaining areas of America, we observe increasing yields. In Europe, rape seeds and canola yields tend to decrease, while they are overall increasing in the parts of the South East of Asia where we have observations. Compared to the GGCM predictions, the statistical emulator underestimate the impacts of climate change for both decreasing and increasing yield cells. On average, the global log-ratio is equal to -0.73.

Groundnuts / peanuts Here we observe significant data distributed across the globe. Most of the data are gathered in the South of America, the East of Northern America, Europe and part of the South East of Asia. We also observe significant points along the coast of Africa. Except for the Amazonian forest where yields are decreasing, yields are increasing for most of the observations. The strength of climate change tends to be overestimated by the statistical emulator in the South of Asia and underestimated in the remaining observations. With a log-ration averaging at -1.11, the statistical emulator makes a strong underestimation compared to most of the other crops studied.

Pulses All the significant observations are gathered in America, Europe and most parts of Asia. In most of the observations, pulses yields are increasing. In general, the crop emulator underestimates the effects of climate change in America (in both the North and the South) and overestimate it in Asia and Europe. With a mean of the log-ratio of -0.20 and a standard deviation equal to 1.39, the statistical emulator makes predictions for pulses that are in general pretty close to the GGCM.

FIGURE 3.13: Change in yields between years 2001 and 2100 for each crop with S1fpint and GGCM models

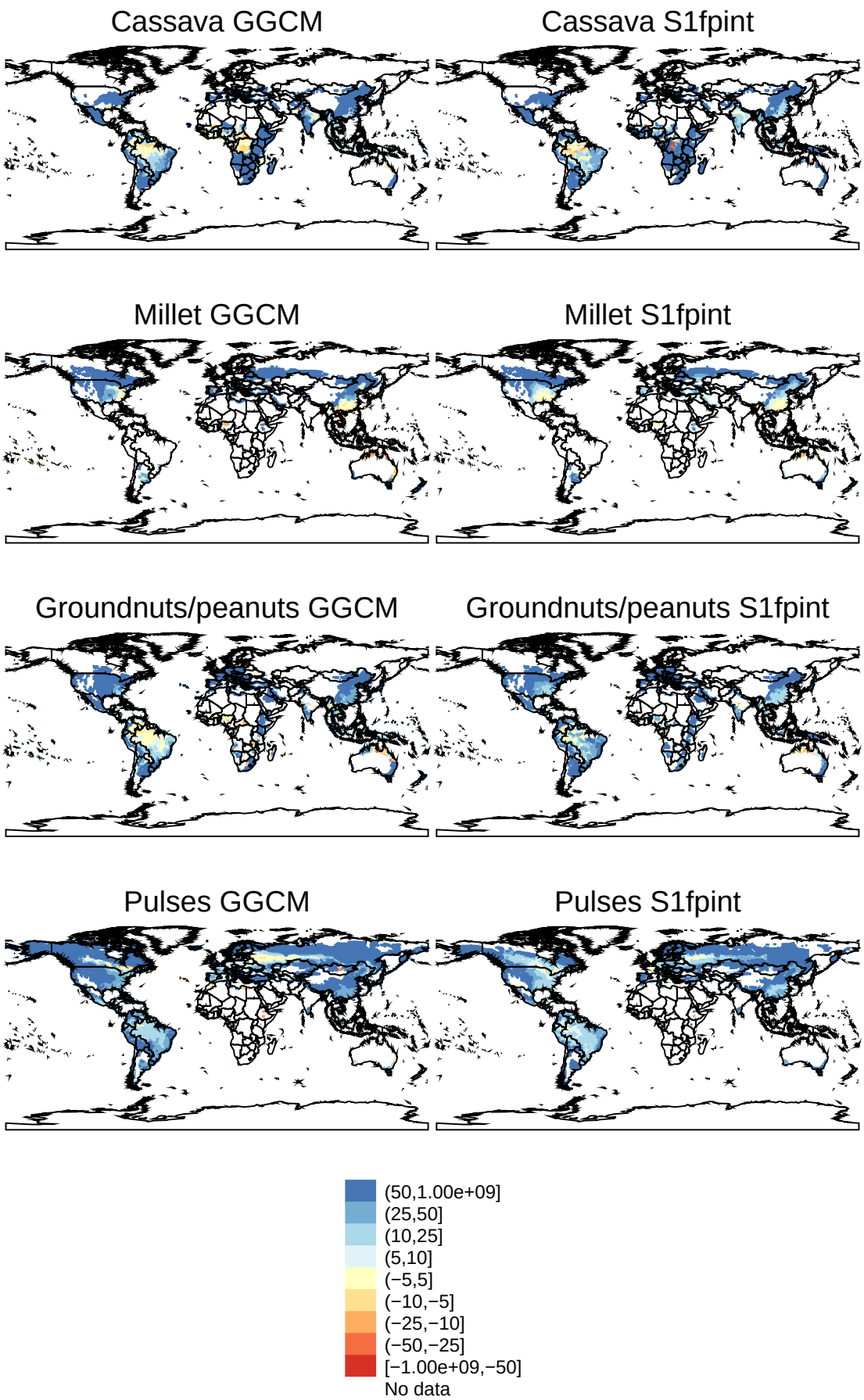


FIGURE 3.14: Change in yields between years 2001 and 2100 for each crop with S1fpint and GGCM models

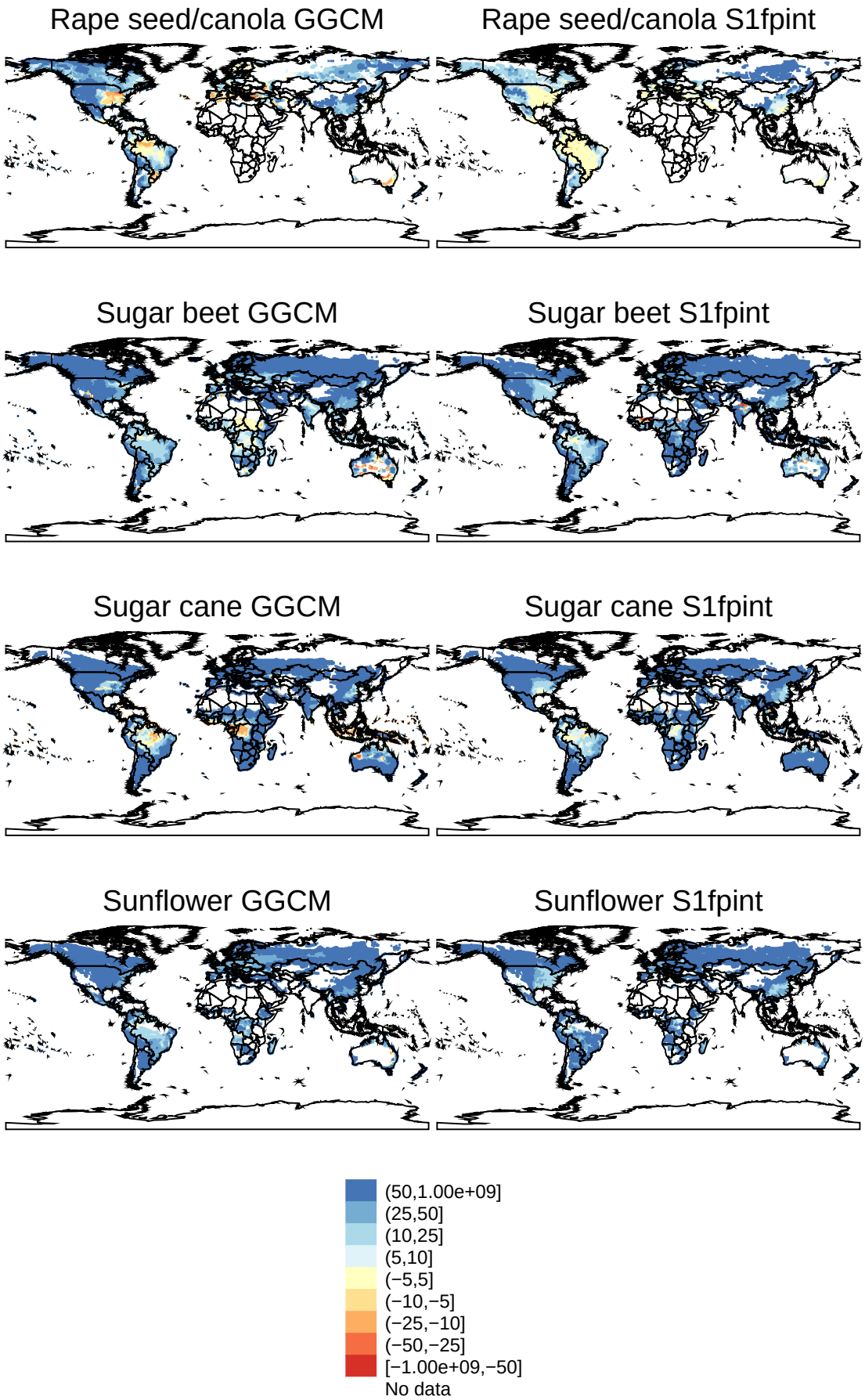


FIGURE 3.15: Comparison (with log ratio) between the statistical emulator and the crop model

