HEC MONTRÉAL

DÉVELOPPEMENT D'UN MODÈLE STOCHASTIQUE AVEC
DÉPENDANCE SPATIALE DU RAYONNEMENT SOLAIRE

Bi Tiessé Théophile Irié

Sciences de la gestion
(Méthodes analytiques de gestion)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M.Sc.)

Juin, 2014

Résumé

Planifier l'installation d'une centrale d'énergie solaire pour plusieurs sites nécessite l'évaluation du risque qui elle-même se traduit par une évaluation de la variabilité dans la production. Ce mémoire présente une approche d'évaluation de ce risque basée sur les copules. Des mesures de rayonnement solaire compilées de 2007 à 2011 et disponibles gratuitement sur le réseau Surfrad du gouvernement Américain sont utilisées pour estimer les modèles nécessaires à cette approche. En effet, ces données ont permis d'estimer pour chaque site associé, un processus stochastique de type ARMA et d'établir une structure de dépendance spatiale des innovations de ces processus. Nous considérons les copules, particulièrement la « R-Vine » pour estimer la dépendance spatiale, en plus du modèle simple qui existe, c'est-à-dire la Normale multivariée. On pourrait donc se poser la question de savoir si les copules permettent une meilleure estimation de cette dépendance. Pour réponse à la question, des simulations d'énergie solaire ont permis de comparer les résultats obtenus avec la « R-Vine » et la Normale multivariée.

Mots-clés : Énergie solaire, séries chronologiques, copules bivariées, copules en vigne régulière ou « R-Vine », simulation, ajustement de modèle.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Je rends gloire à Dieu en qui j'ai mis ma foi et qui me guide dans tout ce que j'entreprends. Je tiens à remercier mon directeur de recherche, Jean-François Plante, pour son soutien moral et financier et particulièrement ses précieux conseils tout au long de la rédaction de ce mémoire. Sa patience et sa disponibilité m'ont aidé à élaborer de façon efficace toutes les étapes et analyses des données du mémoire.

Je saisis l'occasion pour remercier les professeurs du service de l'enseignement des méthodes quantitatives de gestion, pour leur disponibilité et leur professionnalisme. Je remercie le professeur Bruno Rémillard qui a bien voulu mettre à ma disposition quelques programmes informatiques pour mes tests d'adéquation. Et à tous les professeurs du service d'enseignement des méthodes quantitatives de gestion qui m'ont conseillé et inspiré pendant mon séjour à HEC Montréal. Merci à Gilles Caporossi, Sylvain Perron et Erick Delage.

Finalement, je remercie chaleureusement tous les membres de ma famille. Merci à mes trois enfants Gilles, Marc et Anne pour leur soutien durant les périodes difficiles que nous avons traversées. Ces situations feront désormais partie du passé. Et un merci spécial à mon épouse Marie Louise pour avoir traversé avec moi toutes ces périodes émouvantes à HEC Montréal.

Chapitre 1 : Introduction

La production d'une centrale d'énergie solaire est directement reliée à la quantité de rayonnement solaire reçue. Planifier des investissements pour développer cette énergie renouvelable nécessite de bien connaître le potentiel des différents sites convoités, donc de bien comprendre comment le rayonnement solaire varie, dans le temps et l'espace, dans les zones de production.

Des données de rayonnements solaires sont compilées, depuis la fin des années 90, par des organismes tels qu'Environnement Canada et l'Agence Nationale Océanique et Atmosphérique du Département du commerce des États-Unis d'Amérique. Certaines entreprises privées spécialisées dans l'évaluation des ressources d'énergie solaire disposent également de données. Par exemple, Turquoise Technologies a développé des algorithmes pour estimer le rayonnement solaire, quel que soit la localisation du site, et cela à partir des données provenant des satellites. Ces données compilées sont une source pour évaluer le potentiel énergétique des sites concernés et permettent aussi d'estimer la production de l'énergie électrique associée. Dans le cadre de notre étude, nous utilisons les données compilées par l'Agence Nationale Océanique et Atmosphérique Américaine (NOAA) qui sont disponibles et accessibles gratuitement à partir du lien internet *Surfrad (Surface Radiation) Network* sur le site internet de l'Agence (*http://www.esrl.noaa.gov).* La production d'énergie solaire d'un site dépend de certaines incertitudes liées aux conditions atmosphériques et/ou climatiques. Ces données nous permettent d'estimer le potentiel de plusieurs sites simultanément, et donc d'évaluer :

- Le risque lié à la production de l'énergie électrique (solaire) par site. Étant donné que nous avons des données de séries chronologiques, la variance des innovations ajustées aux données jouera un rôle clé. La magnitude de la variabilité dans les innovations d'un site ou des sites conjointement corrélés selon les modèles

d'ajustement aura un effet direct sur l'incertitude dans la prévision de la production énergétique de ce site ou de ces sites d'une période à l'autre.

- La dépendance spatiale des sites. Cette dépendance sera représentée par le calcul des corrélations entre les résidus de différents sites, et cela en modélisant une dépendance entre les innovations des modèles de séries chronologiques estimés pour chaque site. Mais qu'est-ce que la dépendance spatiale permet de comprendre dans notre étude? Considérons deux sites situés dans des espaces géographiques différents et appartenant à un même portefeuille d'investissement. Si nous arrivons à établir une structure de dépendance entre ces sites, il sera possible de comprendre le comportement de la production énergétique d'un site si la production de l'autre site (et vice versa) subit des variations majeures lié aux conditions atmosphériques (importante couverture nuageuse, pluie) et/ou climatiques (longue période hivernale avec moins d'ensoleillement). Donc connaître la structure de dépendance entre les marginales permet aussi d'évaluer le risque qui a été mentionné précédemment.

Représenter la dépendance en calculant les corrélations des innovations suppose l'ajustement d'une distribution normale multivariée aux innovations. Mais ce modèle est-il adapté? Négliger cette corrélation spatiale conduirait à une estimation biaisée du risque. Serait-il hasardeux et insuffisant de lier la structure de dépendance des sites au calcul de la corrélation? Cela simplifierait les propriétés statistiques des données, de l'énergie solaire, observées. Une approche que nous proposons dans ce mémoire pour capter la structure de dépendance entre les innovations des sites est basée sur les copules (Genest et Favre, 2007) et spécifiquement sur la « R-Vine » (Kramer et Schepsmeier, 2011). Cette méthode flexible à construire peut capter l'apparition d'évènements extrêmes simultanés. Mais qu'est-ce que nous gagnons à utiliser la méthode « R-Vine »? Pour répondre à cette question nous allons comparer l'ajustement d'une production d'énergie solaire par une distribution normale multivariée et des modèles « R-Vine » que nous aurons estimés.

Cet objectif nous permet une approche de la question en deux étapes; il s'agit premièrement d'identifier des sites, d'extraire les données compilées de rayonnement solaire pour chaque site sur une période donnée (2007 – 2011) et certaines informations pertinentes de la position géographique des sites afin de construire un modèle physique. Nous estimons ensuite un processus stochastique stationnaire. Les innovations de chaque processus stochastique estimé, permettront d'établir la corrélation géographique entre les sites étudiés. À partir des innovations, de ces mêmes processus stochastiques stationnaires, il sera estimé une structure de dépendance à partir des copules. Deuxièmement, nous analysons l'avantage d'ajuster une production d'énergie solaire avec une distribution normale multivariée au lieu d'une distribution plus flexible basée sur les copules. Cette étape pourrait répondre à un modèle d'affaire concernant un portfolio d'investissements dans le domaine de l'énergie renouvelable.

Boland (2008) s'est intéressé à la modélisation en série chronologique du rayonnement solaire journalier en utilisant les séries de Fourier pour arriver à un processus de type *Autoregressive Moving Average* ou ARMA avec une distribution normale univariée qui a la particularité de s'ajuster à ses données simulées à partir d'un modèle physique d'énergie solaire. Dans notre étude, nous disposons des données réelles de rayonnement solaire journalier sur plus de cinq années. Nous nous inspirons de l'approche proposée par Boland (2008) et Nowicka-Zagrajek et Weron (2002), pour estimer nos modèles de série chronologique par une transformation des données compilées et celles obtenues à partir de nos modèles physiques de rayonnement solaire et cela pour chaque site. L'approche que nous suggérons dans cette étude, et qui diffère de celle de Boland (2008), est de déterminer pour chacun des sites un modèle physique de rayonnement solaire (Iqbal, 1983) qui représente la quantité maximale d'énergie solaire qu'un site recevrait en tenant compte de la latitude, mais en supposant l'absence de l'atmosphère et de la saison. Ensuite, nous estimons un modèle de série chronologique univariée pour chaque site, et déterminons les corrélations entre les résidus de ces sites en définissant une dépendance par les copules, en particulier les R-

vine[1] (Kramer et Schepsmeier, 2011). Nous comparons finalement la production de l'énergie électrique solaire telle qu'estimée par la normale multivariée et par les modèles R-Vine estimés.

Ce mémoire est donc organisé de la façon suivante : dans le chapitre 2, une revue de littérature décrit le potentiel énergétique solaire, l'évaluation des données solaires, les modèles d'énergie solaire élaborés et finalement un exposé sur l'usage des copules. Signalons à ce propos que ce deuxième chapitre décrit les caractéristiques d'un modèle physique d'énergie solaire et les méthodes de calcul de certains paramètres associés. Dans le chapitre 3, nous exposons la méthodologie pour estimer les séries chronologies, après cela nous présentons l'origine et les caractéristiques des données utilisées, ensuite nous déterminons l'élaboration du modèle physique d'énergie solaire et enfin appliquons les séries chronologiques à nos données transformées, présentons une synthèse des résultats avant de conclure. Les modèles de dépendance seront au chapitre 4. Nous exposerons dans ce même chapitre les tests d'adéquation pour les copules qui s'ajustent le mieux aux innovations. Nous simulerons au chapitre 5 une production journalière d'énergie avec la normale multivariée et les modèles déterminés par la R-Vine afin de comparer et de savoir le gain réalisé avec l'approche des copules. Nous présenterons notre conclusion au chapitre 6.

[1] R-Vine : méthode d'estimation régulière des copules en vignes (ou Vine-Copulas).

Chapitre 2 : Revue de littérature

La mesure du rayonnement solaire sur un site constitue une source possible pour évaluer le potentiel énergétique associé à ce site et estimer sa production d'énergie électrique solaire. Cette énergie suscite beaucoup d'intérêt chez des particuliers, des industriels et des producteurs d'électricité, mais il n'est pas courant de disposer des mesures réelles du rayonnement solaire. Ce manque d'informations pertinentes (mesure de données sur la quantité d'énergie captée sur un site) ne facilite pas souvent une évaluation efficiente du potentiel énergétique du site. Selon Younes et *coll.,* (2005) : « [...], il peut avoir un biais dans l'estimation et la prévision d'énergie solaire d'un site s'il n'existe pas de mesure du rayonnement solaire ». Mais pour pallier ce manque d'informations pertinentes, des modèles théoriques (physiques) de rayonnement solaire sont développés pour simuler ce potentiel de production d'énergie solaire. Il importe donc de parcourir différents articles et publications englobant ce domaine d'étude afin de bien cerner ce qui a été réalisé. Dans notre étude, nous disposons des mesures de rayonnement solaire sur les sites comme mentionné en introduction.

Dans ce chapitre, nous présentons les principaux travaux de recherche liés à la production de données solaires, à l'évaluation de ces données et les modèles d'énergie solaire élaborés. Mais avant, nous présenterons de façon générale la disponibilité de l'énergie solaire, quelques aspects de l'utilisation de cette énergie et les enjeux. À la fin de ce chapitre nous présenterons brièvement les copules en insistant sur l'utilisation des copules et l'usage que nous en faisons dans l'analyse de la structure de dépendance entre les potentiels sites de production d'énergie solaire.

2.1. Potentiel énergétique solaire

2.1.1. Disponibilité de l'énergie solaire

La forte diminution des ressources fossiles et la rareté des ressources en eau pour la production de l'énergie électrique, amènent certains États à encourager la production de l'électricité à partir de nouvelles sources d'énergie telles que le rayonnement solaire, la biomasse et le vent. Dans cette perspective, Bastien et Athienitis (2011, page 3) font remarquer une certaine abondance de l'énergie solaire et note que : « sur le plan mondial, le solaire est la source d'énergie renouvelable la plus abondante, elle pourrait fournir 2850 fois la consommation mondiale actuelle en énergie ». De plus ces chercheurs mentionnent le fait « qu'il est évident que les centrales électriques, alimentées aux combustibles fossiles ou à l'uranium ou même l'hydroélectricité, fournissent davantage d'électricité par mètre carré que les énergies solaires. Toutefois, les énergies solaires comportent des avantages énormes. Elles peuvent compenser une plus faible densité énergétique. Elles peuvent être installées en production, distribuées géographiquement et être intégrées à divers types d'infrastructures (génie civil, aéronautiques, construction navale et automobile, etc.) ainsi les énergies solaires n'accaparent pas de territoire additionnel ayant comme seule vocation la production d'énergie » (Bastien et Athienitis, 2011, page 7).

Selon cette même étude de Bastien et Athienitis (2011), la disponibilité abondante d'énergie solaire est également visible en Amérique du nord, particulièrement au Canada et au nord des États-Unis. Malgré leur statut de pays nordiques et froids. Les auteurs précisent que : « certaines villes du Canada et même au nord des États-Unis reçoivent presqu'un tiers de plus d'ensoleillement annuellement que Berlin, alors que l'Allemagne est l'un des chefs de file mondiaux de la production d'électricité par l'énergie solaire » (Bastien et Athienitis, 2011, page 8). Pour ces zones géographiques, c'est-à-dire l'Amérique du nord, ils précisent que : « l'énergie solaire n'est pas seulement une énergie renouvelable et propre qui contribue à la lutte contre les changements climatiques; mais elle est aussi une source d'énergie d'appoint

importante qui augmentent le confort, la sécurité et l'autonomie en matière énergétique » (Bastien et Athienitis, 2011).

2.1.2. Utilisation de l'énergie solaire et enjeux

Cette section résumera les avantages du modèle d'affaire dans l'utilisation de l'énergie produite à partir du rayonnement solaire. L'énergie produite à partir du rayonnement solaire direct, en plus de servir au système de chauffage dans les bâtiments industriels ou commerciaux et à la production d'électricité de façon générale, peut avoir diverses autres utilisations qui sont présentées dans les pages qui suivent. Les informations sont essentiellement tirées de l'analyse de Drake et Hubacek (2007), de l'étude de Bastien et Athienitis (2011) et du rapport annuel de Canmet Energy (2012).

Drake et Hubacek (2007) ont démontré dans leur recherche que la combinaison de l'énergie solaire à l'hydroélectricité, à l'éolien, à la géothermie et d'autres sources d'énergie, pourrait améliorer l'efficacité énergétique et la stabilité de la fourniture en électricité, et par la même occasion réduire la dépendance à moyen terme des énergies fossiles. Ces chercheurs ont analysé la réduction de la variabilité de l'énergie éolienne en combinant celle-ci à l'énergie solaire. Selon les chercheurs, les deux sources d'énergies sont relativement non-corrélées entre elles en termes de période de production. On pourrait espérer cela en cas de ralentissement dans la génération de l'énergie éolienne. Dans ce cas, l'énergie solaire pourrait couvrir le déficit et vice-versa. Selon l'article, une telle compensation de la production d'énergie a été observée sur certains sites de production combinée en Turquie.

Bastien et Athienitis (2011, page 10) font remarquer : « qu'il est possible d'intégrer des modules photovoltaïques aux bâtiments de sorte qu'ils composent une partie intégrante de leur enveloppe. Utilisés comme parement extérieur et proprement installés, ils peuvent remplacer le bardeau d'asphalte d'un toit ou le revêtement extérieur d'un mur et remplir les mêmes fonctions afin d'assurer l'écoulement des

eaux ». Et pour des enjeux économique ils notent que : «Le surcoût des panneaux solaires, intégrés aux bâtiments, se retrouve diminué en évitant l'achat de matériaux conventionnels pour la toiture ou le revêtement extérieur. Ils génèrent de l'énergie et protègent l'enveloppe du bâtiment. Il existe plusieurs configurations de capteurs intégrés au bâtiment utilisant des matériaux différents avec des coûts et efficacités qui peuvent varier» (Bastien et Athienitis, 2011, page 11).