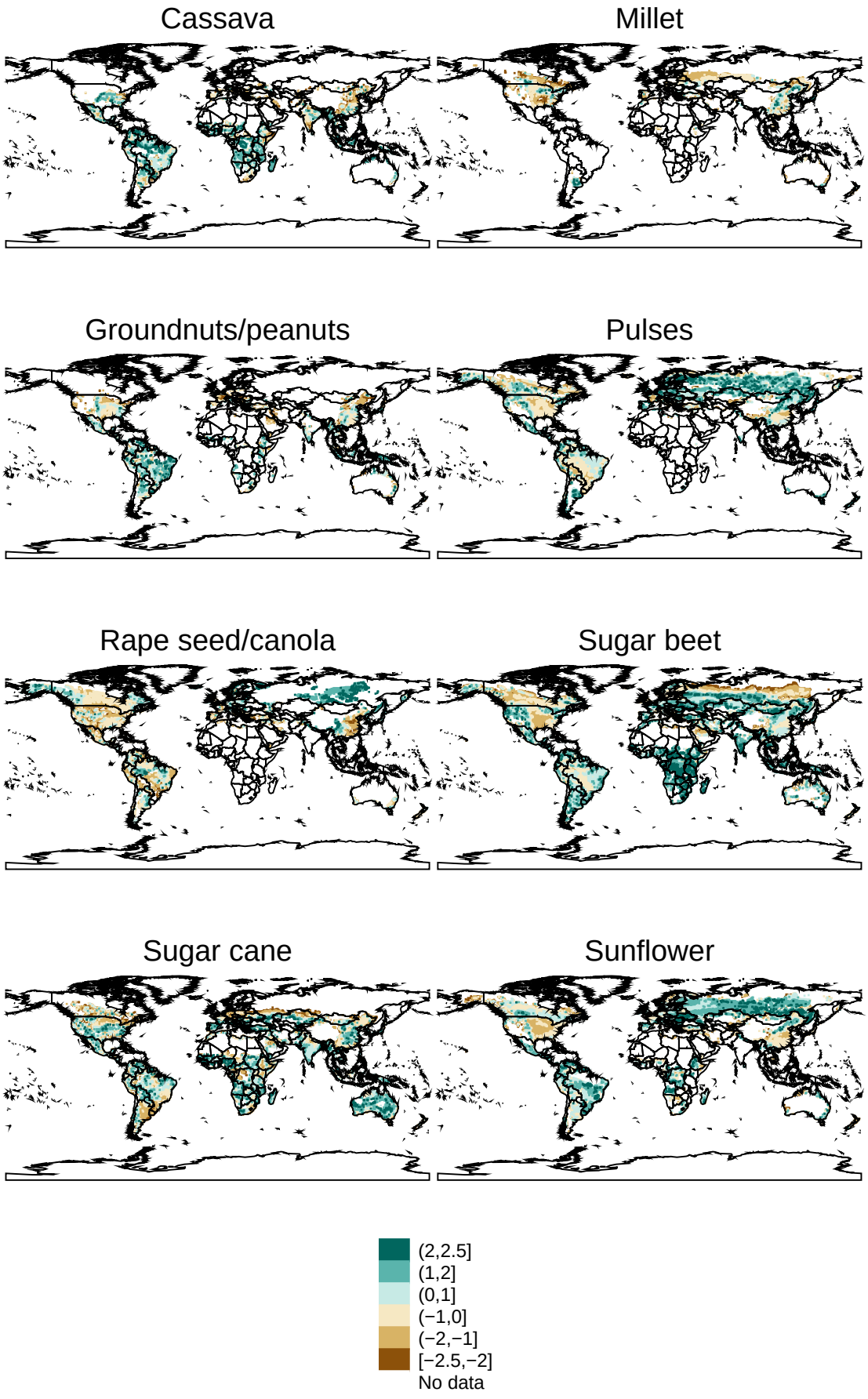


TABLE 3.5: Logarithm ratio of crop yields as predicted by the statistical crop emulator over the GGCM

Crop	Mean	Significance from zero	Variance	Minimum value	Maximum value
Cassava	-0.64	***	2.14	-10.62	14.87
Millet	-1.43	***	1.39	-8.88	9.45
Sunflower	-0.14	***	1.58	-8.22	9.76
Sugar cane	-1.13	***	2.15	-13.63	10.18
Sugar beet	-0.14	***	1.68	-9.13	14.92
Rape seed/canola	-0.73	***	1.73	-10.65	9.47
Groundnuts/peanuts	-1.11	***	2.05	-9.05	8.38
Pulses	-0.20	***	1.39	-8.13	8.61

3.5 Goodness-of-Fit and interpretation

3.5.1 RMSE

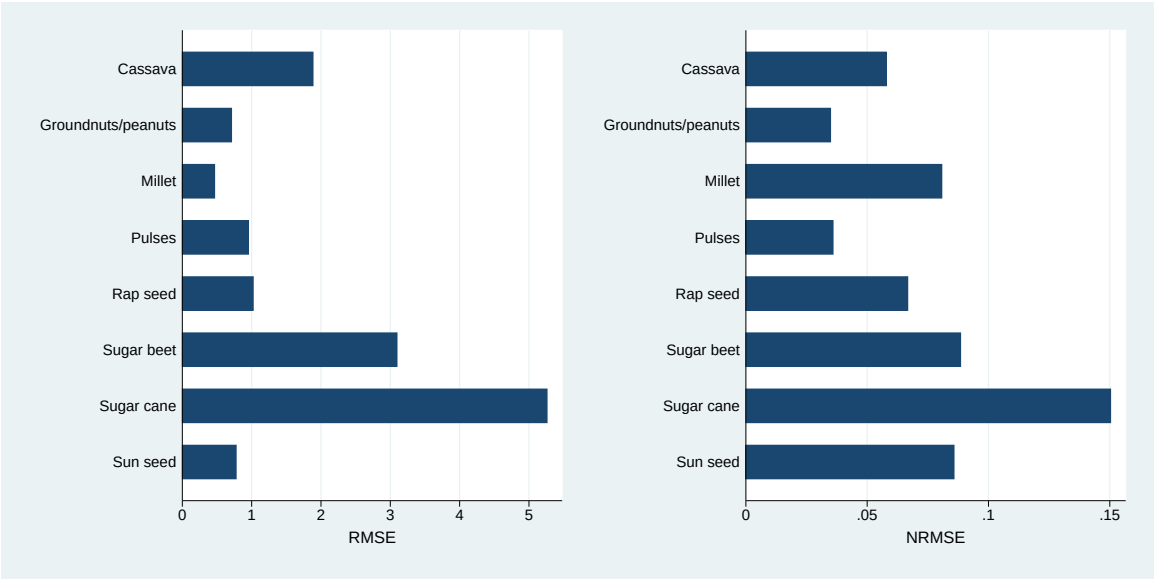
We use the root mean square error (RMSE) to measure the error margin between our estimations and the model. But since the samples scales might differ depending on the climate model, we also employ the normalized root mean square error (NRMSE), which normalizes the RMSE by the range of measured yields. The formulas are described Appendix at section 3.8.2.

3.5.2 Linear correlation

We compute the linear correlation between the values generated by our estimations and the GGCM. A positive correlation close to one means that our estimation reproduces the variations observed in the GGCM. To this end, we calculate the Pearson's correlation coefficient for each crop and for each climate model for the years 2030s, 2050s, 2070s and 2090s.

Table 3.6 shows the results obtained for the coefficients. In most cases, the cor-

FIGURE 3.16: RMSE an NRMSE by crop averaged on the world



relation coefficient is significantly different from 0 and decreasing as far as we go in the future. On average, the correlation coefficients are equal to 0.81 for the 2030s and equal to 0.55 for the 2090s. This can be mainly explained by the fact that there is more uncertainty and noise in the future, so linear correlation is necessary affected. However, despite that decreasing correlation in the future, we still have a significant and positive linear correlation higher than 0.5 for a large majority of the coefficients, which shows a conform prediction of the GGCM by the statistical emulator.

3.5.3 Out-of-sample validation

We use in our estimations two specific GCM (the GFDL ESM2M and Had GEM2 ES) with the scenario RCP8.5, inferring specific weather values. However, we must ensure that the model can still provide accurate projections in the future for out-of-sample weather data. To this end, we re-estimate the yields with exactly the same conditions, except that we exclude data with one of the two GCMs. Once this estimation is done, we use the weather data of the excluded model to project yields in the future and can compare it to the GGCM.

We then compute the RMSE and the NRMSE by comparing the results obtained

TABLE 3.6: Linear correlation between the statistical crop emulator and the GGCM for different segments of time in the future

Crop	2030s	2050s	2070s	2090s
Cassava	0.55^*	0.62^*	0.68^*	0.55^*
Millet	0.93^*	0.91^*	0.91^*	0.52^*
Sunflower	0.83^*	0.90^*	0.77^*	0.45
Sugar cane	0.70^*	0.70^*	0.43	0.71^*
Sugar beet	0.77^*	0.95^*	0.89^*	0.65^*
Rape seed/canola	0.92^*	0.93^*	0.40	0.37
Groundnuts/peanuts	0.91^*	0.79^*	0.81^*	0.69^*
Pulses	0.84^*	0.94^*	0.85^*	0.49

with the GGCM and both GCMs and with the model estimated excluding that GCM. Figure 3.17 summarizes the results obtained. In blue is the statistics computed for the leave-one-GCM-out sample and in blue is the statistics for the overall sample. The GCM displayed is the omitted GCM. There is a clear decline in both RMSE and NRMSE values, for each GCM and for each crop. The greatest difference observed between the restricted sample and the overall sample is for sugar cane for Had GEM2 ES, where the RMSE is increased by 1 and the NRMSE by 0.3.

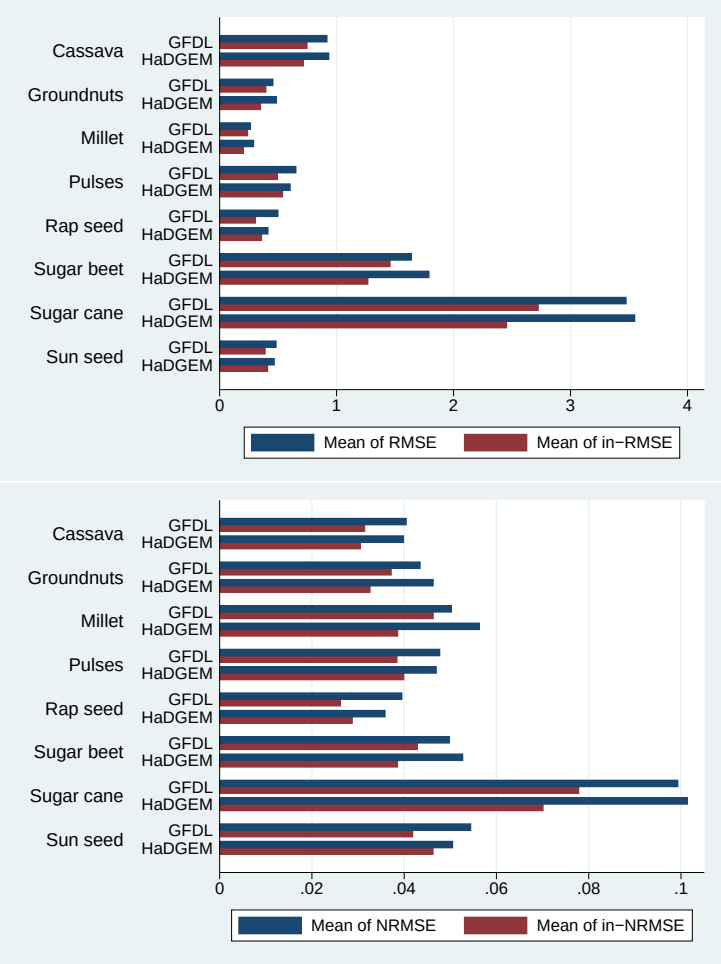
We also plot yields obtained over the 21st century with the excluded GCM and with the overall sample to visualize the difference. Figure 3.18 to figure 3.33 show, for the average on the globe and for the weighted average by area harvested, three graphs : with each GCM excluded and with the overall sample. On each case, the light color line corresponds to the GGCM and the full color corresponds to the statistical emulator. As expected, the correspondence between the GGCM and the statistical emulator is better with the overall sample than with the restricted sample. When a considerable difference is already observed on the global average for the leave-one-GCM-out graph, like for example for the sugar cane, this difference tends to be exacerbated with the weighted average.

TABLE 3.7: Ratio between out-of-sample and in-sample RMSE and NRMSE

Crop	GCM excluded	Ratio for RMSE	Ratio for NRMSE
Cassava	GFDL	1.23	1.29
	HaDGEM	1.30	1.31
Groudnuts	GFDL	1.15	1.17
	HaDGEM	1.39	1.41
Millet	GFDL	1.10	1.08
	HaDGEM	1.43	1.46
Pulses	GFDL	1.32	1.24
	HaDGEM	1.12	1.18
Rape seed	GFDL	1.62	1.51
	HaDGEM	1.15	1.25
Sugar beet	GFDL	1.13	1.16
	HaDGEM	1.41	1.37
Sugar cane	GFDL	1.28	1.28
	HaDGEM	1.45	1.45
Sun seed	GFDL	1.24	1.30
	HaDGEM	1.14	1.09

FIGURE 3.17: RMSE an NRMSE by crop and GCM excluded in comparison with in-sample results

The GCM excluded is not used for the estimation and then re-used as an input to compare the output with the in-sample output



3.5.4 Interpretation

On average for the different grid cells we find that our projections with the statistical emulator are close to the results of the GGCM. The RMSE and NRMSE tests have showed a good robustness (i.e., values under 1 for most of the crops) and the correlation is close to one and significant. On the maps, we have seen uneven distribution of the errors (whether the statistical emulator was under- or overestimating the effects of climate change) that we will deal with at the next section. More importantly, when running the model with out-of-sample data, the statistical emulator still perform.

This means that the crop emulator provides projections very similar to our benchmark, the GGCM, with very restricted and accessible variables. For instance, the use of this emulator could be used to project yields in the future with different climate scenarios or by government agencies to predict yields for the next years and adjust their local and international food policies. By being overall accessible and easy to handle, this model has the advantage of providing the accuracy of the GGCM to the widest audience, including countries with little means.

3.6 Analysis on the dependence performance

3.6.1 Motivation

The crop emulator provides an estimation of the marginal effect of climate change on crop yields at the grid cell level. This statistical estimation aims at reproducing the results of accurate and complex Global Gridded Crop Models which simulates the response to climate change on crop yields with a restricted number of weather variables. The previous parts of this paper show an overall acceptable goodness-of-fit of the statistical emulator to the GGCM. However, the specifications and the analysis of performance so far do not take into account the dependence across regions. As a matter of fact, we can observe on the maps representing the log-ratio of the predictions made by the GGCM and the crop emulator that regions are overall over- and under-estimating the effects of climate change compared to the GGCM.

Dependence across the regions can be of particular importance when considering potential trade between neighbored regions if climate change has opposite effects for the same crop. We have to make sure that the crop emulator provides the appropriate yields distribution in the future (compared to the benchmark, the crop model) so that policymakers, or insurers, could know how to redistribute food in the future depending on which regions are more or less affected. If culture habits change more slowly than climate change, the demand for certain types of food might be the same in the next century, even if the supply is shifted. Correlation is implicitly taken into account with the correlation of weather change between grid cells, but does not appear in our esti-

mation. For example, if the crop emulator tends to underestimate crop yields increase in the East of Africa (i.e., yields increase by 2t/ha instead of 4t/ha) and overestimate them in the West of Africa (i.e., increase of 6t/ha instead of 4t/ha), on average, the difference is null but the dependence pattern is very different between the two models. Typically in our example, we would have a higher correlation between the two areas with the statistical emulator than on the GGCM, the benchmark. If governments were to simulates crop yields in the future to define a trade scheme to offset climate change impacts, modeling the dependence across areas is of the essence. Consequently, we have to check whether our estimations reproduces the same dependence patterns as the GGCM.

3.6.2 Methodology

Data For this part of the study, we focus on Africa. Given the economic and political context, this continent is very likely to be one of the most affected by agricultural changes and losses due to climate change. For computational matters, we subdivide Africa into nine groups and not into countries. We take the mean of crop yield responses between the two climate scenario used for the estimation, for both the GGCM and S1fpint crop emulator. For the GGMC and S1fpint, we have nine groups with crop yields for eight crops over 125 years.

ARMA model For the validity of the model and so as to focus on dependence in changes, we fit our data to an ARMA model and then use the residuals to assess the dependence. The ARMA(p,q) is expressed as followed :

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-1} + \sum_{i=1}^{q} \theta_i \epsilon_{t-1}$$

For the eight crops studied, we look for the model that offers the best performance with the most parsimonious expression and fits to both the crop emulator and the GGCM. We choose ARMA(1,1) or ARMA(1,2) depending on the data (see appendix for the robustness tests), according to the distribution presented on table 3.9.