Dans l'industrie automobile, la vente de quelques automobiles recouvertes de panneaux photovoltaïques sur le toit ou sur le capot font leur apparition. Seulement, l'énergie solaire captée par ces panneaux ne sert qu'à alimenter la batterie et permettre ainsi à la voiture de parcourir quelques dizaines de kilomètres, de façon autonome, grâce à l'énergie solaire accumulée dans une journée. Selon un rapport de Canmet Energy (2012), au Canada de nouveaux systèmes de chauffage des piscines à l'énergie solaire sont mis en marché. Ce rapport note que : « le chauffage solaire des piscines est une option que les propriétaires de piscine exploitent depuis quelques années. Le coût initial du matériel est raisonnable et les frais d'utilisation sont très avantageux».

2.1.3. Centrales de production d'électricité solaire

Le département énergie du Centre National de Recherche Scientifique de France présente dans un rapport de colloque en 2011 «*L'énergie solaire : PV et concentré* » qu'il existe principalement deux techniques de production de l'énergie solaire. Le rapport présente « la technique de production sans concentration du rayonnement solaire, conçues avant tout pour doter les bâtiments, qu'ils soient résidentiels ou professionnels, de sources d'énergies intégrées (toitures, façades), via des capteurs à plans fixes et la technique de production avec une concentration du rayonnement solaire, via des réflecteurs de formes diverses et mobiles (suivi du soleil), destinées à la génération de puissance électrique (fermes photovoltaïques) ou thermique (chaleur haute température utilisée par des centrales solaires spécifiques ». (*Source : www.energie.cnrs.fr*). Le rendement de production varie selon la technique

de production adoptée. Selon le Centre National de Recherche Scientifique (CNRS) en France, ces variations peuvent s'exprimer comme représentées dans le tableau 2.1.

Tableau 2.1 : Rendement moyen d'une centrale de production de l'énergie solaire sur toute sa période d'exploitation.

	Rendement moyen
Solaire non concentré	
▪ Photovoltaïque	5 – 18%
▪ Thermique	40%
Solaire concentré	
▪ Photovoltaïque	20 – 22%
▪ Thermodynamique	15 – 20%

Source : www.imp.cnrs.fr/energie.

Selon le rapport de Canmet-Energy (2012), la capacité d'énergie solaire photovoltaïque installée dans l'ensemble du Canada en 2011 s'élevait à 289 MW ce qui représente un peu plus de 335 GWh de production sur une base annuelle, soit moins de 1% de la production totale d'électricité au pays. Cette capacité reste faible comparativement aux capacités installées dans quelques pays occidentaux et dans le monde en général. Le tableau 1.2 dénombre ce résultat.

Tableau 1.2 : Capacité installée des centrales photovoltaïques (2012).

Pays	**Allemagne**	**Espagne**	**France**	**Japon**	**USA**	**Monde**
Capacité (MW)	24 880	4 210	2 830	4 700	4 200	69 680

Source : Eurobser (Europe), EPIA 2012 : Objectifs nationaux NREAP (Plan d'Action National pour les Énergies Renouvelables).

Il ressort également des données du tableau 2.2 que la production d'énergie solaire de l'Allemagne représentait en 2011 environ 36% de la production enregistrée au niveau mondial.

2.2. Évaluation des données de l'énergie solaire

Une meilleure évaluation du potentiel de production énergétique solaire d'un site nécessite de disposer de mesures directes du rayonnement solaire. Ces mesures nécessitent une instrumentation adéquate du site afin de collecter sur une certaine période des données sûres. Pour certains sites en Amérique du nord, les données sont mesurées à l'aide des satellites, mais cela nécessite des algorithmes assez complexes pour spécifier les différentes composantes du rayonnement solaire nécessaires à la production de l'électricité ou d'autres formes d'énergies solaire. Compte tenu des coûts élevés d'acquisition des données solaires, il est rare d'obtenir des mesures des sites potentiellement convoités. Également la qualité des données collectées peut être problématique si la période de collecte est courte. Il est nécessaire d'appliquer un contrôle qualité avant l'exploitation des mesures réalisées. Pour ce faire les chercheurs (Younes & *coll*., 2005) ont produit un article qui décrit une procédure de validation des données solaires collectées.

Younes & *coll.* (2005) ont utilisé des données de rayonnement solaire stockées et issues d'une dizaine de sites à travers le monde. Selon ces chercheurs : « la procédure de validation est basée sur la création d'une enveloppe incluant l'index de clarté ou diffus du ciel pour l'évaluation physique et statistique des données » (Younes & *coll.* 2005). Les coordonnés définissant l'enveloppe sont enregistrées ce qui permet de fixer les limites d'acceptabilité de l'index de clarté ou diffus. Ces limites d'index viennent impacter le modèle de rayonnement selon que les données soient horaires, journalières, mensuelles ou annuelles.

2.3. Modèles d'énergie solaire

Dans cette section l'analyse des articles s'est faite en deux étapes. La première étape présente le développement des modèles théoriques ou physiques et la deuxième étape présente l'analyse des modèles statistiques incluant la modélisation par les séries chronologiques. Nous présentons vers la fin de cette section un bref aperçu sur les copules et leur nécessité dans notre étude.

2.3.1. Modèles physiques

Piedallu et Gégout (2007) présente dans leurs travaux que l'une des approches empiriques de la modélisation physique des rayonnements solaires est basée sur le modèle de Liu et Jordan (1960). Selon les auteurs, ce modèle tient compte de l'angle solaire et la transmissivité de l'atmosphère (ou opacité de l'atmosphère). La composante directe de l'énergie solaire prend en compte certains paramètres aléatoires tels que l'effet des nuages, les saisons, la végétation et des certains paramètres liés à la topographie des sites pour simuler l'évolution physique horaire de l'énergie. Cette composante directe du rayonnement solaire qui nous intéresse dans le modèle de Liu et Jordan (1960) est définie suivant cette forme :

$$R_{dir} = S_h \; R_{out} \; \tau^{M} \; \cos(i),$$

où (i) représente l'angle d'incidence entre le rayon solaire et la surface du sol. Cet angle tient compte des paramètres topographiques évalués selon la formule de Campbell (Campbell 1981), le cosinus de l'angle (i) se calcule selon la formule :

$$\cos(i) = \cos\alpha \sin\chi \cos(\beta - \beta s) + \sin\alpha \cos\chi.$$

χ est la pente des radiations (degrés), βs représente l'angle d'exposition au rayonnement solaire et S_h est une valeur binaire d'ombrage calculée pour chaque heure de la journée. Les angles (α) et (β) sont respectivement l'angle solaire et l'azimut solaire. La position du soleil dans le ciel est fonction de l'heure et de la latitude. Cette position est définie par deux angles caractérisant l'altitude et l'azimut solaire :

$$\sin\alpha = \sin\phi \; \sin\delta + \cos\phi \; \cos\eta \; \cos\delta,$$

avec ϕ la latitude pour chaque cellule, η étant l'heure solaire, δ la déclinaison solaire, qui varie en fonction du jour julien J :

$$\delta = 23.45 \; \sin(360\,(284 + J)/365),$$

l'azimut solaire (β) est l'angle entre le soleil et le nord. Nous avons utilisé la formule d'Oke (Oke, 1987) :

$$\cos \beta = (\sin \delta \cos \phi - \cos \delta \sin \phi \cos \eta) / \cos \alpha.$$

Le calcul du flux solaire à la sortie de l'atmosphère (R_{out}, W/m²) est évalué avec le modèle de Kreith et Kreider (1978). Selon ces auteurs, ce flux solaire est fonction de la constante solaire S_c (fournie par World Radiation Center, 1367 W/m²), et le jour de l'année (J) :

$$R_{out} = S_c \ (1 + 0.034 \cos (360 \ J/365)).$$

Le coefficient de transmissivité τ^M représente la fraction du rayonnement incident à la surface de l'atmosphère qui atteint le sol le long d'une trajectoire verticale. La combinaison de tous ces paramètres donne un modèle qui suit un processus sinusoïdal.

Iqbal (1983) propose dans son livre un modèle physique de rayonnement solaire direct journalier qui s'exprime comme suit :

$$H_0 = \frac{24 I_{sc}}{\pi} \left[1 + 0.33 \cos \frac{360 \ J}{365} \cos \phi \cos \delta \sin w_s + \frac{2\pi w_s}{360} \sin \phi \sin \delta \right].$$

I_{sc} correspond à la constante solaire exprimé précédemment par S_c, w_s est l'angle horaire moyen de lever du soleil pour un mois donné. Les autres paramètres correspondent à ceux définis un peu plus haut. Cette expression H_0 du rayonnement solaire proposé par Iqbal (1983) nous paraît un peu plus simple à manipuler si les coordonnées du site sont bien définies.

Solanki et Sangani (2007) ont développé un modèle physique de rayonnement solaire direct journalier similaire à celui de Liu et Jordan (1960). Ce modèle s'exprime par la différence de radiations globales et de radiations diffuses journalières, indexé par l'inverse de la position du soleil par rapport au site.

$$I_N = \frac{G_{daily} - D_{daily}}{\sin(\alpha)}.$$

Ce modèle est appelé EAC (*Elevation Angle Constant*). Les paramètres essentiels du modèle sont l'altitude du site, l'angle de lever du soleil, la déclinaison

solaire et la latitude de site. La prise en compte de tels aspects dans le modèle de Solanki et Sangani (2007) s'apparentent aussi à ceux que Liu et Jordan (1960) proposent. Solanki et Sangani ont effectué des mesures de rayonnement solaire sur une douzaine de sites à travers le monde y compris certains dans la province d'Ontario. Ces chercheurs ont démontré dans leur étude que : « L'évaluation du modèle et des observations ont montré que plus de 90% des estimations se sont avérées avec des erreurs de plus ou moins 5% par rapport aux valeurs observées » (Solanki et Sangani, 2007). Il importe de préciser que dans leurs travaux, Solanki et Sangani (2007) analysent l'efficacité de la qualité des modèles physiques de rayonnement solaire, y compris le modèle proposé plus haut par Liu et Jordan.

Piedallu et Gégout (2008) ont évalué l'efficacité de la distribution topographique du rayonnement solaire direct afin d'améliorer l'application du modèle à un site donné. Ils ont comparé l'efficacité de différentes méthodes d'estimation dans l'espace topographique de rayonnement solaire, du plus simple (basé sur la pente solaire et les valeurs transformées en cosinus) au plus élaboré en utilisant un programme de GIS (système d'information géographique) adapté à la présence des nuages dans le ciel pendant les périodes d'ensoleillement. Ils ont défini un index de transmittance qui tient compte de la présence des nuages pour améliorer l'efficacité du modèle théorique. Ils arrivent à la conclusion que la méthode de pente est moins efficace que les calculs de GIS, mais la différence entre ces méthodes diminue sur une échelle locale. Selon les auteurs : « L'utilisation d'une méthode appropriée pour l'évaluation de rayonnement solaire, y compris la présence des nuages et les caractéristiques topographiques, devrait augmenter la précision du modèle en y intégrant les facteurs écologiques ». il précisent également que : « ce processus peut être employé pour calculer certaines variables, comme l'équilibre réel dans l'espace distribué d'évapotranspiration ou de l'eau, qui sont les principaux conducteurs des espèces croissantes et en particulier dans le cadre du changement climatique actuel » (Piedallu et Gégout, 2008).

2.3.2. Modèles stochastiques

Boland (2008) propose un modèle physique du rayonnement solaire dont les paramètres sont définis par une transformation en série de Fourier ce qui permet de représenter la courbe de rayonnement solaire périodique. Cela permet également de spécifier d'une part la composante périodique et cyclique et d'autre part la composante régulière du rayonnement pour une période d'ensoleillement (horaire ou journalière). Cette méthode permet au chercheur de visualiser les déviations du modèle par rapport aux observations. Ensuite à partir du modèle théorique établie, il détermine un modèle de série chronologique par la méthode de Box-Jenkins. Il fait différents tests d'adéquation pour choisir le modèle de série chronologie qui s'ajuste bien à ses données. Les modèles qu'il a établi dans son étude sont soit des processus autorégressifs ou AR, soit des processus autorégressifs à moyenne mobile ou ARMA.

Nowicka-Zagrajek et Weron (2002) ont modélisé et élaboré la prévision de la charge électrique de l'État de la Californie en appliquant à leurs données un processus ARMA. Ils ont montré que les innovations de ce processus présentaient une distribution différente de la Gaussienne. Ces chercheurs ont montré que c'était une distribution logistique qui s'ajustait aux innovations générées, avant de faire des prévisions de la demande en énergie électrique de l'ensemble du réseau de la Californie. Bien que nous soyons ici dans une analyse de planification d'un système électrique, nous pourrons nous inspirer de la première étape de cette étude qui modélise l'énergie solaire d'un site donné.

Azami et *coll.* (2009) abondent dans le même sens que Boland (2008). Ils proposent une analyse par modèle de série chronologique du rayonnement solaire dans les régions tropicales. L'analyse que ces chercheurs proposent porte sur des observations compilées, sur le site de Bangui en Malaisie, sur une période de six (6) mois. Ils sont arrivés à un modèle de processus ARMA (1, 0). Leur approche semble nous apprendre qu'il est judicieux de modéliser l'énergie solaire par un processus simple tel que l'ARMA.

Le modèle stochastique développé par Lauret et Boland (2010) se démarque des modèles précédents. Les chercheurs utilisent un modèle théorique à partir d'une fonction logistique dont les paramètres dépendent essentiellement des conditions atmosphériques. Ensuite, ils appliquent la méthode statistique basée sur la déduction bayésienne. Les chercheurs utilisent des paramètres d'ajustement de la distribution Gaussienne dans le modèle statistique pour montrer le biais des erreurs moyennes entre les données théoriques et celles observées sur un site. L'une des différences majeures entre la méthode bayésienne et la méthode classique est que la déduction bayésienne offre un cadre de mise à jour continuelle des données antérieures. En d'autres termes, toutes les données antérieures ne sont pas perdues et les paramètres peuvent être employés en tant qu'information préalable pour évaluer les données à venir, ce qui réduit les biais dans le modèle stochastique de rayonnement solaire. Bien qu'avantageux dans la réduction du biais dans le modèle, cette méthode demande d'établir des probabilités sur les conditions atmosphériques tels que la présence des nuages afin de définir le « *clearness index,* (k_d) ». Nous n'allons pas utiliser ce processus dans l'établissement de nos modèles afin d'éviter la complexité dans notre démarche (définition des fonctions de densité conditionnelle). Dans la revue de littérature, les auteurs se sont limités jusqu'ici soit à la modélisation stochastique du rayonnement solaire d'un site à la fois, soit d'un système électrique d'un État donné. Or, avec plusieurs centrales ou l'estimation du potentiel de production d'énergie solaire de plusieurs sites, il est pertinent d'établir une structure de dépendance afin de connaître les productions conjointes des sites. L'un des objectifs de notre étude est donc de répondre à cet aspect qui constitue un jalon important de la mesure de risque conjoint des sites dépendants.

2.4. Copules

2.4.1. Généralité sur les copules

Le concept de copule fut introduit pour la première fois par Sklar en 1959. Son but était de résoudre un problème de probabilité énoncé par Maurice Fréchet, dans le

cadre de ses travaux avec Berthold Schweizer sur les espaces métriques aléatoires. Mais c'est au milieu des années 1980 qu'une étude approfondie fut réalisée par Christian Genest et ses collaborateurs sur les copules. Quelques années plutard, Genest et Mackay (1986) décrivent une classe de distributions bivariées, particulièrement facile et simple à manipuler, appelées les copules archimédiennes. Plusieurs développements ont par la suite été effectués par Genest et Rémillard (2004), Genest et Fabre (2007) et Genest et *coll.*, (2009). Ces dernières années, les copules sont devenues un outil standard couramment rencontré dans la littérature concernant l'étude de la structure de dépendance des produits financiers et également dans l'évaluation des risques dans les secteurs de l'assurance. Il semble donc normal que la modélisation de la dépendance à l'aide de copules soit aujourd'hui très fréquente. Mais, l'utilisation de copules est cependant difficile dans des dimensions supérieures à deux. En plus de l'analyse de la structure de dépendance bivariée, nous sommes intéressés dans notre étude à utiliser les copules pour analyser la structure de dépendance pour une dimension supérieure à deux. Cet aspect n'a pas encore étudié selon les travaux réalisés dans l'estimation et la modélisation des données solaires. Dans cette analyse précise de la structure de dépendance de dimension supérieure à deux et pour notre contribution dans le domaine de la production de l'énergie solaire, nous nous inspirons des travaux de Kramer et Schepsmeier (2011). Cette étude (Kramer et Schepsmeier, 2011, Introduction des copules en vigne ou Vine-Copulas) permet l'utilisation des copules bivariées pour construire la structure de dépendance entre les données d'une dimension supérieure à deux.