Vine Copula To represent the dependence across the nine groups of countries in Africa, we employ a Vine copula. This copula family allows a flexible fit of multivariate

TABLE 3.8: Groups of countries used

Group	Countries aggregated
North 1	Western Sahara - Algeria - Morocco - Tunisia - Libyan Arab Jamahiriya - Egypt
North 2	Senegal - Gambia - Guinea-Bissau - Mali - Burkina Faso - Niger - Chad - Sudan - Eritrea - Mauritania
North 3	Guinea - Sierra Leone - Liberia - Cote d'Ivoire - Ghana - Togo - Benin - Nigeria - Cameroon - Central African Republic
East	Kenya - Somalia - Ethiopia - United Republic of Tanzania - Djibouti
Center	Gabon - Equatorial Guinea - Democratic Republic of the Congo - Congo - Rwanda - Burundi - Uganda
South 4	Mozambique - Madagascar
South 3	Angola - Zambia - Malawi
South 2	Zimbabwe - Namibia - Botswana
South 1	Swaziland - South Africa - Lesotho

TABLE 3.9: ARMA models chosen for each crop

Crop	ARMA model
Cassava	(1,1)
Millet	(1,2)
Sunflower	(1,2)
Sugar cane	(1,1)
Sugar beet	(1,1)
Rape seed	(1,1)
Groundnuts	(1,1)
Pulses	(1,2)

dependence. Using the model developed by Joe (1996), Bedford and Cooke (2001), Aas et al. (2009), Czado (2010), we use the Vine Copula form as followed :

$$f(x_1, ..., x_d) = \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,(i+j)|(i+1),...,(i+j-1)} \prod_{k=1}^{d} f_k(x_k)$$

3.6.3 Results

We fit the crop yields obtained for the nine regions for both the GGCM and the S1fpint and then compare the results to each other. If the copula outputs are significantly similar, the statistical model does not omit dependence. We run the goodness of fit test of the copula object fitted to one the models to the data of the actual model, and look at the significance of the difference between S1fpint (or GGCM) when GGCM (respectively S1fpint) has been used to fit the copula. To complete this goodness-of-fit, we also simulate the values from the copula fitted, compute the difference for each vector (ie. each group of countries) and see whether the mean is significantly different from zero.

Table 3.10 shows the goodness-of-fit of the copula model estimated to a certain data set. This goodness-of-fit uses the information matrix equality of White (1982). This estimation requires a data set used to estimate the copula and the copula object estimated. We run the goodness-of-fit of the copulae estimated with the GGCM and the S1fpint models and check if the dependence is not significantly different. To measure the similarity between the statistical emulator and the GGCM, we also run the goodness-of-fit by crossing data from GGCM and S1fpint in the copula estimation and in the goodness-of-fit estimation. For example, we can use the GGCM to estimate the copula model, and then run the goodness-of-fit using the data from the S1fpint. If it is not significantly different, it means that the copula estimated with one model is also very close to the dependence trend of the other data set.

To evaluate the goodness of fit of our statistical model regarding the dependence pattern, we also look at the difference in the series of values for the copulas fitted with the GGCM and the S1fpint emulator. To this end, we simulate values for each crop from the copula fitted to both the GGCM and the S1fpint, and we then run a Student test to measure the significance of the difference between both series. Table 3.11 shows the p-value of the test, using either 50 or 100 values simulated from the two copulas. In

TABLE 3.10: Goodness-of-fit for fitted copula with different data sets

Crop	Data used to fit copula	Data compared	p-value
Cassava	GGCM	GGCM	0.43
	S1fpint	S1fpint	0.89
	GGCM	S1fpint	0.045
	S1fpint	GGCM	0.78
Millet	GGCM	GGCM	0.265
	S1fpint	S1fpint	0.49
	GGCM	S1fpint	0.37
	S1fpint	GGCM	0.81
Sunflower	GGCM	GGCM	-
	S1fpint	S1fpint	0.76
	GGCM	S1fpint	-
	S1fpint	GGCM	0.66
Sugar cane	GGCM	GGCM	-
	S1fpint	S1fpint	0.76
	GGCM	S1fpint	-
	S1fpint	GGCM	0.21
Sugar beet	GGCM	GGCM	0.295
	S1fpint	S1fpint	-
	GGCM	S1fpint	0.365
	S1fpint	GGCM	-
Rapeseed	GGCM	GGCM	0.21
	S1fpint	S1fpint	0.585
	GGCM	S1fpint	0.18
	S1fpint	GGCM	0.64
Groundnut	GGCM	GGCM	0.39
	S1fpint	S1fpint	0.565
	GGCM	S1fpint	0.685
	S1fpint	GGCM	-
Pulses	GGCM	GGCM	0.865
	S1fpint	S1fpint	-
	GGCM	S1fpint	0.15
	S1fpint	GGCM	0.905

most cases where we observe a significant different (ie. at the 10% significance level), it is in the case of 100 points simulated, corresponding to a 100-year prediction. In these cases, and when no significant difference is observed for 50 points, it shows that the statistical emulator is good for predictions out of 50 years but diverges too much for longer predictions. This is for example the case for groundnuts in five areas (South and North mostly). The only case where we observe significant difference for both 50 and 100 simulations is for sugar cane and for the eastern part of Africa. Apart from very isolated cases like this one, significant differences are rarely observed and mostly for the cases with 100 observations. Hence, in a large majority of the cases, the crop emulator provides predictions in the future with a similar dependence pattern as the GGCM. More generally, the overall fitting makes more sense than interpreting each region for each crop separately since we are looking at the interdependence across Africa. Here, groundnuts would be the most questionable case with five regions out of nine being significantly different as from a sample of 100 years.

TABLE 3.11: p-value of the t-test between the data generated from the copulas fitted on GGCM and S1fpint

Crop	Number of simulations	Center	East	North 1	North 2	North 3	South 1	South 2	South 3	South 4
Cassava	50	0.62	0.71	0.35	0.33	0.11	0.47	0.16	0.81	0.44
	100	0.65	0.12	0.88	0.15	0.26	0.54	0.38	0.54	0.70
Millet	50	**0.08**	0.31	0.98	0.30	0.47	0.77	**0.02**	0.99	0.89
	100	0.19	0.43	0.55	0.50	**0.05**	0.42	0.55	0.13	0.63
Sunflower	50	0.84	0.50	0.14	0.39	0.72	0.38	0.87	0.63	0.98
	100	**0.09**	**0.05**	0.15	0.82	0.85	0.31	0.75	0.28	0.11
Sugar cane	50	**0.07**	0.58	0.13	0.31	0.65	0.66	0.71	0.13	0.29
	100	**0.04**	0.45	0.10	0.70	0.42	**0.06**	0.85	0.43	0.31
Sugar beet	50	0.46	0.99	0.40	0.38	0.81	0.37	0.45	0.47	0.37
	100	0.47	0.63	0.51	0.91	0.70	0.13	0.18	0.55	**0.01**
Rape seed	50	0.66	0.53	0.39	0.23	0.80	0.94	0.91	0.82	0.83
	100	0.98	0.51	0.75	0.71	0.90	0.31	**0.01**	0.93	**0.04**
Groundnuts	50	0.35	0.87	0.65	0.49	0.39	0.65	0.60	0.75	0.83
	100	0.36	**0.04**	**0.00**	0.93	0.60	**0.02**	**0.07**	0.25	**0.01**
Pulses	50	0.58	0.21	**0.03**	0.65	0.75	0.41	0.88	0.71	0.48
	100	0.74	0.97	0.13	0.40	0.81	0.30	0.51	0.90	0.36

3.7 Concluding remarks

This statistical emulator made for cassava, millet, sunflower, sugar cane, sugar beet, rape seed/canola, groundnuts/peanuts and pulses at the grid-cell level provides an accurate estimation of crop yields in the future initially made by the GGCM. Using statistical estimations for each crop, soil and climate scenario, we obtain estimates for the marginal impacts of changes in mean temperature, mean precipitation and CO_2 concentration that could be generalized to other climate scenarios. Our measure of the goodness-of-fit (through many common tools such as RMSE, log-ratio between the two models, out-of-sample validation, etc) shows an acceptable fit to the 'perfect' model represented by the GGCM. Moreover, in addition to the goodness-of-fit averaged on the grid cells, we have showed that the crop emulator also captures the dependence pattern of the crop model. Hence, the crop emulator as designed here, could be used as an input for future regional food redistribution or insurance scheme.