2.4.2. Copules en vigne (ou Vine-Copulas)

Comme mentionné précédemment, l'introduction de copules en vigne ou Vine-Copulas a été faite par Kramer et Schepsmeier (2011). Brechmann et Schepsmeier (2013), quant à eux proposent une approche flexible dans la modélisation des dépendances complexes pour des distributions multivariées. Ces chercheurs et bien d'autres tels qu'Allen & *coll.* (2013), ont mis en relief les limites de la copule multivariée classique c'est-à-dire la normale multivariée (ou gaussienne) utilisée

fréquemment et qui peut être très restrictive, car cette copule ne tient pas souvent compte des caractéristiques telle que l'asymétrie. Brechmann et Schepsmeier (2013) note que : « la Vine-Copulas vient pallier les limitations et permettre la modélisation des dépendances complexes tout en bénéficiant de la richesse des copules bivariées ». Ils ont développé des outils permettant une analyse exploratoire des données à plusieurs variables et une sélection de copules bivariées ainsi que pour la sélection des familles de copules en vigne ou Vine-Copulas. Dans la méthode proposée par Brechmann et Schepsmeier (2013), les modèles peuvent être estimés de manière séquentielle ou par maximum de vraisemblance conjoint. Ils ont mis au point des algorithmes d'échantillonnage et les méthodes graphiques pour visualiser les regroupements de copules bivariées sous forme d'arbres.

2.5. Synthèse

À travers la revue de littérature, deux principaux constats s'imposent : le premier constat concerne la détermination des modèles physiques de rayonnement solaire ou d'énergie solaire, les études antérieures fournissent suffisamment d'éléments pour établir notre modèle, nous nous référerons principalement aux travaux de Liu et Jordan (1960) et aux calculs du flux solaire proposés par Kreith et Kreider (1978) et Oke (1987). Le deuxième constat est que la modélisation, du rayonnement solaire, en série chronologique est assez récente et met en évidence que les modèles développés concernent un site à la fois. Ceci permet une analyse du risque de production énergétique à partie de l'énergie solaire, un aspect n'ayant pas été clairement mis en évidence dans les travaux de Boland (2008) et Azami & *coll*. (2009). Notre contribution sera d'une part de présenter une analyse de la variabilité de la production d'énergie solaire des sites par la détermination de modèles de séries chronologiques s'inspirant des travaux de Boland (2008) et Azami & *coll.* (2009). D'autre part, le fait que nous travaillons sur plusieurs sites (portefeuille de sites), nous analyserons la structure de dépendance spatiale des sites en modélisant la dépendance spatiale à partir des résidus des sites. Nous effectuerons précisément une comparaison entre le modèle de normale multivariée et un modèle R-Vine plus flexible afin d'évaluer s'il vaut la

peine d'utiliser des modèles plus complexes. Cet aspect de structure de dépendance nous conduira à ressortir les risques conjoints de production énergétique pour un ensemble de sites liés. Il est intéressant de noter à ce propos que cette étape de notre étude n'a pas encore été abordée dans les études et publications que nous avons analysées.

Chapitre 3 : Modèles de séries chronologiques univariées

Comme nous l'avons mentionné en introduction, au chapitre 1, la mesure du risque de production d'énergie solaire d'un site donné, dépend de l'évaluation de la variabilité dans la production. Cette variabilité peut dépendre des innovations d'un processus stochastique ajusté aux données solaires de ce site. Dans ce chapitre, nous allons ajuster un modèle de séries chronologiques aux données de rayonnement solaire des sites. Ces processus séries chronologiques permettent d'estimer des résidus nécessaires pour la suite de notre analyse. Mais avant, il faut :

1. Définir une méthodologie pour expliquer le choix du processus série chronologique pour notre étude. Et présenter brièvement les séries chronologiques et leurs propriétés essentielles pour notre étude.

2. Expliquer le processus de construction d'un modèle physique de rayonnement solaire. Ces données théoriques nous permettent de capturer une certaine saisonnalité dans les données.

3. Identifier clairement notre source des données de rayonnement solaire mesuré, en faire une description précise, présenter les différents sites auxquels elles sont associées et spécifier le choix des sites pour notre étude.

4. Expliquer les transformations effectuées sur les données réelles versus les données générées à partir du modèle physique.

5. Et enfin appliquer aux données transformées le processus série chronologique choisi.

Une fois, la série chronologique estimée et ses paramètres bien spécifiés, les résidus seront testés afin de s'assurer qu'ils respectent les conditions d'un processus « bruit blanc ». Pour capturer les effets saisonniers dans nos données, nous utilisons un modèle physique de rayonnement solaire plutôt qu'une approche par les séries de Fourier comme le propose Boland (2008).

Ce chapitre est donc organisé en trois sections. La première section présente la méthodologie pour construire les modèles de série chronologique et le modèle physique. La deuxième section présente les données de rayonnement solaire mesurées, leurs sources ainsi que les sites auxquels elles sont associées. La dernière section présente les opérations sur les données, l'ajustement du processus stochastique et l'analyse des résultats.

3.1. Méthodologie

3.1.1. Définition et propriétés des séries chronologiques

La définition et les propriétés qui suivent sont tirées de Hamilton (1954). Nous les présentons sous une forme simplifiée.

Définition

Une série chronologique (ou série temporelle) est une succession d'observations au cours du temps : $\{x_t : t = 1, 2, \ldots, n, \ldots\} = (x_1, x_2, \ldots, x_n, \ldots)$. Par rapport aux autres types de données statistiques, la particularité des séries chronologiques tient à la présence d'une relation d'antériorité qui ordonne l'ensemble des données. Les temps d'observations sont souvent équidistantes les unes des autres : il existe des séries horaires, journalières, mensuelles, trimestrielles et annuelles, ces séries sont souvent indexée par t avec $t \in \mathbb{Z}$.

Caractéristiques et propriétés

Une série chronologique (processus stochastique ou série temporelle) peut résulter, selon Enders (2010), d'une ou de la combinaison des différentes composantes suivantes :

- Une tendance (ou trend) qui représente l'évolution à long terme de la série, elle traduit aussi le comportement « moyen » de la série.

- Une composante saisonnière (ou saisonnalité) qui correspond à un phénomène qui se répète à intervalles de temps réguliers.

- Une composante résiduelle (innovation, bruit, choc ou résidu) qui correspond à des fluctuations irrégulières, en général de faible intensité mais de nature aléatoire. On parle aussi d'aléas. Ceci peut être engendré par des phénomènes accidentels (exemple : des conditions météorologiques exceptionnelles, un crash financier, une crise pétrolière, etc.).

- Une autre composante a trait au phénomène cyclique, qui est souvent présente en climatologie et en économie (exemple : récession et/ou expansion). Il s'agit d'un phénomène se répétant mais contrairement à la saisonnalité, cela se produit sur des durées qui ne sont pas fixes et généralement plus longues. «Sans informations spécifiques, il est généralement très difficile de dissocier tendance et cycle » (Enders, 2010).

Dans le cas qui nous intéresse, la stationnarité de la série et la propriété « bruit blanc » des innovations sont souhaitables. Il est nécessaire que le modèle de série temporelle soit de type stationnaire (ce qui implique que le comportement de la série ne dépend pas du temps) Enders (2010). On suppose qu'il n'y a ni tendance, ni cycle dans les données, ou que ceux-ci ont un effet négligeable. La composante saisonnière sera retirée à l'aide du modèle physique. Ainsi, il restera à s'assurer que les résidus des

séries estimées sont « bruit blanc », c'est-à-dire une moyenne nulle, une variance constante et non auto corrélée. Par conséquent, la stationnarité de la série et le fait que les résidus (ou chocs) soient « bruit blanc » constituent les deux propriétés essentielles que nous exploiterons.

Choix du processus série chronologique

Comme nous l'avons mentionné dans le chapitre 2 (revue de littérature), des travaux sur la modélisation de l'énergie solaire en série temporelle ont été élaborés par Boland (2008) et Azami & *coll.* (2009). Ces chercheurs ont établi que les processus autorégressifs à moyenne mobile ou ARMA semblent s'ajuster correctement au données de rayonnement solaire. Notre démarche sera donc d'explorer cette voie afin de nous convaincre que nos données peuvent également s'ajuster à ce type de processus stochastique. Si les résultats de nos estimations avec le processus ARMA s'avèrent concluants, il serait donc simple pour nous d'estimer les résidus. Nous choisissons donc le processus ARMA comme point de départ pour estimer nos modèles de série chronologique.

La particularité de ce processus est qu'il ajuste à la fois les composantes autorégressives et les composantes à moyenne mobile (Hamilton, 1994). La stationnarité de ce processus est définie par la composante autorégressive (Jenkins & *coll.*, 2008, pages 51 à 55). Si nous souhaitons vérifier que les résidus de ce processus sont « bruit blanc » il faudrait réaliser le test de McLeod & Li (1983) sur ces résidus.

3.1.2. Construction du modèle physique de rayonnement solaire

À la section 2.3.1, des modèles théoriques de rayonnement solaire ont été présentés. Nous pouvons citer le modèle de Liu et Jordan (1960) qui est composé du rayonnement normal direct et du rayonnement diffus. Ce modèle dépend essentiellement des paramètres tels que : la latitude, l'opacité de l'atmosphère (ou la transmittance) et l'angle d'incidence entre les rayons solaires et la surface du sol. Le

modèle de Solanki et Sangani (2007) s'inspire de celui de Liu et Jordan, à la différence d'indexer le rayonnement solaire journalier à la position du soleil par rapport au site. Quant au modèle de Piedallu et Gégout (2008), il est assez complexe et intègre des paramètres de saison, de topographie et même d'écologie, y compris les paramètres initiaux identifiés par Liu et Jordan (1960).

Le modèle qui retient notre attention est celui de Liu et Jordan (1960). À la base c'est un modèle simple et de forme sinusoïdale. Hormis le paramètre d'opacité de l'atmosphère, ce modèle dépend essentiellement de la position du site par rapport au soleil. Certains paramètres tels que la constante solaire et le flux solaire à la sortie de l'atmosphère dans le modèle sont fournis par Kreith et Kreider (1978). En ce qui concerne l'opacité de l'atmosphère, nous n'avons pas pu les obtenir. Par conséquent nous posons comme hypothèse que les conditions d'opacité atmosphérique son idéale (c'est-à-dire le paramètre d'opacité : la transmittance est égale à 1). La formule générale du modèle physique de rayonnement solaire direct est la suivante :

$$y_x = 1367\left[a_0 + a_1 \cos\left[\frac{\pi(x - t_0)}{12}\right]\right].$$

Nomenclature

y_x	Rayonnement solaire pour une période x de la journée en W/m^2.
Y	Rayonnement solaire moyen journalier, moyenne des y_x en W/m^2.
a_0 et a_1	Composantes de l'angle solaire.
L	Latitude du site.
δ	Déclinaison solaire.
ndays	Nombre de jours dans l'année (dépend si année bissextile ou non).
J	Jour julien.
t_0	Heure solaire vraie (correspond à l'heure au moment du midi solaire).

Les composantes de l'angle solaire sont a_0 et a_1. Ces composantes sont fonction de la latitude (L) des sites et de l'angle de déclinaison (delta) du soleil. L'expression de

l'heure vraie correspondant à l'heure au moment du midi solaire est donnée par t_0, cette heure est fonction de la longitude (l).

$$a_0 = \sin\left(\frac{\pi L}{180}\right)\sin\left(\frac{\pi\delta}{180}\right),\ a_1 = \cos\left(\frac{\pi L}{180}\right)\cos\left(\frac{\pi\delta}{180}\right),\ \delta = 23.45\sin\left[\frac{2\pi(J+284)}{ndays}\right],$$

$$t_0 = 12 + \left[\frac{24(l)}{360}\right],$$

avec J est le jour Julien et « *ndays* » est le nombre de jours dans une année, ce nombre varie selon que l'année est bissextile ou non. L'expression $\frac{\pi(x-t_0)}{12}$ correspond à l'angle horaire du soleil à un temps *x* d'une journée. Ce temps est choisi afin de correspondre au temps des mesures du rayonnement solaire sur le site (section 3.2.1). Le rayonnement solaire théorique journalier moyen pour un site est la moyenne des y_x pour cette journée. À partir de l'estimation de la durée moyenne journalière d'ensoleillement des sites (Figure 3.2), nous pouvons estimer l'énergie solaire moyenne journalière en (kWh/m^2).

3.2. Données du rayonnement solaire

Dans cette section nous présentons une description des données du rayonnement solaire mesurées que nous avons obtenues pour notre étude. Leurs origines et les différents sites auxquels elles sont liées, vous seront présentés dans les paragraphes suivants.

3.2.1. Origine et caractéristiques des données

Les données d'énergie solaire utilisées dans cette étude sont extraites de Surfrad. *Surfrad (Surface Radiation) Network* est un réseau créé en 1993 par le laboratoire Américain de recherche sur les conditions atmosphériques pour réaliser des mesures et collecter :

- Les données météorologiques telles que la vitesse des vents et la température.

- Les données sur l'imagerie du ciel et le couvert nuageux (*Aerosol Optical Depth*).

- Les données de rayonnements solaires (rayonnement solaire direct et diffus) et les paramètres sur la rotation de la terre autour du soleil.

Ce réseau est constitué de sept stations (ou sites) opérationnelles qui sont répartis dans quelques États Américains. Ces sites sont : Penn-State (État de Pennsylvanie), Bondville (État de l'Illinois), Boulder (État du Colorado), Fort Peck (État du Montana) Goodwin Creek (État du Mississipi), Sioux Fall (État du Sud Dakota) et Desert Rock (État du Nevada). Les informations disponibles sur ce réseau sont gratuites et accessibles au grand public. Les données utiles pour notre étude sont les mesures du rayonnement solaire qui sont compilées et stockées sur ce réseau.

Données et leurs caractéristiques

Les mesures de rayonnement solaire direct constituent les données d'intérêt pour notre étude. Signalons à ce propos que le rayonnement solaire direct constitue la composante du rayonnement solaire total à la surface terrestre capable, après transformation, de produire l'énergie électrique. Nous avons extrait les données de rayonnement solaire direct et les paramètres de la position de la station à partir du site internet de NOAA[2] (*www.esrl.noaa.gov/gmd/grad/surfrad*), sur le réseau *Surfrad*. Les principales caractéristiques de ces données sont :

- Les périodes de mesure complète de rayonnement solaire direct pour la majorité des sites s'étalent de 1995 à 2014 (mesure continue pour les années à venir). En Pennsylvanie les données complètes sont de 1998 à 2013 et au Sud Dakota les données complètes sont de 2003 à 2013.

[2] NOAA : National Oceanic & Atmospheric Administration (U.S. Department of Commerce)

Tous les sites n'ont pas été opérationnels dans les mêmes périodes. Nous avons alors opté pour un niveau homogène de période pour les données mesurées par site. Nous avons choisi pour notre étude les mesures du rayonnement solaire allant de 2007 à 2011, c'est-à-dire une période de cinq (5) ans.

- L'examen des données compilées par site présente : une mesure journalière avec une fréquence variant selon les années. Ainsi, de 2007 à 2008 la fréquence est d'une mesure aux 3 minutes et d'une mesure à la minute de 2009 à 2011.

Pour uniformiser les fréquences des mesures, nous avons effectué des opérations en divisant par 3 les mesures réalisées en 1 minute. Nous avons également choisi de travailler avec des mesures de rayonnement solaire moyen par jour. Connaissant la durée moyenne d'ensoleillement de chaque zone (Figure 3.2), nous avons estimé en (kWh/m^2) l'énergie journalière moyenne pour chaque site.

Cartographie des sites

Les cartes de la figure 3.1 et de la figure 3.2 présentent respectivement les stations (ou sites) du réseau Surfrad et les durées moyennes d'ensoleillement journalière sur le territoire Américain. Le tableau 3.1 présente la liste des sites et les États concernés.

Tableau 3.1 : Liste des sites du réseau Surfrad sur l'ensemble du territoire des États-Unis d'Amérique.

Ordre	Sites	États
1	Penn State	Pennsylvanie
2	Bondville	Illinois
3	Boulder	Colorado
4	Fort Peck	Montana
5	Goodwin Creek	Mississipi
6	Sioux Fall	Sud Dakota
7	Desert Rock	Nevada

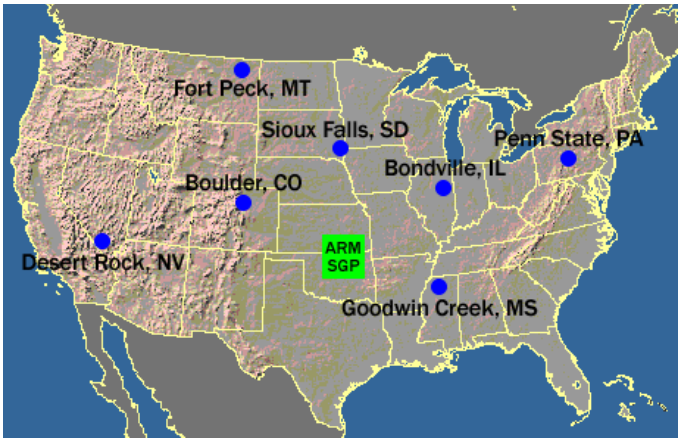

Figure 3.1 : Situations géographiques des sites du réseau Surfrad sur le territoire des États-Unis d'Amérique. *Source : (www.esrl.noaa.gov).*

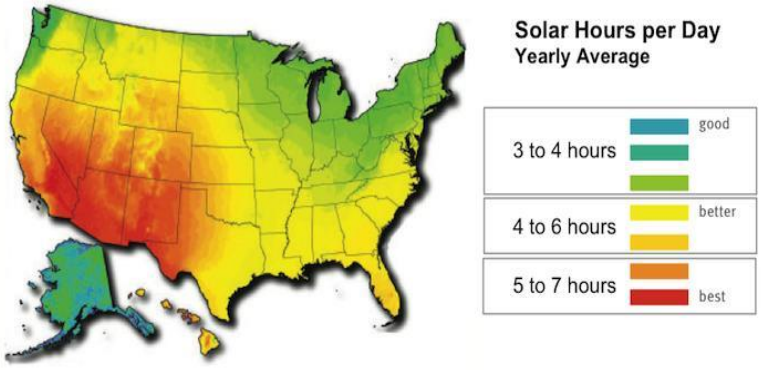

Figure 3.2 : Heures solaires par jour sur le territoire des États-Unis d'Amérique. *Source : (www.freecleansolar.com).*

3.2.2. Choix des sites

Après analyse des positions géographiques et des données des sept sites qui constituent le réseau Surfrad, nous en avons retenu six pour notre étude. Étant donné que Desert Rock au Nevada est situé dans une zone montagneuse différente des six autres sites, nous ne l'avons pas inclus. Malgré la distance géographique de ce site par rapport aux autres sites du réseau Surfrad (voir tableau A.2 à l'annexe A.2), le relief pourrait influencer une certaine dépendance du site de Desert Rock avec les autres sites. Les six sites choisis pour l'étude sont les suivants : Penn-State (Pennsylvanie),

Bondville (Illinois), Boulder (Colorado), Fort Peck (Montana), Goodwin Creek (Mississipi) et Sioux Fall (Sud Dakota). Nous pouvons observer les cordonnées de ces sites dans le tableau A.1 en annexe.

3.3. Application du processus ARMA

3.3.1. Transformation sur les données

Un ajustement efficace de processus stochastique stationnaire nécessite des séries qui ne présentent aucune tendance et aucun effet saisonnier. Du fait que les données de rayonnement solaire mesurées par site soient des données brutes, nous avons fait des transformations. Nous avons calculé un ratio en divisant les valeurs mesurées du rayonnement solaire et les valeurs du modèle théorique (qui capture les tendances). Cette première transformation permet de réduire les effets saisonniers existant. Ensuite nous appliquons une deuxième transformation en calculant l'inverse du logarithme de un moins la valeur du ratio (- log [1-ratio]). Cette deuxième transformation permet l'obtention de valeurs symétriques. La deuxième transformation est appelée *«log ratios »* dans le programme R élaboré. L'estimation du processus ARMA se fait donc avec les données « *log ratios* ». les données de chaque site choisi ont été transformées par le même processus.

3.3.2. Construction du processus ARMA et test de spécification

Construction du processus ARMA (p, q)

Il convient de rappeler l'expression analytique du modèle ARMA (p, q). Elle se présente sous la forme : (x_t, t∈Z), $x_t = \delta + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$, avec ϵ_t i.i.d ~ N (0, σ_ϵ^2). Pour estimer les processus stochastiques stationnaires ARMA (p, q), nous utilisons une fonction générale « *auto.arima* » du « package

(*forecast*) » dans le logiciel R. Cette fonction a été développée par Hyndman et Khandakar (2008) et utilise le maximum de vraisemblance pour estimer les paramètres du processus. L'application de la fonction « *auto.arima* » retourne le meilleur modèle ARIMA (p, d, q) basé sur les critères d'information d'Akaike (AIC) ou Bayésien (BIC) pour la qualité de l'ajustement. La fonction effectue une recherche sur les modèles possibles dans la limite des paramètres fournis. Si le modèle est stationnaire sans différentiation, la fonction retourne une valeur du paramètre d = 0. Sinon, elle effectue un test de racine unitaire pour déterminer la valeur de d. Le test de racine unitaire est spécifié selon trois critères au choix avec la fonction « *auto.arima* ». Soit le critère Kwiatkowski – Phillips – Schmidt – Shin « *kpss* » (Kwiatkowski & *coll.*, 1992), soit celui de Dickey – Fuller augmenté « *adf* » (Dickey & Fuller, 1979), ou soit le critère de Phillips – Perron « *pp* ». Avec les transformations effectuées sur nos données, il est possible de constater une valeur de d nulle donc l'absence de racine unitaire dans nos modèles. Les autres paramètres dont nous avons tenu compte pour définir nos modèles sont contenus dans la nomenclature à la page suivante.

Pour le test de la racine unitaire, nous avons choisi le test de Dickey – Fuller augmenté pour évaluer l'existence d'une racine unitaire dans les modèles ARIMA. C'est un test simple et couramment utilisé pour connaître le nombre de différentiation avant qu'un processus soit stationnaire. La fonction « *auto.arima* » nécessite également de connaître les paramètres de retard p et q avant d'estimer les modèles ARIMA. Comment obtenons-nous les valeurs de ces paramètres de retard? Pour chaque site, nous examinons la fonction d'autocorrélation appliquée aux données « *log ratios* » pour le choix du retard maximum q et celle d'autocorrélation partielle pour le choix du nombre maximum de retard p.

Selon les graphiques des fonctions d'autocorrélation et d'autocorrélation partielle (corrélogramme ou corrélogramme partiel) les lignes verticales montrent la corrélation entre les valeurs journalière de « *log ratios* » et une version décalée de ces mêmes valeurs. Les barres horizontales indiquent les valeurs critiques pour un test

supposant que ces corrélations (ou corrélations partielles) sont nulles. Si l'autocorrélation est présente les lignes verticales dépassent les barres horizontales sur le corrélogramme (corrélogramme partiel). Le nombre maximum de p ou de q sera choisi en fonction du dernier chiffre sur l'axe des abscisses où une barre verticale dépasse les barres horizontales significativement avant qu'il soit constaté une tendance vers les limites horizontales à long terme. Une fois le nombre maximum de retard p et celui de q identifiés et les autres critères du modèle spécifiés, nous pouvons estimer le modèle à partir de la fonction, du code R, suivante :

library (forecast)
{Model <- auto.arima (data, stepwise = TRUE, trace = TRUE, ic, test, max.p, max.q)}.

Légende :

data	Données pour modéliser le processus (« log ratios » pour l'étude).
stepwise	Si « TRUE », sélection rapide des étapes.
trace	Si « TRUE », affichage des modèles ARIMA possibles.
ic	Test des critères d'information. Sous forme de vecteur.
test	Test d'hypothèse pour la racine unitaire (Dickey-Fuller).
max.p	Nombre maximum de retard p à déterminer (fonction d'autocorrélation partielle).
max.q	Nombre maximum de retard q à déterminer (fonction d'autocorrélation).

Test de spécification des innovations

Une fois le modèle de séries chronologiques déterminé, un test de spécification sur les innovations est fait pour s'assurer que les résidus suivent un processus « bruit blanc ». Le test de McLeod & Li (1983), appliqué aux résidus, permet de valider si oui ou non les innovations suivent un processus « bruit blanc ». Nous utilisons une fonction développée par McLeod & Li (1983) pour réaliser ce test sur les innovations des processus ARMA. Cette fonction retourne un p-value. Si le p-value est supérieur à

5%, on ne peut pas rejeter l'hypothèse nulle « H_0 : innovation bruit blanc ». La fonction de McLeod & Li (1983) se présente de la façon suivante :

library (FitAR) {LjungBoxTest (DatInnov, k = 0, lag.max =< 30, StartLag = 1)}.

Légende :

DatInnov	Résidus après avoir estimé le modèle.
Lag.max	Nombre maximum de Lag (retard) à utiliser (30 maximum).
StartLag	Initialiser le début à 1.

3.3.3. Modèles et interprétation des résultats

Les sites retenus pour l'étude sont : Penn-State (Pennsylvanie), Bondville (Illinois), Boulder (Colorado), Fort Peck (Montana), Goodwin Creek (Mississipi) et Sioux Fall (Sud Dakota). Pour chaque site, nous construisons un modèle physique d'énergie solaire journalière que nous comparons aux données réelles brutes de l'énergie solaire mesurée. Nous convertissons ces énergies en (kWh/m^2). Le modèle physique de rayonnement solaire journalier est estimé en fonction des coordonnées du site (paramètres a_0 et a_1) et son heure solaire (section 3.1.2). Le tableau 3.2 suivant contient les heures solaires calculées par site et un exemple des valeurs des paramètres a_0 et a_1. Il faut rappeler la formule théorique du rayonnement solaire pour une période de temps *x* de la journée :

$$y_x = 1367\left[a_0 + a_1 \cos\left[\frac{\pi(x - t_0)}{12}\right]\right].$$

Tableau 3.2 : Heures solaires (t_0), paramètres a_0 et a_1 pour une déclinaison solaire correspondant à ($\delta = 2.29$) au jour julien 160.

Sites	**Situation**	a_0	a_1	t_0 **(heure)**
Penn-State	Pennsylvanie	0.25	0.69	17.20
Bondville	Illinois	0.25	0.70	17.89
Boulder	Colorado	0.25	0.71	19.00
Fort Peck	Montana	0.29	0.61	19.00
Goodwin Creek	Mississipi	0.22	0.76	17.99
Sioux Fall	Sud Dakota	0.27	0.67	18.44

Les paramètres a_0 et a_1 dépendent de la déclinaison solaire delta (δ) et de la latitude des sites (voir les formules à la page 24). Si nous considérons par exemple le $160^{\text{ième}}$ jour d'une année de 365 jours, nous estimons $\delta = 2.29$. À partir de cette valeur de δ, nous pouvons calculer les paramètres a_0 et a_1 contenus dans le tableau 3.2. Nous observons également dans le tableau B.1 en annexe B.1 ces mêmes valeurs pour les latitudes des sites. Les heures solaires calculées et contenues dans le tableau 3.2 sont exprimées en fonction de la longitude des sites mentionnés.

Le rayonnement solaire théorique par jour est calculé en prenant la moyenne des rayonnements solaires théoriques pour chaque période de temps *x* de la journée. Les mesures réelles ont été faites à des fréquences de 1 minute à 3 minutes, nous avons ensuite uniformisé ces fréquences de mesure à 3 minutes (voir section 3.2.1). Pour être conforme dans le calcul théorique, nous avons fait coïncider les périodes de calcul avec les périodes de mesure sur les sites, c'est-à-dire à chaque période de 3 minutes. Les graphiques du modèle physique par site versus les données mesurées montrent que les deux courbes ont la même allure à la différence du modèle physique qui est lisse et qui présente un écart en amplitude par rapport aux données mesurées (figure 3.3 et suivantes). Cet écart est dû au fait que nous avons posé comme hypothèse pour le modèle physique les conditions idéales de l'opacité de l'atmosphère (transmittance = 1), ce qui implique un ciel dans atmosphère. Les ordres de grandeur des valeurs

journalières de l'énergie solaire calculées ne sont pas très loin de la répartition de l'énergie selon les zones géographiques à la figure B.2 de l'annexe B.2 à la page 110.

Nous présentons ensuite le graphique des données «*log ratios* » pour chaque site. Nous rappelons que « *log ratios* » est obtenu à partir des transformations sur les données du modèle physique et celles des mesures sur le site. Nous constatons une stationnarité dans l'évolution graphique de ces données et cela pour tous les sites. On retrouve ces figures dans les pages suivantes. Pour estimer les processus autorégressif à moyenne mobile (ARMA), nous produisons les fonctions d'autocorrélation et d'autocorrélation partielle à partir des données « *log ratios* ». Sur les corrélogrammes, nous identifions respectivement le nombre maximum des retards q et p. Ainsi nous estimons le modèle ARMA par site en appliquant la fonction « *auto.arima* » aux données « *log ratios* » avec tous les éléments de la fonction définis.

Une fois le modèle ARMA estimé par site, les résidus sont ensuite testés pour vérifier leurs caractéristiques de processus « bruit blanc » en appliquant le test de McLeod et Li (1983) qui utilise la fonction « *LjungBoxTest* ». Nous constatons que pour les six sites étudiés, les p-value retournés par ce test sont supérieurs au seuil de 5% prédéfinit. Le graphique d'autocorrélation des résidus semble nous informer qu'il n'existe pas de corrélation entre les résidus et leur courbe montre une certaine stationnarité avec une moyenne nulle. Ce constat est fait pour les modèles de chacun des six sites étudiés.

Dans les applications qui suivent, la période des estimations et des mesures du rayonnement solaire par site est choisie de 2007 à 2011 ce qui correspondant à 5 années de données. Sur l'axe des abscisses des graphiques qui suivent, en dehors des corrélogrammes, la variable représentée correspond aux années allant de 2007 à la fin 2011.

3.3.4. Site de Penn-State (Pennsylvanie)

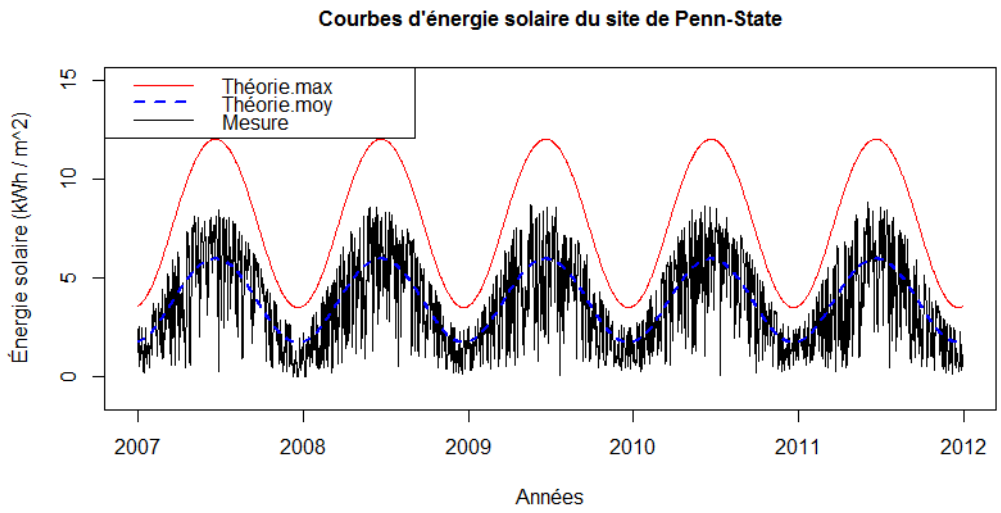

Figure 3.3 : Courbes des modèles théoriques et des données mesurées du site de Penn-State.