3.8 Appendix

3.8.1 The fractional polynomial model

A fractional polynomial model of degree m is defined by the following equation :

$$Y = \alpha_0 + \sum_{i=1}^{m} \alpha_i X^{(p_i)} + \mu$$

$$X^{(p_i)} = \begin{cases} X^{p_i} & if \quad p_i \neq 0 \\ lnX & if \quad p_i = 0 \end{cases}$$

Following Royston and Sauerbrei (2008), powers are chosen among the set $\{$-2, -1, -0.5, 0, 0.5, 1, 2, 3$\}$ and the maximum allowed degree will be $m = 2$. The selection is made using the Royston and Altman model-selection algorithm.

3.8.2 Goodness-of-fit

The Root-Mean-Square Error (RMSE) is a measure of the differences between values predicted by the statistical model and the crop model. With N predictions, predicted values $\hat{y}$ and a dependent variable $\hat{y}$, it is calculated as followed :

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y} - y)^2}{N}}$$

The normalized RMSE (NRMSE) allows the comparison between datasets with different scales (y_{max} and y_{min} being respectively the maximum and the minimum values of the observed data) :

$$NRMSE = \frac{RMSE}{y_{max} - ymin}$$

3.8.3 Geographic dependence

TABLE 3.12: Results of the t-tests to compare the average log-ratios with zero

Crop	Mean	$P(\|T\| > \|t\|)$
Cassava	-0.64	0.00
Millet	-1.43	0.00
Sunflower	-0.14	0.00
Sugar cane	-1.13	0.00
Sugar beet	-0.14	0.00
Rape seed/canola	-0.73	0.00
Groundnuts/peanuts	-1.11	0.00
Pulses	-0.20	0.00

FIGURE 3.18: Average cassava crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

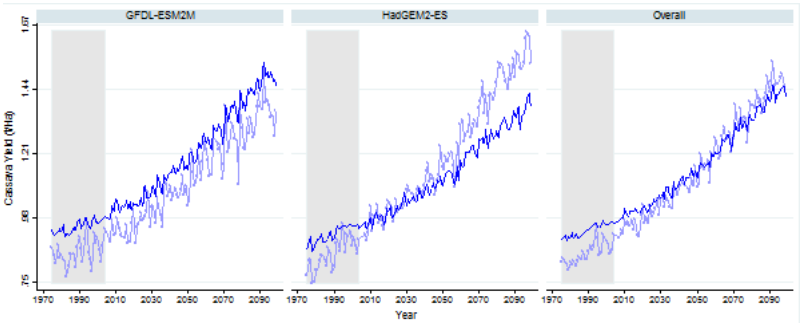


FIGURE 3.19: Average cassava crop yields projections weighted by area harvested

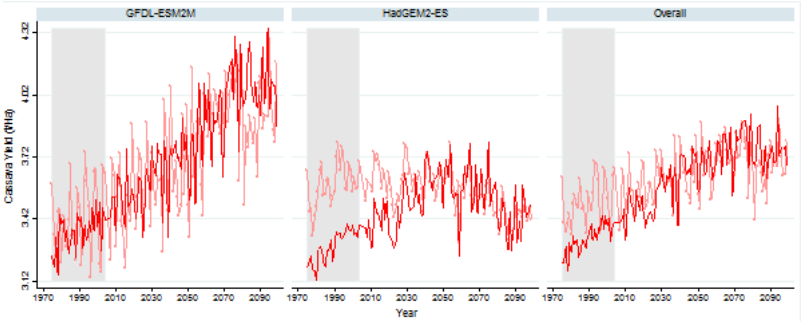


FIGURE 3.20: Average millet crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

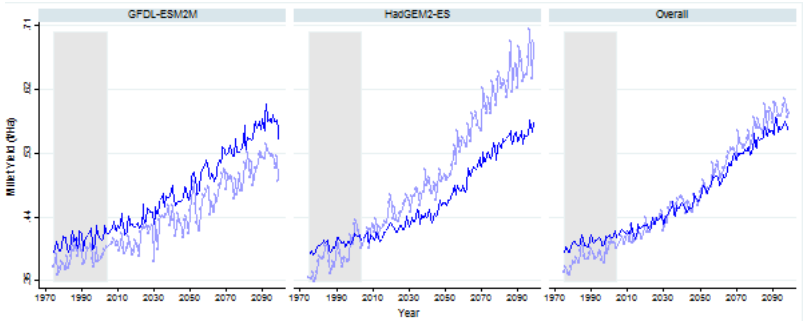


FIGURE 3.21: Average millet crop yields projections weighted by area harvested

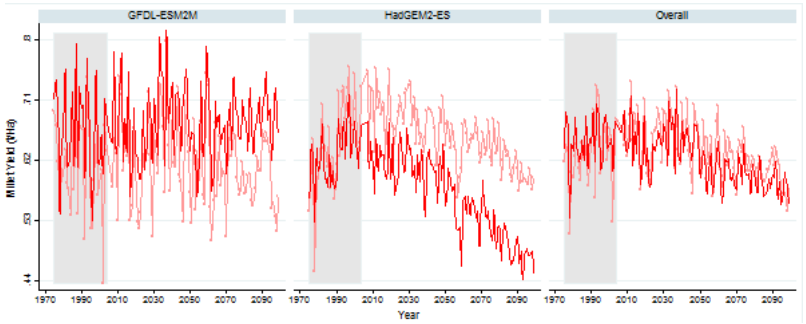


FIGURE 3.22: Average sunflowers crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

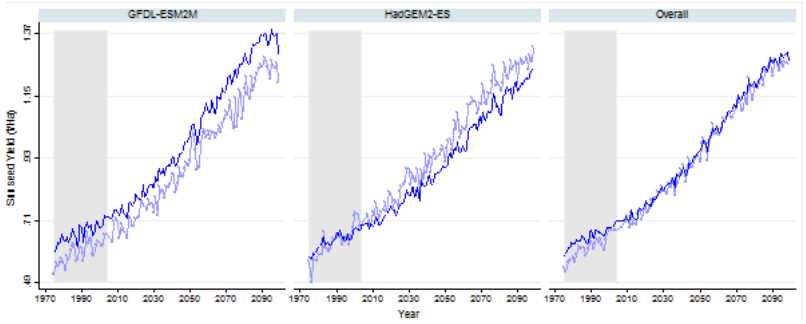


FIGURE 3.23: Average sunflowers crop yields projections weighted by area harvested

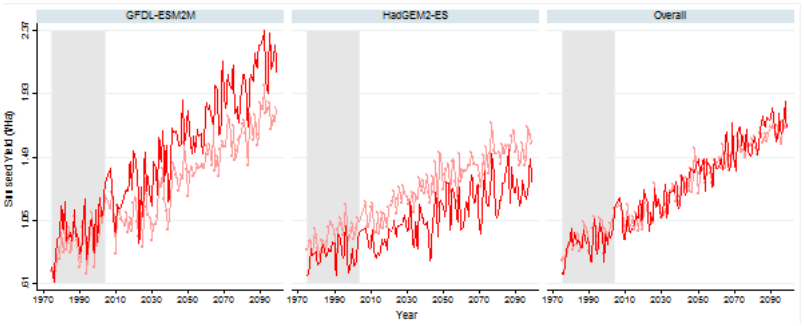


FIGURE 3.24: Average sugar cane crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

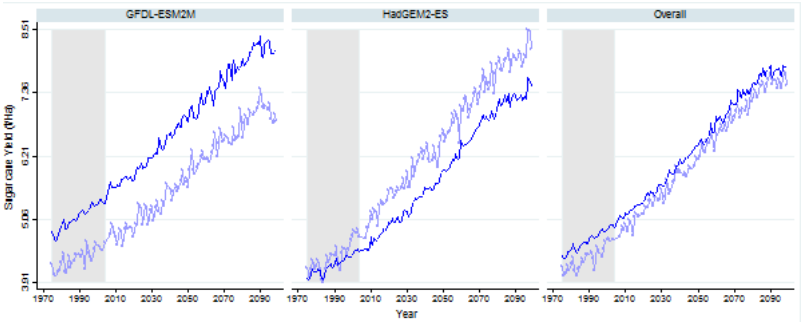


FIGURE 3.25: Average sugar cane crop yields projections weighted by area harvested

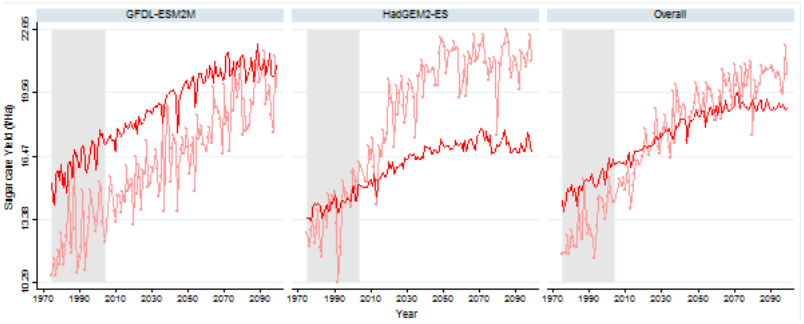


FIGURE 3.26: Average sugar beet crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

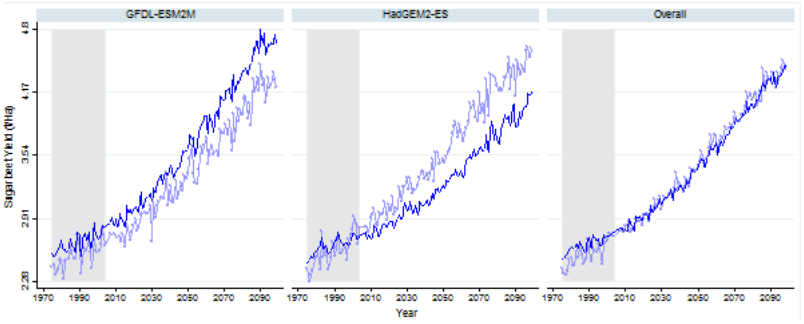


FIGURE 3.27: Average sugar beet crop yields projections weighted by area harvested

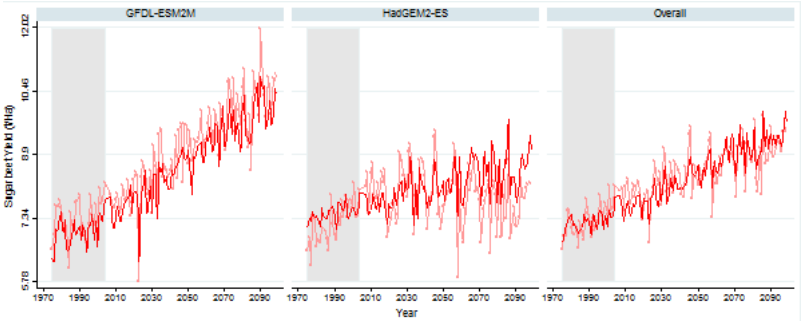


FIGURE 3.28: Average rape seeds / canola crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

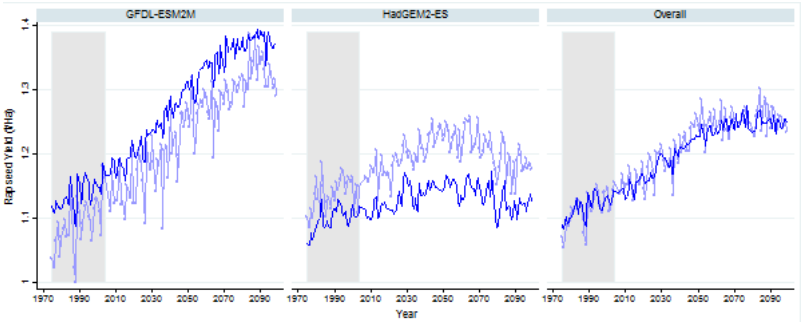


FIGURE 3.29: Average rape seeds / canola crop yields projections weighted by area harvested

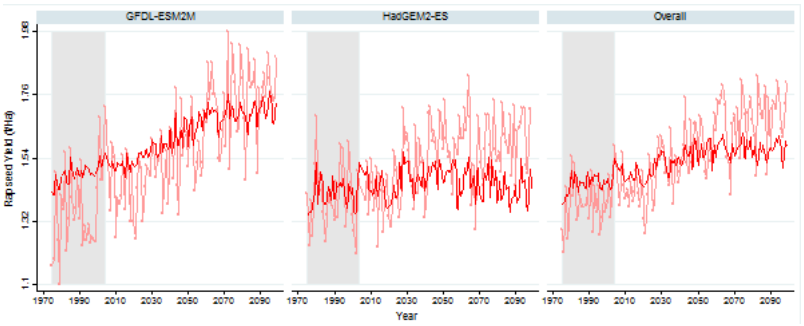


FIGURE 3.30: Average groundnuts / peanuts crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

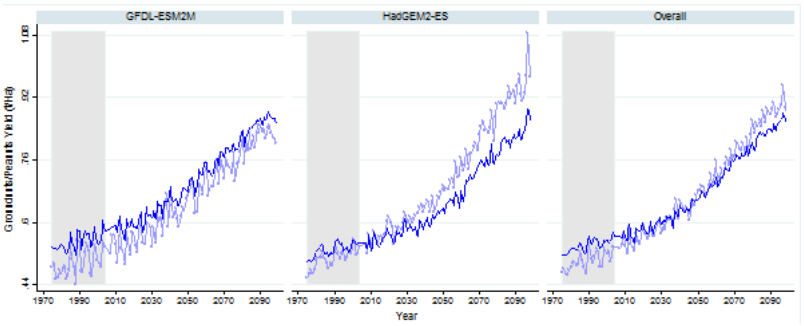


FIGURE 3.31: Average groundnuts / peanuts crop yields projections weighted by area harvested

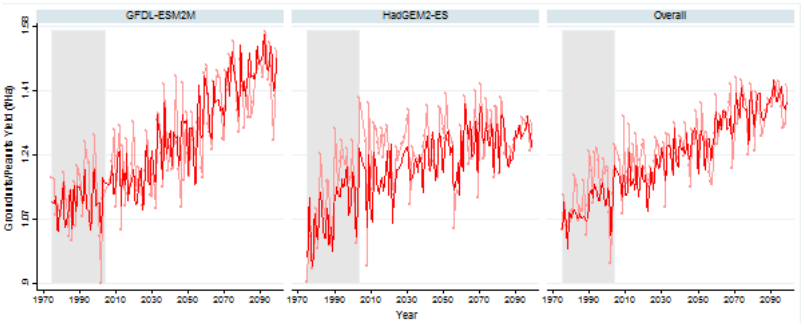


FIGURE 3.32: Average pulses crop yields projections from GGCM and statistical emulator with leave-one-GCM-out validation and overall sample