La figure 3.3 fournit l'allure de modèles physiques et des données mesurées du rayonnement solaire. Les trois courbes ont une évolution identique en dehors de l'amplitude élevée du modèle théorique maximum où nous supposons un ciel sans atmosphère. La courbe théorique moyenne (courbe bleue pointillée) est estimée en multipliant les valeurs générées par le modèle théorique maximum par un facteur moyen de réduction. Ce facteur est calculé à partir de la formule : 1- exp (-moyenne (*log ratios*)) pour le site de Penn State et les autres sites (le facteur multiplicateur est estimé à partir de cette formule et il est contenu dans le tableau B.2 en annexe). La courbe théorique moyenne semble illustrer, comparativement à la courbe théorique maximum, une tendance centrale. Cette courbe théorique moyenne semble se trouver au centre de la variation des vraies données mais avec une allure de la courbe théorique maximum. La figure 3.4 présente quant à elle, l'allure du « *log ratios* » des données transformées du site de Penn State. Visuellement, nous observons une certaine stationnarité de la transformation des données.

Avec « *log ratios* » les fonctions d'autocorrélation et d'autocorrélation partielle sont estimées pour choisir le nombre de retard p et q du modèle ARMA à estimer.

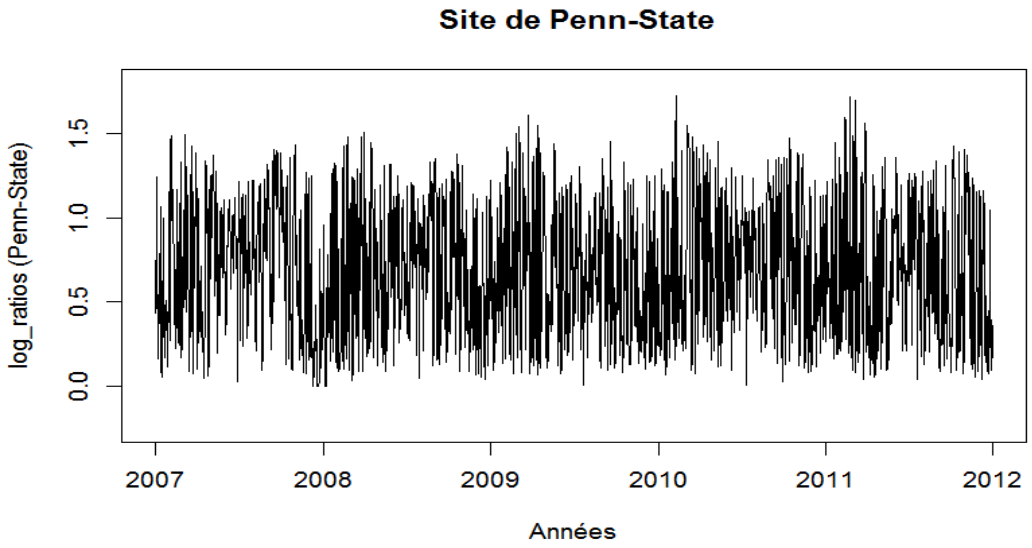

Figure 3.4 : Représentation des données de Penn-State transformées en « *log ratios* ».

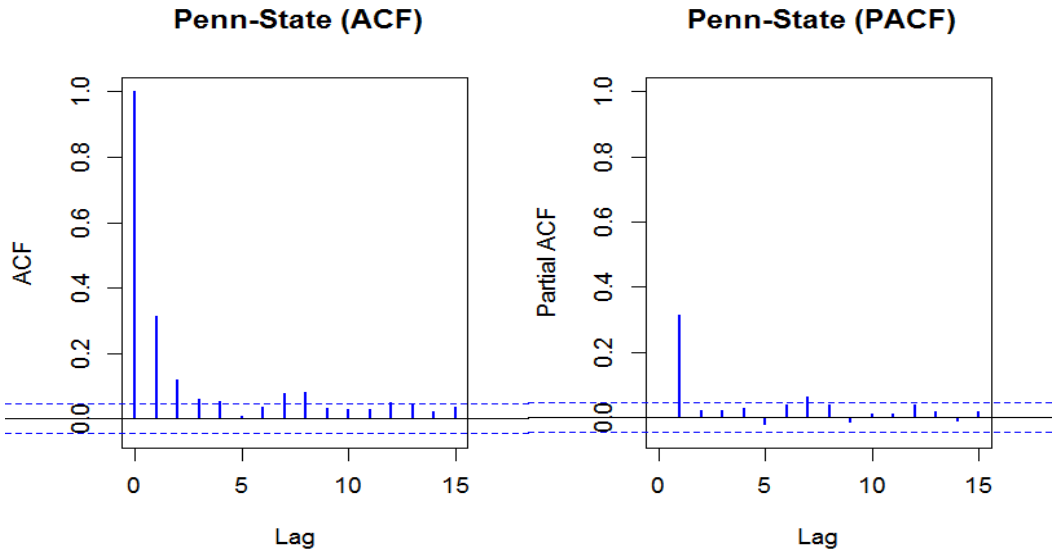

Figure 3.5 : Fonctions d'autocorrélation et d'autocorrélation partielle des données « *log ratios* » de Penn-State.

Comme expliqué à la section 3.3.2, Le choix de retard maximum p est lié au corrélogramme partiel. En observant les lignes verticales à la figure 3.5, nous pouvons fixer max.p à 7. Il en est de même pour le corrélogramme auquel est lié le choix du retard maximum q. Nous fixons également max.q à 8.

En appliquant la fonction de Hyndman et Khandakar (2008) aux données « *log ratios* » du site de Penn-State et en précisant les autres éléments de la fonction « *auto.arima* », les processus contenus dans le tableau 3.3 sont générés avec les paramètres et les valeurs du test des critères d'information d'Akaike AIC et Bayésien BIC.

Tableau 3.3 : AIC et BIC des processus ARIMA générés à partir des données « *log ratios* » du site de Penn-State

Modèle	**Critères d'information**	
	AIC	**BIC**
ARIMA (1, 0, 1)	1683.0	1705.0
ARIMA (1, 0, 0)	1681.8	1698.4
ARIMA (0, 0, 1)	1702.9	1719.5

Au tableau 3.3, il semble que le modèle de processus ARIMA (1, 0, 0), s'ajuste le mieux aux données « *log ratios* » du site de Penn-State selon les critères d'information AIC et BIC. Ce modèle a une moyenne non nulle et correspond aussi à un processus ARMA (1, 0) avec p =1 et q = 0. Le tableau 3.4 suivant contient l'ensemble des paramètres estimés du modèle ARMA (1, 0).

Tableau 3.4 : Paramètres du modèle ARMA (1, 0) estimé pour Penn-State ainsi que les critères d'information et le logarithme du maximum de vraisemblance.

Coefficients		**Erreur Standard (s.e)**
φ	0.3142	0.0222
c	0.6870	0.0131

σ^2	0.1465	**BIC**	1698.4
Log(L)	-837.6	**AIC**	1681.8

Le logarithme du maximum de vraisemblance de la fonction de densité du processus estimé est Log(L). Le coefficient *c* représente la valeur espérée du processus (moyenne). Le modèle centré est représenté par la première équation et en substituant les valeurs du tableau 3.4, nous avons l'expression analytique du ARMA (1, 0) à la deuxième équation.

$$x_t = (1 - \varphi) * c + \varphi * x_{t-1} + \epsilon_t.$$
$$x_{t(Penn)} = 0.47 + 0.3142 x_{t-1(Penn)} + rpenn_t,$$

où lorsque nous faisons l'analogie avec l'expression analytique d'un processus ARMA(1, 0), $rpenn_t$ correspond aux résidus du processus ARMA (1, 0) défini pour le site de Penn State à la période t avec variance ($rpenn_t$) = 0.1465. Vérifions si les résidus du modèle estimé du site de Penn-State sont « bruit blanc ». En appliquant la fonction de McLeod et Li (1983) décrite précédemment, nous avons les résultats consignés dans le tableau 3.5. Le p-value est de 11% ce qui est supérieur au seuil de 5%. À la figure 3.6 nous observons une moyenne de 0 et la fonction d'autocorrélation des résidus de ARMA (1, 0).

Tableau 3.5 : Moyenne, écart type et variance des résidus du site de Penn-State, ainsi que le p-value du test de McLeod & Li (1983).

Sites	**Résidus**	**Moyenne**	**RMSE**	**var**	**p-value**
Penn State	*rpenn*	0	0.3828	0.1465	0.11

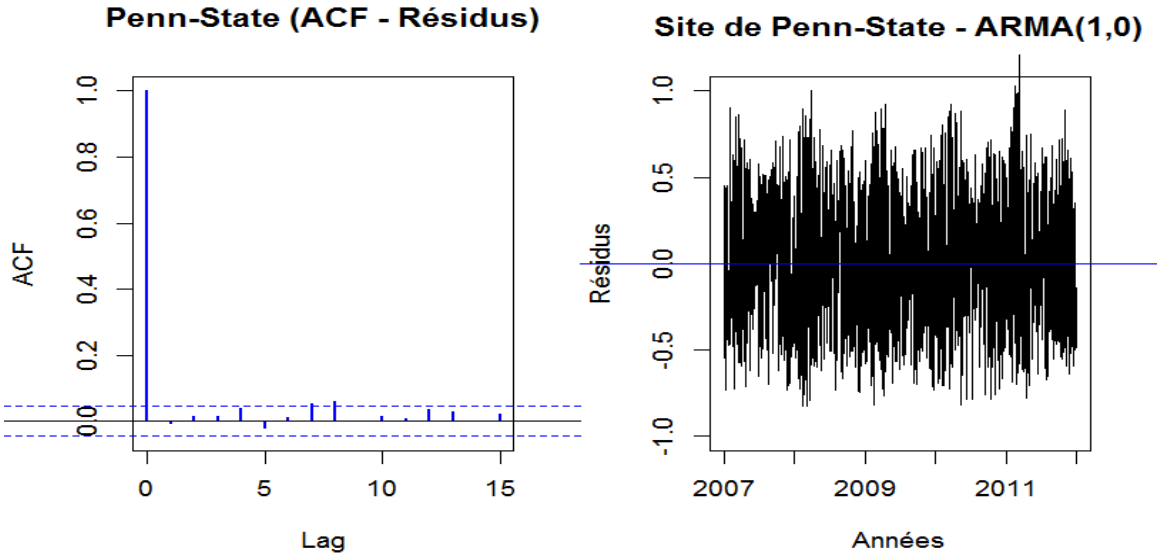

Figure 3.6 : Fonction d'autocorrélation et représentation des résidus du processus ARMA (1, 0) du site de Penn-State.

L'hypothèse des résidus « bruit blanc » du modèle ARMA (1, 0) du site de Penn-State semble se consolider compte tenu des résultats obtenus après le test de McLeod & Li (1983). Nous reprenons le même cheminement pour les autres sites, c'est-à-dire Bondville, Boulder, Fort Peck, Goodwin Creek et Sioux Fall.

3.3.5. Site de Bondville (Illinois)

La même démarche est appliquée pour les autres sites en suivant les différentes étapes élaborées précédemment pour le site de Penn State. Nous observons à la figure 3.7 l'allure des courbes du modèle théorique et des données mesurées sur le site. Les mêmes commentaires de la figure 3.3 s'appliquent à la figure 3.7. La figure 3.8 présente une certaine stationnarité de la série de données « *log ratios* » du site de Bondville.

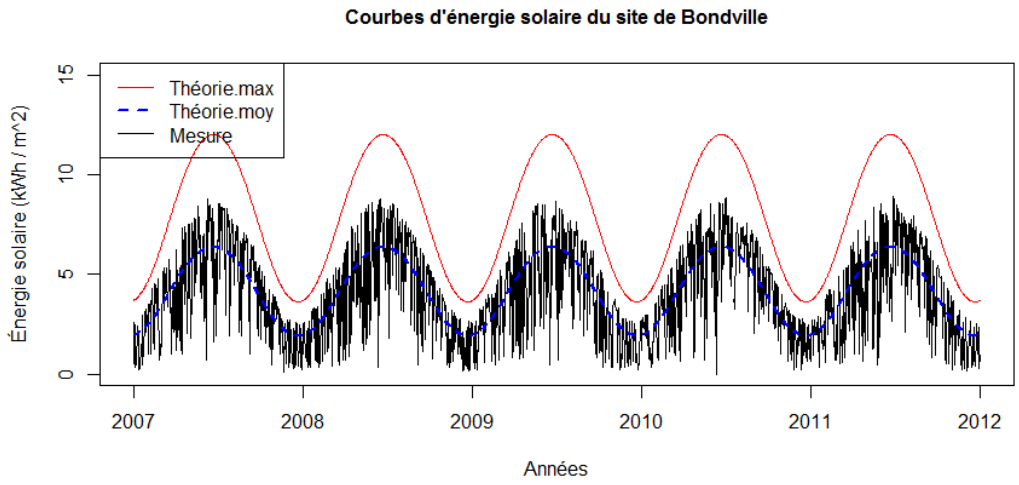

Figure 3.7 : Courbes des modèles théoriques et des données mesurées du site de Bondville.

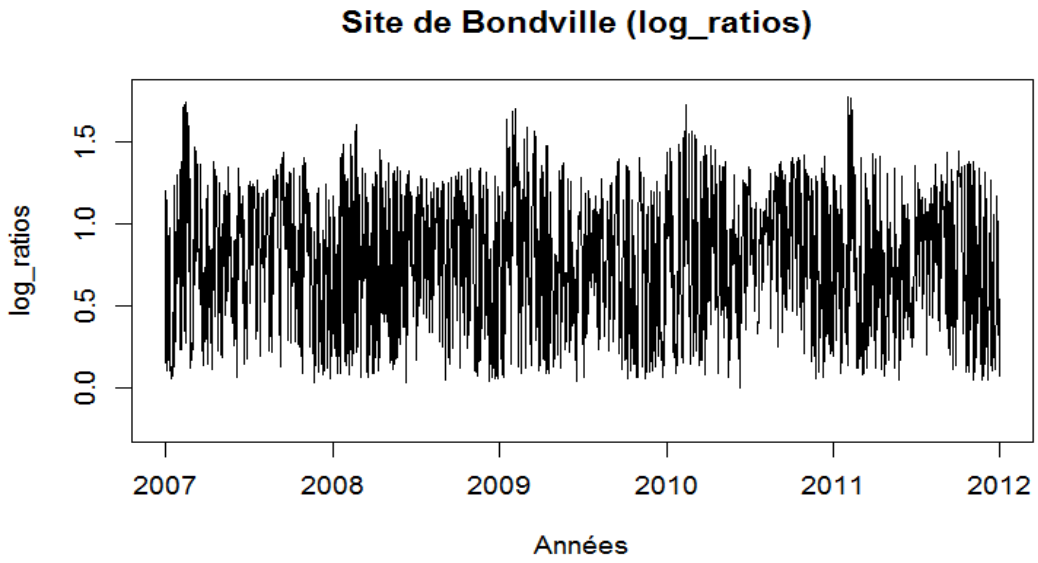

Figure 3.8 : Représentation des données de Bondville transformées en « *log ratios* ».

Les corrélogrammes de la figure qui suit permettent de choisir le nombre maximum de retards p et q.