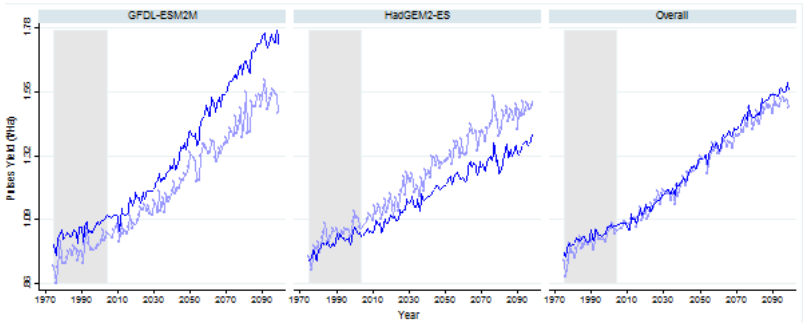


FIGURE 3.33: Average pulses crop yields projections weighted by area harvested

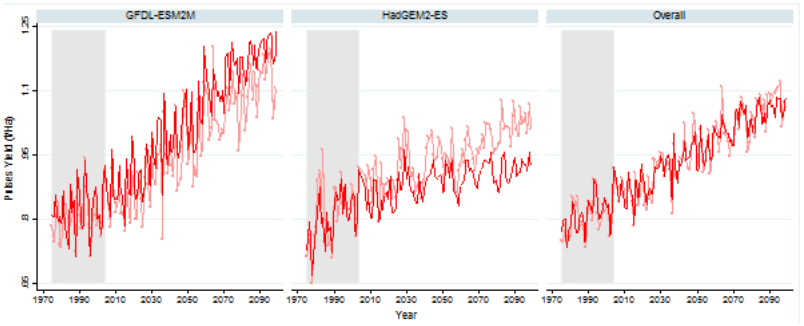


FIGURE 3.34: African climate zones (Source : United Nations)**??**

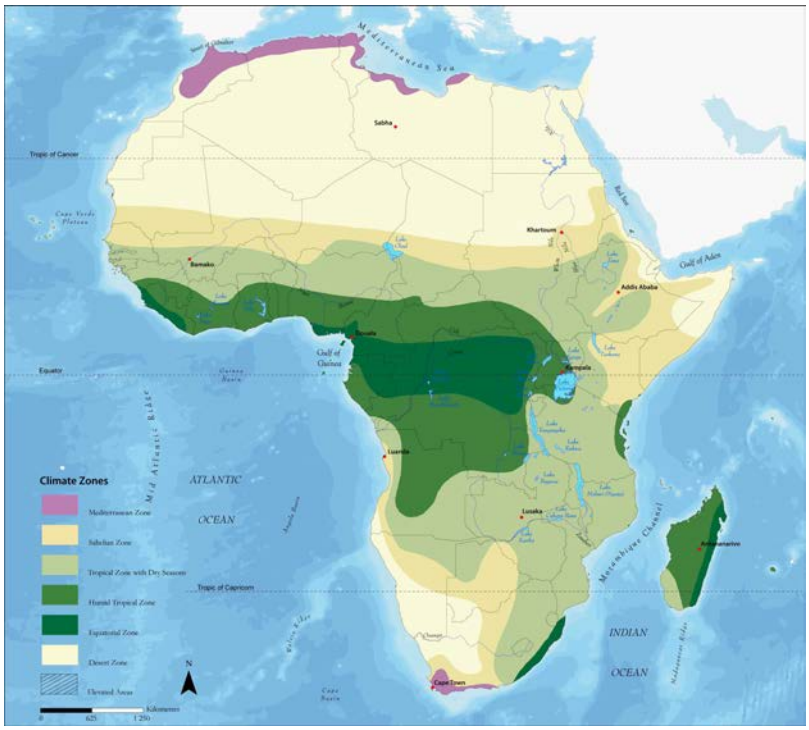


FIGURE 3.35: R Vine Copula matrix fitted to the crop emulator data

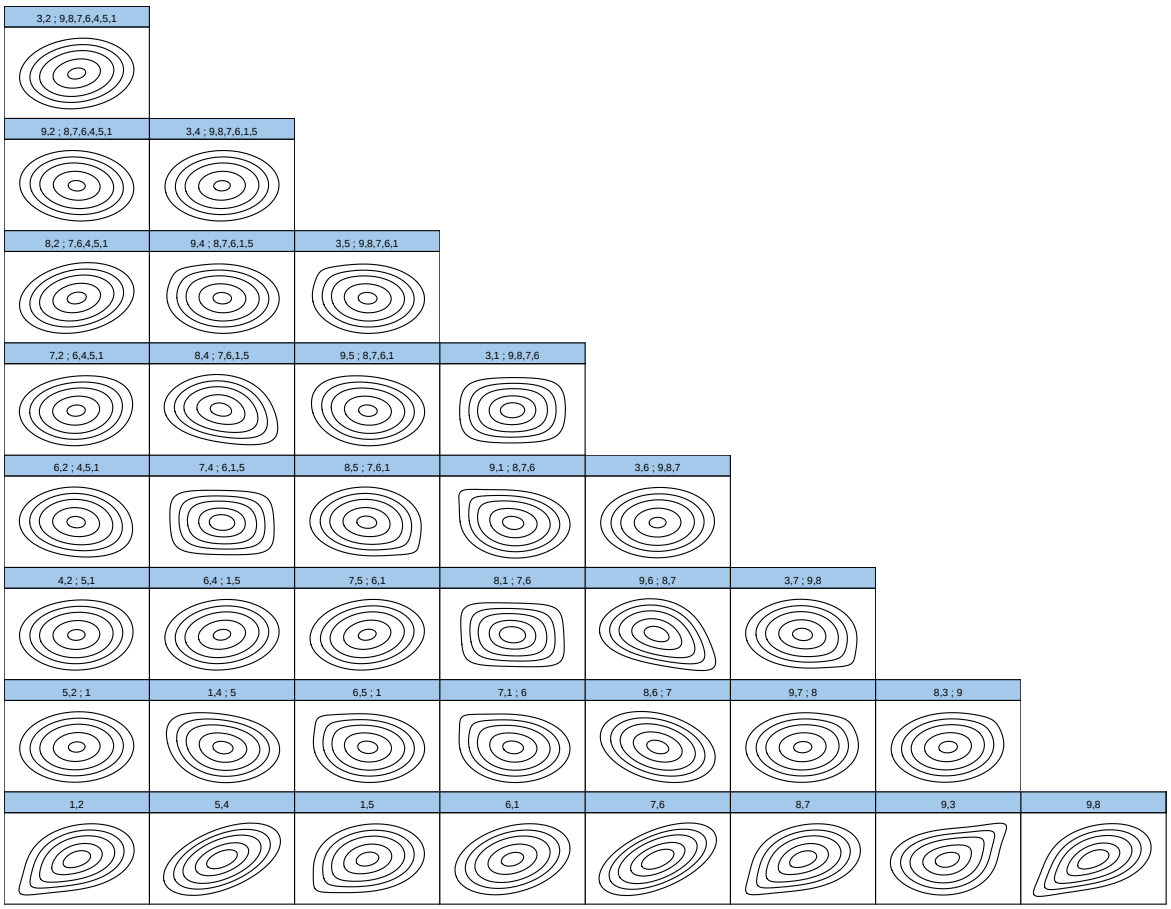


FIGURE 3.36: R Vine Copula matrix fitted to the GGCM data

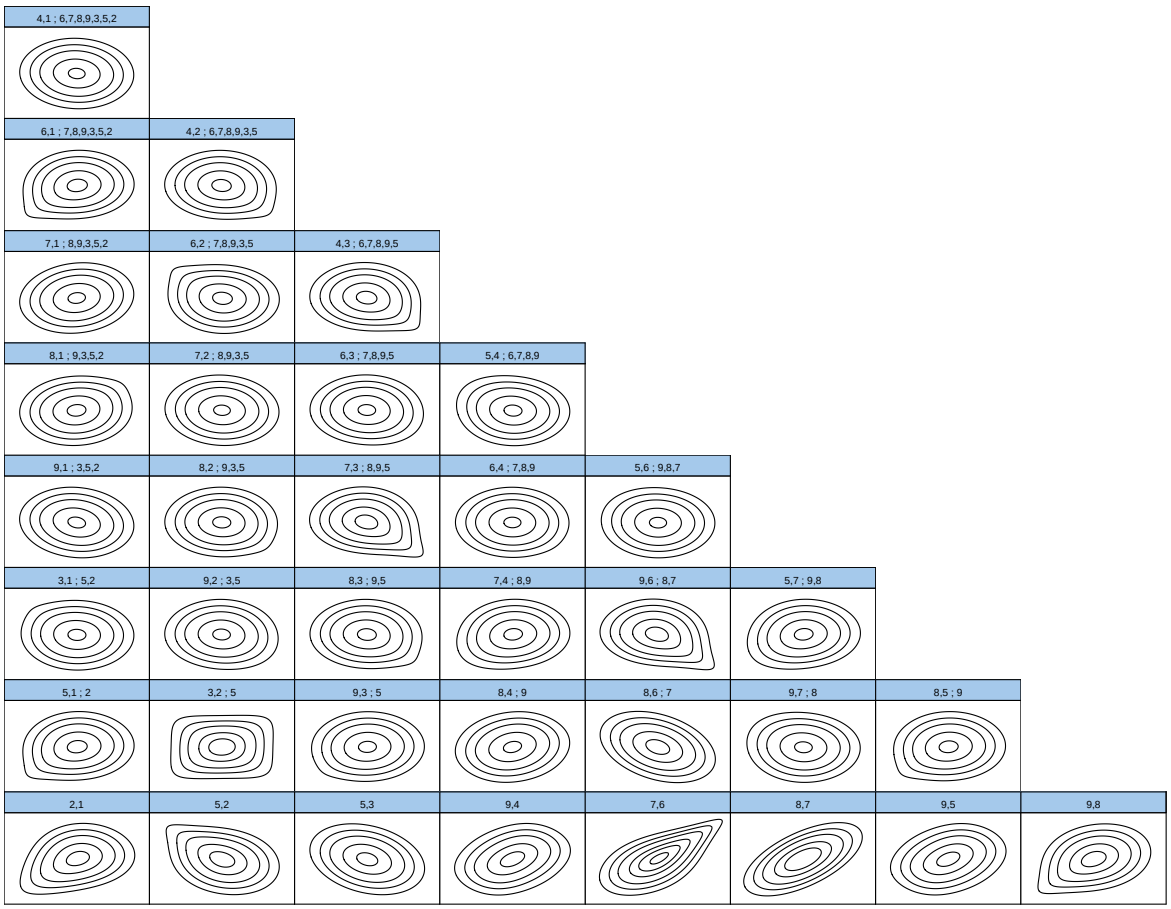


TABLE 3.13: p-values of the Box tests on the residuals of the chosen ARMA model for each crop, by group of countries - Case of the crop emulator

Crop	cassava	millet	sunflower	sugar cane	sugar beet	rape seed	groundnuts	pulses
ARMA model	(1,1)	(1,2)	(1,2)	(1,1)	(1,1)	(1,1)	(1,1)	(1,2)
Center	0.42	0.99	0.31	0.58	0.51	0.36	0.54	0.99
East	0.41	0.26	0.75	0.99	0.54	0.95	0.70	0.71
North 1	0.37	0.86	0.13	0.85	0.03	0.14	0.10	0.31
North 2	0.99	0.99	0.54	0.94	0.39	0.73	0.49	0.47
North 3	0.52	0.65	0.67	0.74	0.17	0.22	0.67	0.79
South 1	0.49	0.97	0.56	0.73	0.96	0.30	0.96	0.98
South 2	0.15	0.99	0.16	0.95	0.37	0.10	0.34	0.88
South 3	0.62	0.60	0.22	0.70	0.62	0.88	0.70	0.85
South 4	0.12	0.97	0.58	0.99	0.19	0.27	0.11	0.94