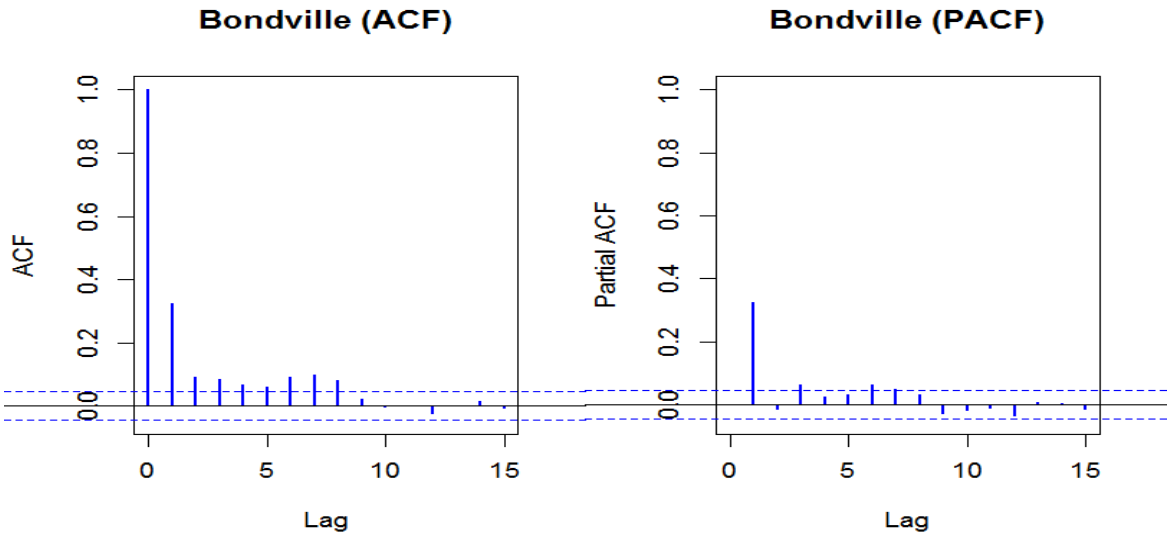

Figure 3.9 : Fonctions d'autocorrélation et d'autocorrélation partielle des données « *log ratios* » du site de Bondville.

À La figure 3.9, nous pouvons choisir le nombre maximum de retard p à la valeur max.p = 6 sur le corrélogramme partiel. En ce qui concerne le nombre maximum de retard q, il peut être estimé à max.q = 8 sur le corrélogramme. Le tableau 3.4 contient la liste des modèles ARIMA possibles générés en exécutant la fonction « *auto.arima* » appliquée aux données « *log ratios* » de Bondville.

Tableau 3.6 : AIC et BIC des processus ARIMA générés à partir des données « *log ratios* » du site de Bondville.

Modèle	**Critères d'information**	
	AIC	**BIC**
ARIMA (2, 0, 2)	1781.4	1814.4
ARIMA (1, 0, 0)	1789.6	1806.2
ARIMA (0, 0, 1)	1800.0	1816.6
ARIMA (1, 0, 2)	1777.2	1805.7
ARIMA (1, 0, 1)	1790.7	1812.7

Au tableau 3.6, le modèle qui semble s'ajuster aux données « *log ratios* » de ce site est le processus ARIMA (1, 0, 2) avec une moyenne non nulle. Ce processus correspond également à un ARMA (1, 2). Les paramètres de ce processus sont rapportés dans le tableau 3.7.

Tableau 3.7 : Paramètres du modèle ARMA (1, 2) estimé à Bondville, les critères d'information et le logarithme du maximum de vraisemblance.

Coefficients		**Erreur Standard (s.e)**
φ	0.8897	0.0587
θ_1	-0.5636	0.0658
θ_2	-0.2205	0.0376
c	0.7788	0.0180

σ^2	0.1543	**BIC**	1805.7
Log(L)	-884.1	**AIC**	1777.2

L'expression analytique du processus ARMA (1, 2) est décrit par les équations suivantes, la deuxième équation est déterminée en substituant les valeurs contenues dans le tableau 3.7 :

$$x_t = (1-\varphi) * c + \varphi * x_{t-1} + \theta_1 * \epsilon_{t-1} + \theta_2 * \epsilon_{t-2} + \epsilon_t.$$

$$x_{t(Bond)} = 0.09 + 0.8897 * x_{t-1(Bond)} - 0.5636 * rbond_{t-1} - 0.2205 * rbond_{t-2} + rbond_t,$$

avec $rbond_t$ correspondant aux résidus du processus ARMA (1, 2) de Bondville à la période t avec variance ($rbond_t$) = 0.1543. Les résidus sont également testés et les résultats sont contenus dans le tableau 3.8. Le p-value est estimé à 29% ce qui est supérieur au seuil de 5%.

Tableau 3.8 : Moyenne, écart type et variance des résidus du ARMA (1, 2) du site de Bondville y compris le p-value du test de McLeod & Li (1983).

Sites	Résidus	Moyenne	RMSE	var	p-value
Bondville	*rbond*	0	0.3928	0.1543	0.29

Nous pouvons supposer que les innovations de la série ARMA (1, 2) du site de Bondville sont « bruit blanc ». La figure 3.10 semble supporter cette hypothèse.

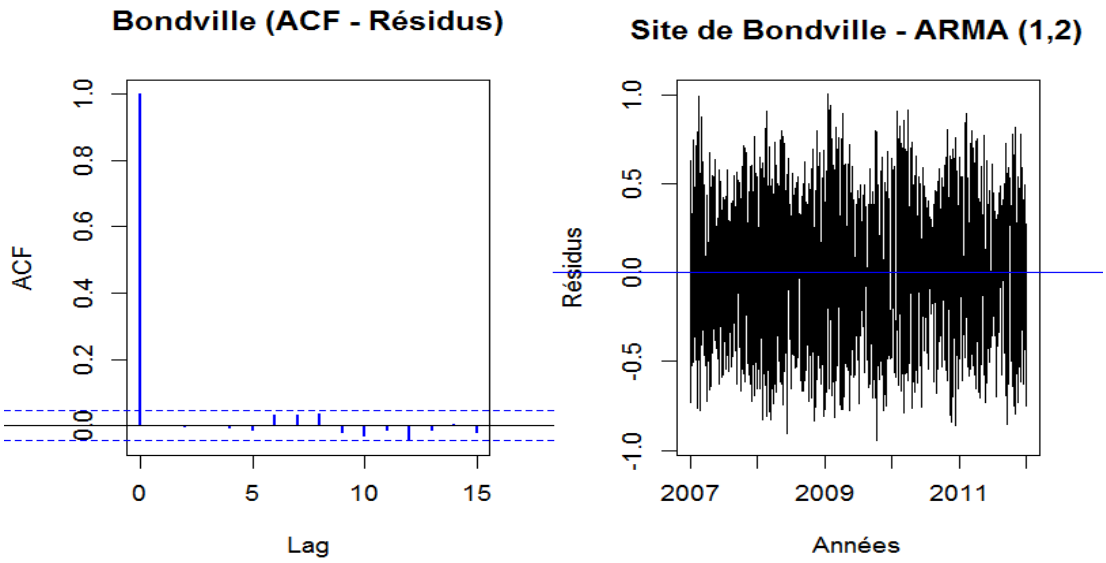

Figure 3.10 : Fonction d'autocorrélation et représentation des résidus du processus ARMA (1, 2) du site de Bondville.

3.3.6. Site de Boulder (Colorado)

Nous appliquons la même formule de la section 3.3.4 aux données « *log ratios* » du site de Boulder dans le Colorado. La figure 3.11 et la figure 3.12 illustrent, respectivement l'allure du modèle physique comparé aux données mesurées et de l'état stationnaire visuel des données de Boulder transformées en « *log ratios* ».

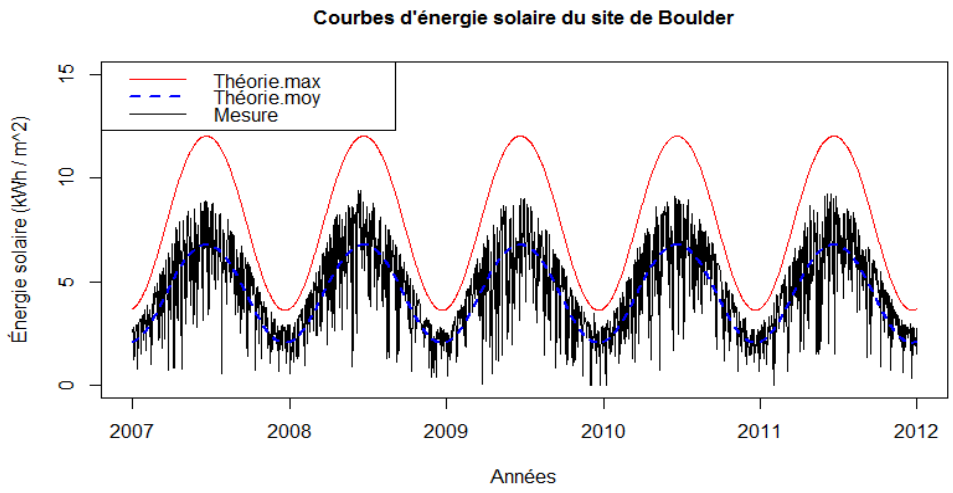

Figure 3.11 : Courbes des modèles théoriques et des données mesurées sur le site de Boulder.

Avec les données « *log ratios* » nous estimons les fonctions d'autocorrélation et d'autocorrélation partielle. Sur les corrélogrammes à la figure 3.13 de la page suivante, nous pouvons choisir max.p = 8 et max.q = 14. La fonction « *auto.arima* » appliquée aux données « *log ratios* » de Boulder génère les processus contenus dans le tableau 3.9.

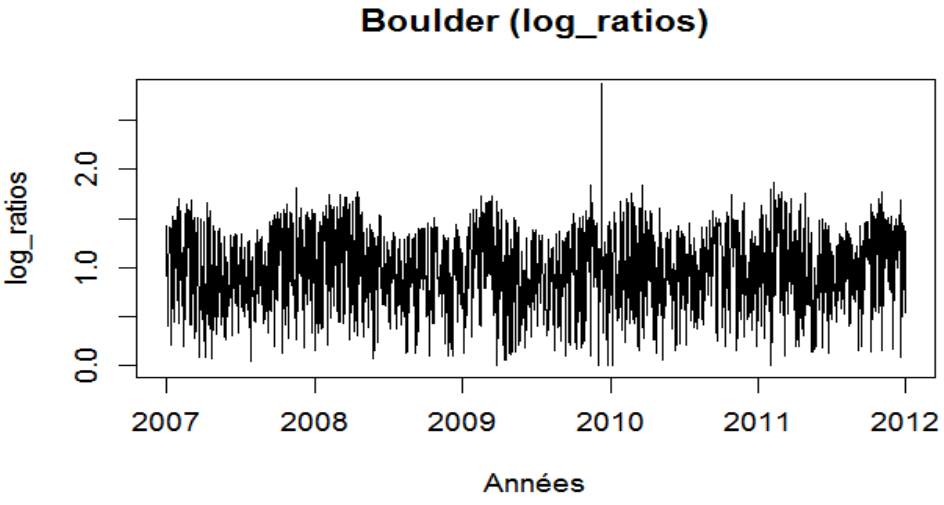

Figure 3.12 : Représentation des données de Boulder transformées en «*log ratios* ».

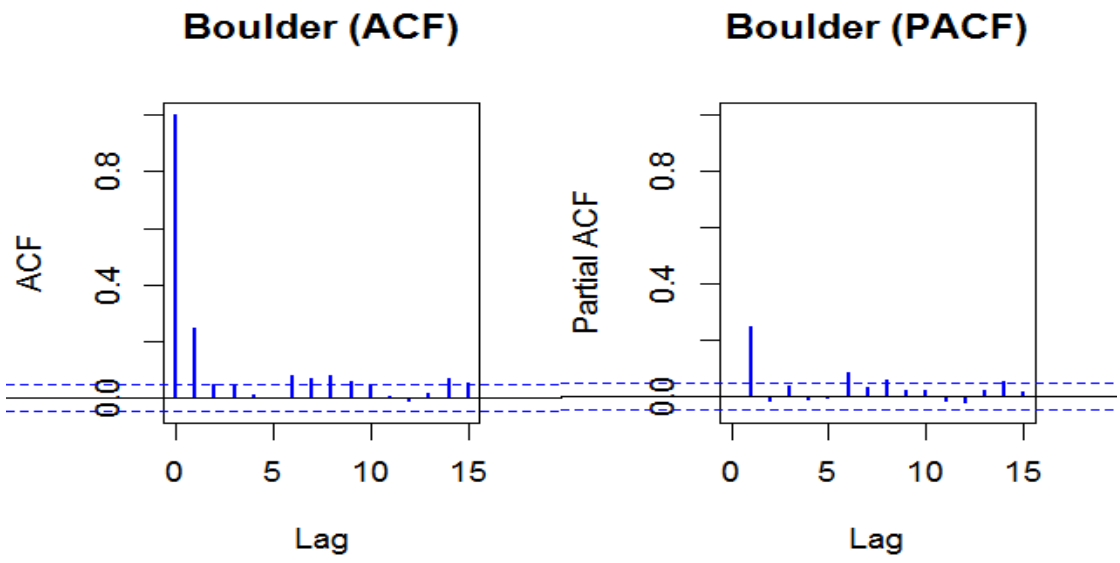

Figure 3.13 : Fonctions d'autocorrélation et d'autocorrélation partielle des données « *log ratios* » du site de Boulder.

Tableau 3.9 : AIC et BIC des processus ARIMA générés à partir des données « *log ratios* » du site de Boulder

Modèle	**Critères d'information**	
	AIC	**BIC**
ARIMA (1, 0, 0)	1824.0	1840.6
ARIMA (0, 0, 1)	1825.2	1841.7
ARIMA (1, 0, 1)	1825.6	1847.7

Avec les critères d'information AIC et BIC indiqués au tableau 3.9, nous choisissons le processus ARIMA (1, 0, 0) avec une moyenne non nulle. Il correspond à un processus ARMA (1, 0), ses paramètres sont compilés dans le tableau 3.10.

Tableau 3.10 : Paramètres du modèle ARMA (1, 0) estimé pour Boulder, les critères d'information et le logarithme du maximum de vraisemblance.

Coefficients		**Erreur Standard (s.e)**
φ	0.2477	0.0227
c	0.9882	0.0124

σ^2	0.1584	**BIC**	1840.6
Log(L)	-908.6	**AIC**	1824.0

Le modèle centré du ARMA (1, 0) du site de Boulder et son expression analytique en substituant des valeurs du tableau 3.10 :

$$x_t = (1 - \varphi) * c + \varphi * x_{t-1} + \epsilon_t.$$
$$x_{t(Bould)} = 0.74 + 0.2477 * x_{t-1(Bould)} + rboul_t.$$

La variance des résidus (*rboul*) à l'instant t est : var ($rboul_t$) = 0.1584. La vérification « bruit blanc » des innovations de ce processus donne les résultats dans le tableau 3.11 qui suit :

Tableau 3.11 : Moyenne, écart type et variance des résidus du ARMA (1, 2) du site de Boulder y compris le p-value du test de McLeod & Li (1983).

Sites	**Résidus**	**Mean**	**RMSE**	**var**	**p-value**
Boulder	rboul	0	0.3980	0.1584	0.47

Le p-value de 47% est supérieur à la valeur seuil de 5%. Il semble vérifier l'hypothèse « bruit blanc » des innovations du site de Boulder (Figure 3.14).

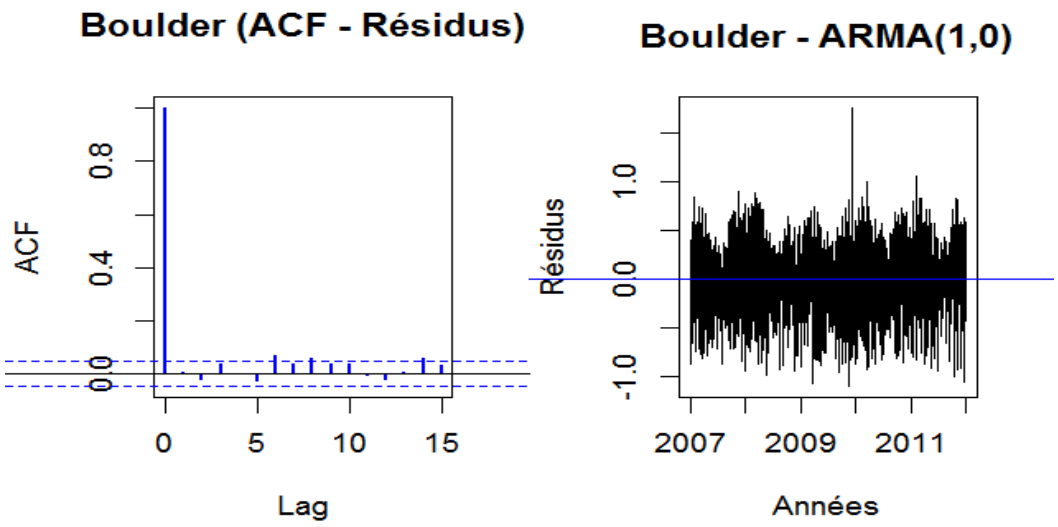

Figure 3.14 : Fonction d'autocorrélation et représentation des résidus du ARMA (1, 0) du site de Boulder.

Les données du site de Fort Peck sont également analysées comme celles des sites précédents.

3.3.7. Site de Fort Peck (Montana)

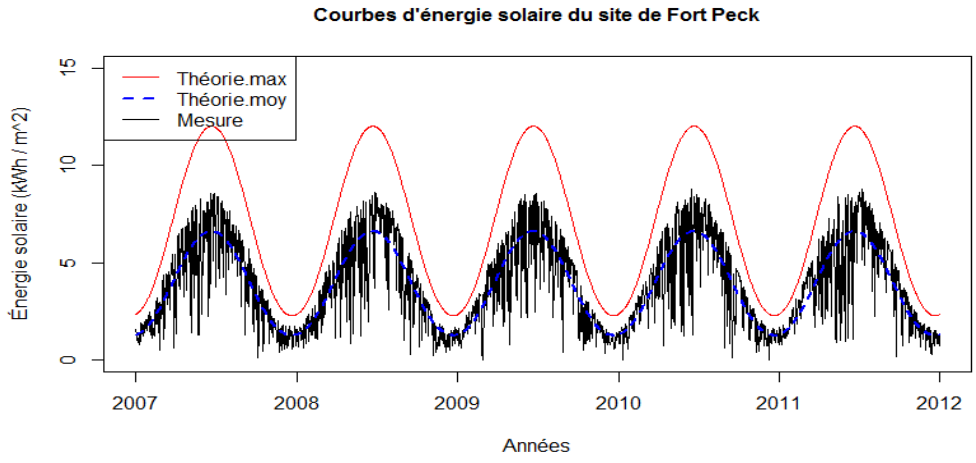

Figure 3.15 : Courbes des modèles théoriques et des données mesurées du site de Fort Peck.

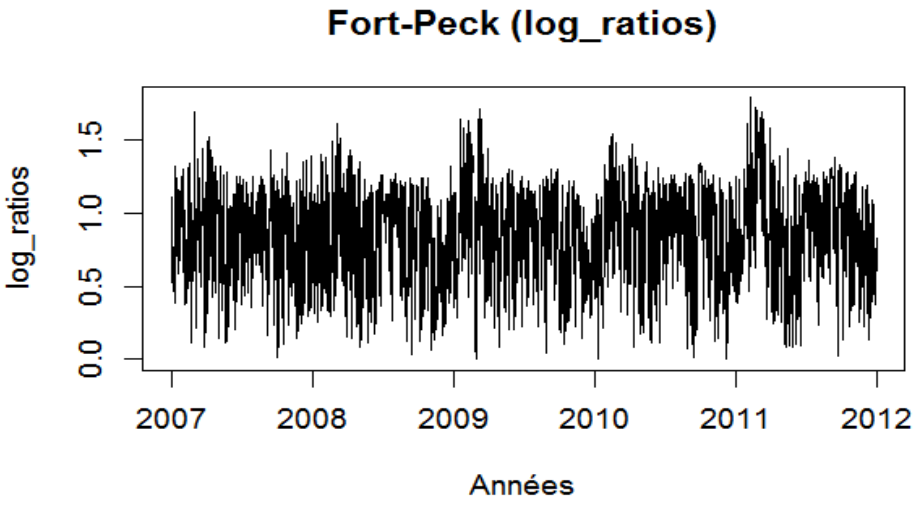

Figure 3.16 : Représentation des données de Fort Peck transformées en « *log ratios* ».

Les modèles théoriques et les données mesurées sont représentés à la Figure 3.15. Il semble avoir la stationnarité des données transformées en « *log ratios* » pour le site de Fort Peck, ceci est illustré à la Figure 3.16.

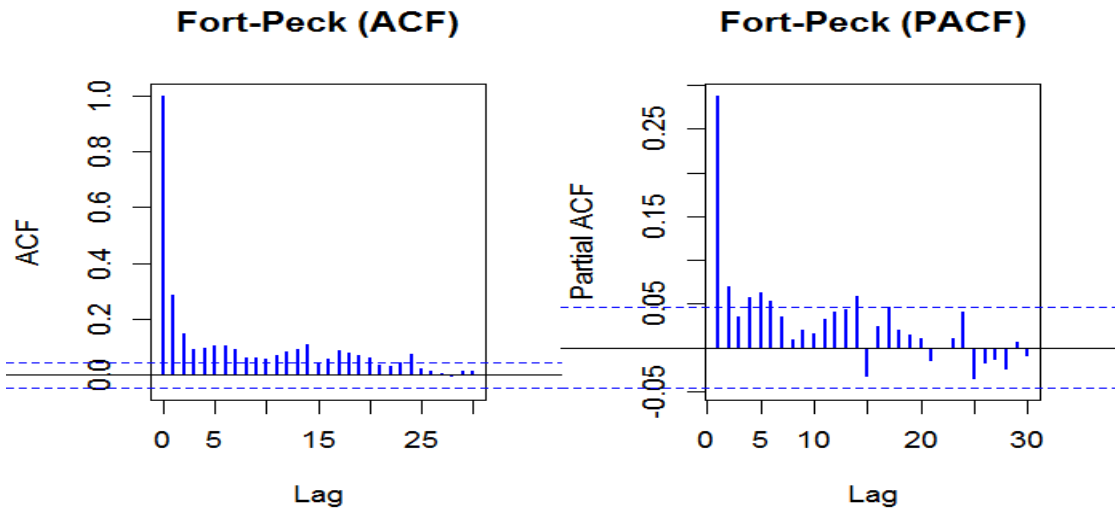

Figure 3.17 : Fonctions d'autocorrélation et d'autocorrélation partielle des données « *log ratios* » de Fort Peck.

Les corrélogrammes à la figure 3.17 permettent d'estimer max.p et max.q. En observant les graphiques de cette figure, nous choisissons max.p à 14 et max.q à 24. Les modèles possibles générés sont contenus dans le tableau 3.12.

Tableau 3.12 : AIC et BIC des processus ARIMA générés à partir des données « *log ratios* » du site de Fort Peck.

Modèle	**Critères d'information**	
	AIC	**BIC**
ARIMA (2, 0, 2)	1194.7	1227.8
ARIMA (1, 0, 0)	1383.9	1245.3
ARIMA (0, 0, 1)	1254.3	1270.9
ARIMA (1, 0, 1)	1213.0	1239.9
ARIMA (1, 0, 2)	1199.3	1226.9

Le modèle qui s'ajuste aux données « *log ratios* » du site de Fort Peck semble être le processus ARIMA (1, 0, 2) avec une moyenne non nulle, c'est également un processus ARMA (1, 2). Le tableau 3.13 contient les paramètres de ce modèle.

Tableau 3.13 : Paramètres du modèle ARMA (1, 2) estimé à Fort Peck, les critères d'information et le logarithme du maximum de vraisemblance.

Coefficients		**Erreur Standard (s.e)**
φ	0.9444	0.0196
θ_1	-0.7030	0.0308
θ_2	-0.1527	0.0243
c	0.8547	0.0202

σ^2	0.1121	**BIC**	1226.9
Log(L)	-592.5	**AIC**	1199.3

L'expression de ce modèle ARMA (1, 2) centré du site de Fort Peck correspond à :

$$x_{t(Fort)} = 0.05 + 0.9444 * x_{t-1(Fort)} - 0.7030 * rfort_{t-1} - 0.1527 * rfort_{t-2} + rfort_t,$$

avec variance ($rfort_t$) = 0.1121. Le test de McLeod & Li (1983) sur les résidus donne les résultats compilés dans le tableau 3.14. Le p-value est estimé à 41%. Cette valeur est supérieure au seuil de 5%. L'hypothèse des résidus « bruit blanc » semble possible. Ceci s'observe aussi visuellement sur la figure 3.18.

Tableau 3.14 : Moyenne, écart type et variance des résidus du ARMA (1, 2) du site de Fort Peck y compris le p-value du test de McLeod & Li (1983).

Sites	Résidus	Moyenne	RMSE	var	p-value
Fort Peck	*rfort*	0	0.3348	0.1121	0.41

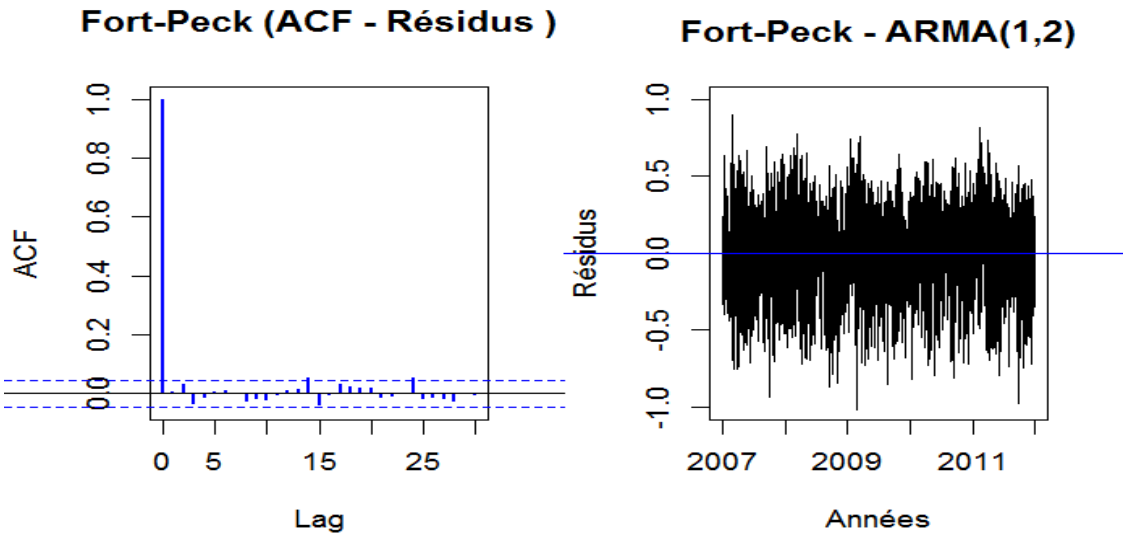

Figure 3.18 : Fonction d'autocorrélation et représentation des résidus ARMA (1, 2) du site de Fort Peck.

3.3.8. Site de Goodwin Creek (Mississipi)

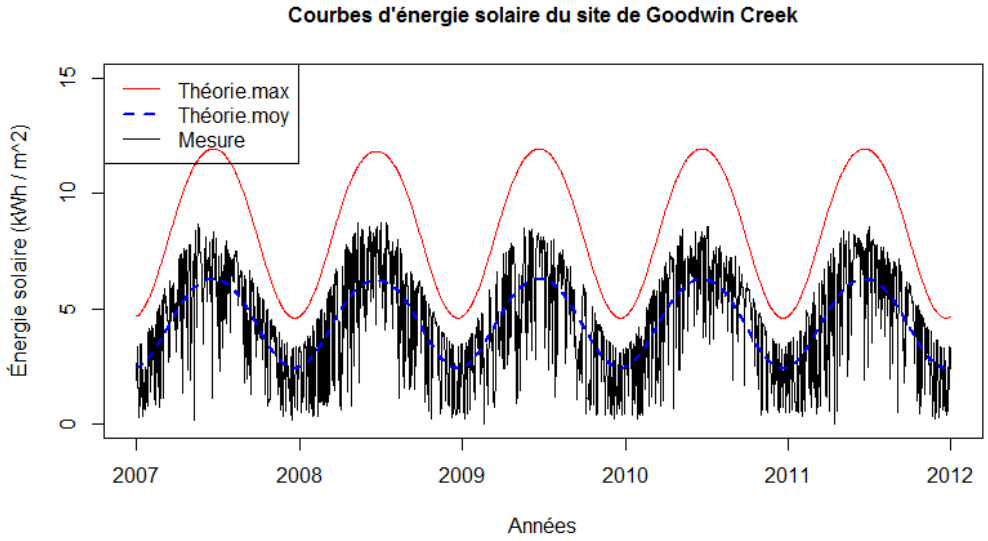

Figure 3.19 : Courbes des modèles théoriques et des données mesurées au site de Goodwin Creek.

Les mêmes commentaires s'appliquent au site de Goodwin Creek (voir la figure 3.19 et la figure 3.20).

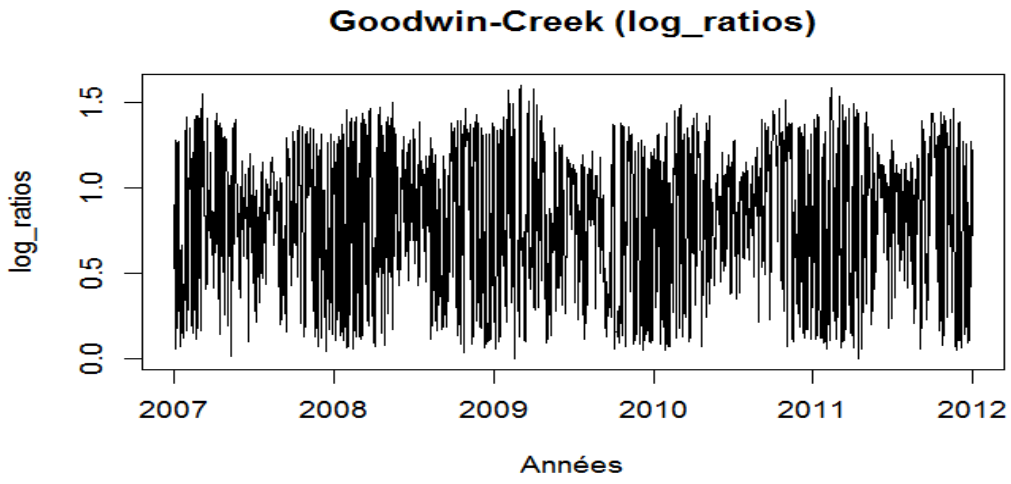

Figure 3.20 : Représentation des données de Goodwin Creek transformées en « *log ratios* ».

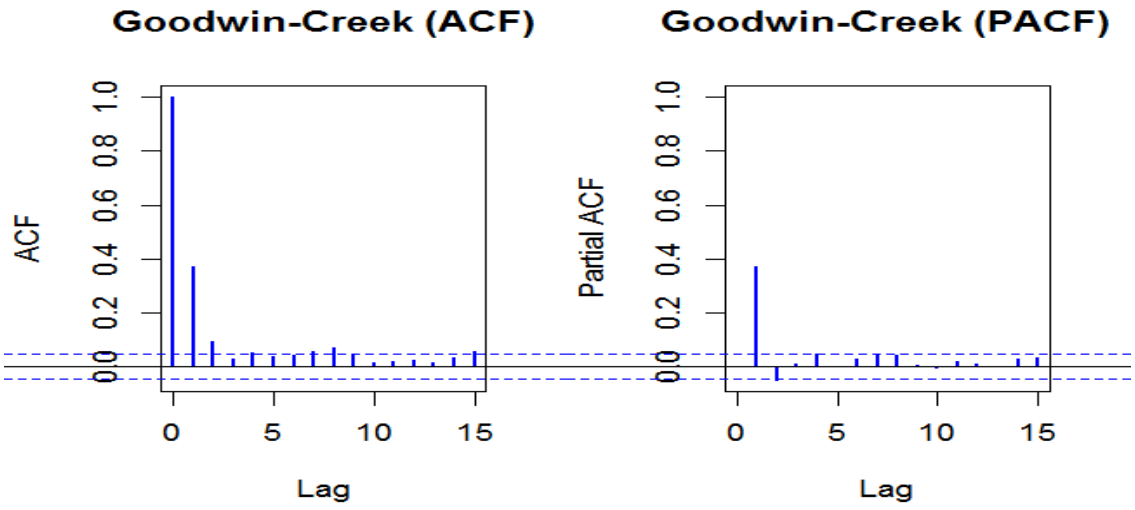

Figure 3.21 : Fonctions d'autocorrélation et d'autocorrélation partielle des données transformées en « *log ratios* » de Goodwin Creek.