TABLE 3.14: p-values of the Box tests on the residuals of the chosen ARMA model for each crop, by group of countries - Case of the GGCM

Crop	cassava	millet	sunflower	sugar cane	sugar beet	rape seed	groundnuts	pulses
ARMA model	(1,1)	(1,2)	(1,2)	(1,1)	(1,1)	(1,1)	(1,1)	(1,2)
Center	0.65	0.00	0.44	0.03	0.52	0.55	0.29	0.80
East	0.64	0.98	0.02	0.32	0.18	0.13	0.18	0.03
North 1	0.61	0.10	0.28	0.66	0.89	0.99	0.02	0.94
North 2	0.62	0.92	0.74	0.38	0.31	0.77	0.80	0.14
North 3	0.71	0.07	0.33	0.66	0.12	0.29	0.99	0.19
South 1	0.03	0.99	0.16	0.41	0.39	0.17	0.38	0.15
South 2	0.57	0.93	0.23	0.44	0.91	0.91	0.45	0.59
South 3	0.86	0.06	0.68	0.87	0.67	0.69	0.80	0.95
South 4	0.36	0.99	0.87	0.59	0.83	0.61	0.58	0.93

Bibliographie

[1] E. BLANC et E. STROBL. “Assessing the Impact of Typhoons on Rice Production in the Philippines”. In : *Journal of Applied Meteorological and Climatology* 55.4 (2016), p. 993–1007. DOI : 10.1175/JAMC-D-15-0214.1.

[2] E. BLANC et B. SULTAN. “Emulating maize yields from global gridded crop models using statistical estimates”. In : *Agric. For. Meteorol.* (2015), p. 134–147.

[3] Potsdam Institute for CLIMATE IMPACT RESEARCH. *LPJmL - Lund-Potsdam-Jena managed Land*. URL : https://www.pik-potsdam.de/research/projects/activities/biosphere-water-modelling/lpjml.

[4] Y. KANG, S. KHAN et X. MAA. “Climate change impacts on crop yield, crop water productivity and food security – A review”. In : *Progress in Natural Science* 19.12 (déc. 2009), p. 1665–1674. URL : https://doi.org/10.1016/j.pnsc.2009.08.001.

[5] D. B. LOBELL et M. B. BURKE. “On the use of statistical models to predict crop yield responses to climate change”. In : *Agric. For. Meteorol.* 150.11 (oct. 2010), p. 1443–1452. URL : https://doi.org/10.1016/j.agrformet.2010.07.008.

[6] M. MORIONDO, C. GIANNAKOPOULOS et M. BINDI. “Climate change impact assessment: the role of climate extremes in crop yield simulation”. In : *Climatic Change* 104 (fév. 2011), p. 679–701. DOI : DOI10.1007/s10584-010-9871-0.

[7] F.T. PORTMANN, S. SIEBERT et P. DÖLL. “MIRCA2000 – global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling”. In : *Glob. Biogeochem. Cycles* (2010), p. 24.

[8] *Price Volatility in Food and Agricultural Markets: Policy Responses*. Rapp. tech. FAO et OECD, June 2011.

[9] P. ROUDIER et al. “The impact of future climate change on West African crop yields: What does the recent literature say?” In : *Global Environmental Change* 21.3 (août 2011), p. 1073–1083. URL : https://doi.org/10.1016/j.gloenvcha.2011.04.007.

[10] W. SCHLENKER et M. J. ROBERTS. “Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change”. In : *Proc. Natl. Acad. Sci* 106.37 (sept. 2009), p. 15594–15598. URL : https://doi.org/10.1073/pnas.0906865106.

[11] Soil Survey STAFF. *Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys*. Natural Resources Conservation Service, U.S. Department of Agriculture Handbook 436, 1999.

[12] L. WARSZAWSKI et al. “The inter-Sectoral impact model intercomparison project (ISI-MIP): project framework”. In : *Proc. Natl. Acad. Sci* 9.111 (2014), p. 3228–3232.

[13] M. YESUF et R. BLUFFSTONE. “Risk Aversion in Low-Income Countries: Experimental Evidence from Ethiopia”. In : *American Journal of Agricultural Economics* 91.4 (2009), p. 1022–1037.

W :/agoburdhun/Documents/Dissertation/bibliocrop.bib

université PARIS-SACLAY | ÉCOLE DOCTORALE Sciences de l'homme et de la société (SHS)

Titre : Trois chapitres sur la gestion et la corrélation du risque, et le risque météorologique

Mots clés : Risque, agriculture, corrélation, catastrophes naturelles, développement

Résumé : La thèse étudie le risque météorologique et économique sous différents angles principalement dans les pays en développement. Elle se décompose en trois chapitres indépendants analysant dans diverses situations la corrélation des risques liés aux aléas météorologique et climatique ou économique, et étudie le potentiel de la région géographique étudiée pour mettre en place un système d'assurance contre le risque étudié. En effet, cette thèse étudie des risques très susceptibles d'être fortement corrélés : que cela concerne le risque météorologique ou climatique, ou le risque lié à la volatilité des prix, les villes voire pays voisins sont exposés aux mêmes risques et de façon simultanée. Cet aspect essentiel compromet la mutualisation du risque, paramètre primordial du modèle économique de l'assurance. A travers les trois chapitres de la thèse, nous étudierons le bénéfice lié à la mutualisation de ces risques a priori relativement corrélés. Le premier chapitre étudie la corrélation des prix du maïs en Tanzanie. A l'aide d'un modèle Copula-GARCH, la dépendance entre les cours du maïs des 20 marchés principaux du pays est modélisée et nous pouvons voir si le prix moyen du maïs est lissé en agrégeant les marchés. Cela permet de voir si l'intégration des marchés permet une efficace mutualisation du risque lié à la volatilité des prix. Le second chapitre s'attache au risque cyclonique dans les îles Pacifique sud et son impact sur les infrastructures. Ce papier propose une modélisation des cyclones tropicaux dans la région étudiée et la distribution de probabilité des cyclones associés à leur force, permettant ainsi de tenir compte du climat actuel pour modéliser les coûts. Avec les données liées aux infrastructures, nous calculons le coût des cyclones, y compris pour les événements extrêmes de très faible probabilité. Le troisième chapitre propose une extension d'un émulateur statistique des rendements agricoles selon des variables climatiques. Nous modélisons l'impact de l'accroissement marginal de la température, des précipitations ou de la concentration en CO2 en faisant une estimation statistique sur des modèles de culture et non sur des données historiques. Cela permet de prendre en compte des effets extrêmes sur des valeurs météorologiques pas ou peu observées jusqu'à présent. La robustesse du modèle est évaluée, entre autres, à l'aide de copules pour comparer la dépendance spatiale entre le modèle et notre émulateur statistique et vérifier que notre estimation capture bien la dépendance géographique.

Title : Three essays on risk management and correlation, and meteorological hazard

Keywords : Risk, agriculture, correlation, natural disasters, development

Abstract : The PhD dissertation studies meteorological and economic hazard under different angles and mostly in developing countries. It is composed of three independent chapters analyzing different situations dealing with meteorological and climatic or economic hazard correlation. It estimates the potential of the studied regions for implementing an insurance scheme for the risk. Indeed, this thesis studies risks very likely to be highly correlated: whether this is for the meteorological or climatic hazard, or the price volatility risk, neighbored cities or even countries are exposed to the same risk simultaneously. This essential aspect jeopardizes risk mutualization, a key parameter of the economic insurance model. Through the three chapters of this thesis, we study the benefits linked to the mutualization of a priori correlated risks. The first chapter deals with maize price correlation in Tanzania. Using a Copula-GARCH model, we model the dependence among the 20 main markets of the country and assess if the mean maize price is smoothed by aggregating the markets. Hence, we see whether markets integration allows an efficient risk mutualization against the risk of price volatility. The second chapter deals with tropical storms risk in the South Pacific islands and their impact on infrastructures. This paper proposes an artificial tropical cyclones modeling in the region studied as well as the probability distribution of the cyclone's occurrence and strength. This enables us accounting for the current climate for modeling costs. With data on infrastructures, we calculate the cost due to tropical storms, including for very low probability extreme events. The third chapter proposes an extension for a statistical emulator of crop yields depending climatic variables. We model the marginal impact of an increase of temperature, precipitations and CO2 concentration by running a statistical estimation on crop models rather than historical data. It allows accounting for extreme effects caused by meteorological data values not observed so far. The model robustness is assessed, among others, with copulas to compare the spatial dependence between the model and our statistical emulator and check that our estimation captures the geographic dependence.