À la figure 3.21, le max.p = 4 et max.q = 8. Le tableau 3.15 donne les modèles possibles générés avec la fonction de Hyndman et Khandakar (2008).

Tableau 3.15 : AIC et BIC des processus ARIMA générés à partir des données « *log ratios* » du site de Goodwin Creek.

Modèle	**Critères d'information**	
	AIC	**BIC**
ARIMA (2, 0, 2)	1715.7	1748.7
ARIMA (2, 0, 1)	1714.1	1741.6
ARIMA (1, 0, 0)	1716.0	1732.5
ARIMA (0, 0, 1)	1727.9	1744.4
ARIMA (1, 0, 1)	1713.0	1735.0

Le processus ARIMA (1, 0, 1) s'ajuste aux données « *log ratios* » du site de Goodwin Creek avec une moyenne non nulle. C'est aussi un processus ARMA (1, 1). Les paramètres du modèle sont dans le tableau 3.16.

Tableau 3.16 : Paramètres du modèle ARMA (1, 1) estimé du site de Goodwin, les critères d'information et le logarithme du maximum de vraisemblance.

Coefficients		**Erreur Standard (s.e)**
φ	0.2518	0.0587
θ	0.1350	0.0598
c	0.8252	0.0138

σ^2	0.1503	**BIC**	1735.0
Log(L)	-852.4	**AIC**	1713.0

L'expression analytique de ce modèle ARMA (1, 1) correspond à l'équation qui suit :

$$x_{t(Good)} = 0.6174 + 0.2518 * x_{t-1(Good)} + 0.1350 * rgood_{t-1} + rgood_t,$$

avec la variance des résidus (*rgood* à la période t) var ($rgood_t$) = 0.1503. Le test de spécification des innovations semble supposer qu'elles suivent un processus « bruit blanc » (le tableau 3.17 et la figure 3.22).

Tableau 3.17 : Moyenne, écart type et variance des résidus du site de Goodwin Creek y compris le p-value du test de McLeod & Li (1983).

Sites	**Résidus**	**Moyenne**	**RMSE**	**var**	**p-value**
Goodwin Creek	*rgood*	0	0.3877	0.1503	0.59

Une observation visuelle à la figure 3.22, permet de dire que les résidus de la série chronologique ARMA (1, 1) du site de Goodwin Creek semblent suivre un processus « bruit blanc ».

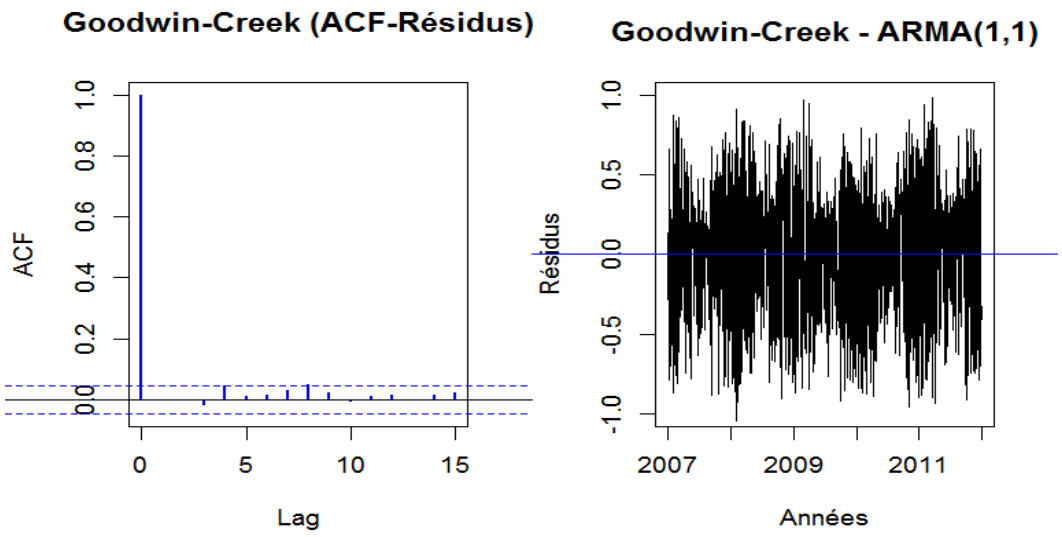

Figure 3.22 : Fonction d'autocorrélation et représentation des résidus du ARMA (1, 1) du site de Goodwin Creek.

3.3.9. Site de Sioux Fall (Sud Dakota)

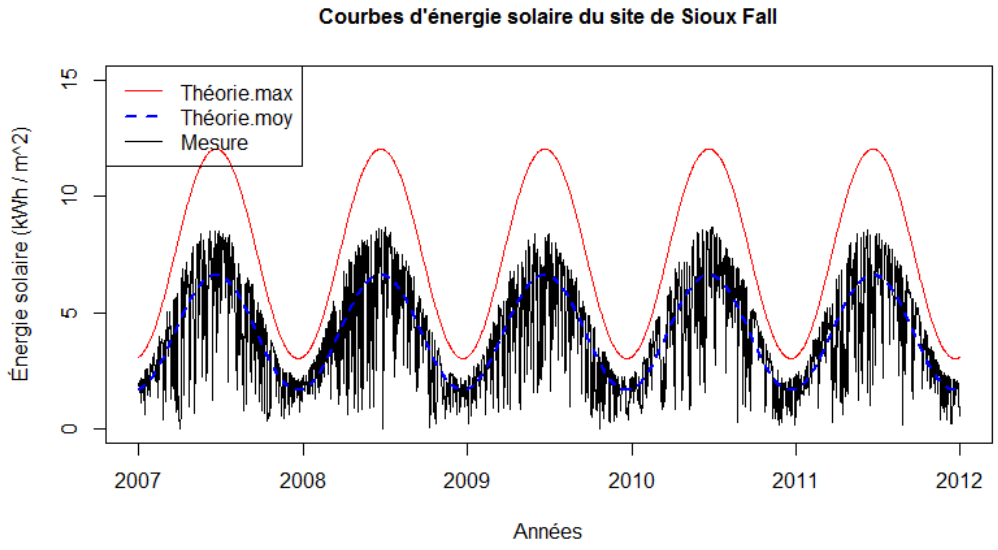

Figure 3.23 : Courbes des modèles théoriques et des données mesurées du site de Sioux Fall.

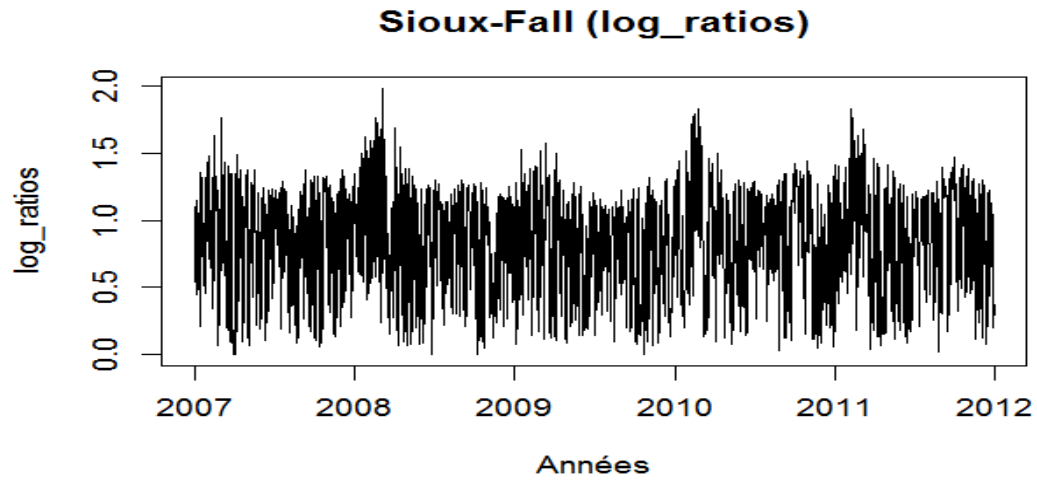

Figure 3.24 : Représentation des données de Sioux Fall transformées en « *log ratios* ».

Les mêmes commentaires des figures précédentes s'appliquent au site de Sioux Fall (voir la figure 3.23 et la figure 3.24).

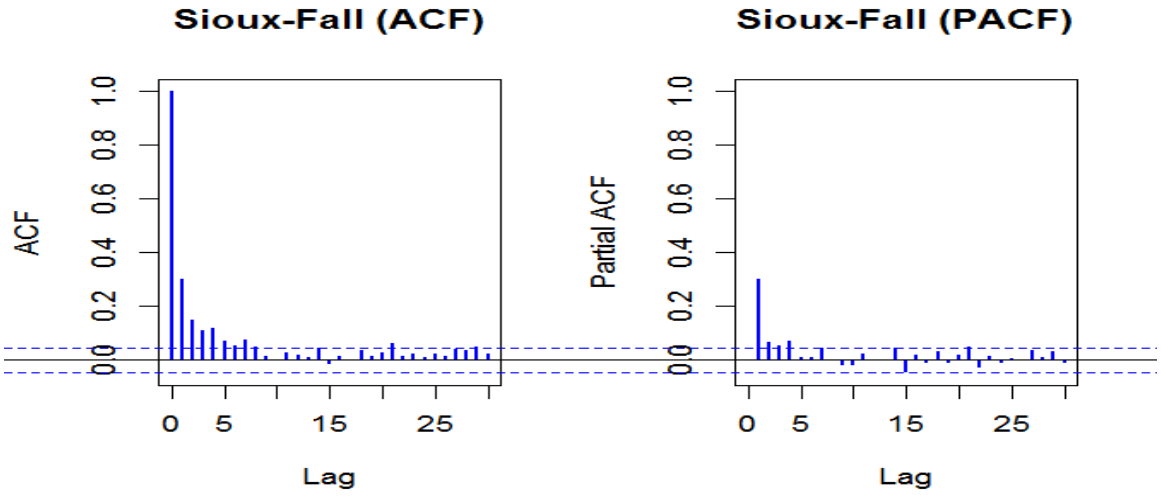

Figure 3.25 : Fonctions d'autocorrélation et d'autocorrélation partielle des données « *log ratios* » de Sioux Fall.

À la figure 3.25, le nombre de retard (Lag) est pris à 30 parce que la fonction d'autocorrélation s'estompe à partir du Lag 21, ce qui est différent dans les analyses

précédentes. Nous choisissons ici le max.p à 15 et le max.q = 21. Les modèles possibles générés sont consignés dans le tableau 3.18 avec les différents critères d'information AIC et BIC.

Tableau 3.18 : AIC et BIC des processus ARIMA générés à partir des données « *log ratios* » du site de Sioux Fall.

Modèle	**Critères d'information**	
	AIC	**BIC**
ARIMA (2, 0, 3)	1674.4	1712.9
ARIMA (2, 0, 2)	1674.4	1707.5
ARIMA (1, 0, 2)	1670.9	1689.5
ARIMA (1, 0, 0)	1689.1	1705.6
ARIMA (0, 0, 1)	1715.6	1732.1
ARIMA (0, 0, 2)	1694.0	1716.0
ARIMA (1, 0, 3)	1672.7	1705.8
ARIMA (1, 0, 1)	1677.0	1699.1

Le modèle qui semble s'ajuster est le processus ARIMA (1, 0, 2) avec une moyenne non nulle. C'est un ARMA (1, 2). Les paramètres du modèle sont dans le tableau 3.19.

Tableau 3.19 : Paramètres du modèle ARMA (1, 2) estimé à Sioux Fall, les critères d'information et le logarithme du maximum de vraisemblance.

Coefficients		**Erreur Standard (s.e)**
φ	0.8089	0.0592
θ_1	-0.5405	0.0653
θ_2	-0.1003	0.0335
c	0.8447	0.0168

$\boldsymbol{\sigma^2}$	0.1453	**BIC**	1689.5
Log(L)	-829.9	**AIC**	1670.9

L'expression algébrique du modèle centré du ARMA (1, 2) de Sioux Fall est de :

$$x_{t(Sioux)} = 0.1614 + 0.8089 * x_{t-1(Sioux)} - 0.5405 * rsiou_{t-1} - 0.1003 * rsiou_{t-2} + rsiou_t,$$

avec la variance ($rsiou_t$) = 0.1584. En appliquant le test de spécification de McLeod et Li (1983) aux résidus de ce modèle, nous avons les résultats compilés dans le Tableau 3.20 suivant :

Tableau 3.20 : Moyenne, écart type et variance des résidus du processus ARMA (1, 2) du site de Sioux Fall ainsi que le p-value du test de McLeod & Li (1983).

Sites	Résidus	Moyenne	RMSE	var	p-value
Sioux Fall	*rsiou*	0	0.3812	0.1453	0.49

Le p-value de 49% est supérieur à la valeur seuil de 5%. Les résidus de la série suivent un processus « bruit blanc ». Cette hypothèse est illustrée à la figure 3.26 qui montre une moyenne nulle de ces résidus et une autocorrélation inexistante.

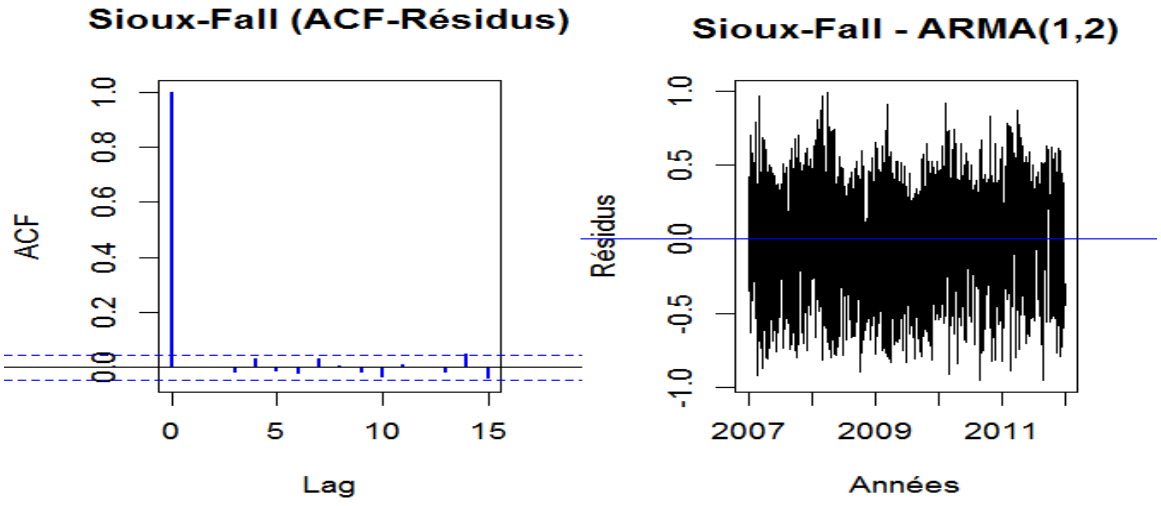

Figure 3.26 : Fonction d'autocorrélation et représentation des résidus du processus ARMA (1, 2) du site de Sioux Fall.

3.4. Synthèse des résultats et conclusion

Trois tableaux regroupent l'ensemble des modèles de séries chronologiques représentant chacun des six sites étudiés.

3.4.1. Résultats

Les tableaux 3.21, 3.22 et 3.23 présentent la synthèse des modèles de processus ARMA par site. Nous avons respectivement un résumé des processus et les écarts types entre les valeurs des résidus, un résumé des expressions algébriques de chaque processus et un résumé de l'ensemble des variances des résidus et les résultats du test de McLeod & Li.

Tableau 3.21 : Récapitulatif des processus ARMA et les écarts types des résidus.

N°	Sites	Modèles TS	RMSE
1	Penn State	ARMA (1,0)	0.3828
2	Bondville	ARMA (1, 2)	0.3928
3	Boulder	ARMA (1, 0)	0.3980
4	Fort Peck	ARMA (1, 2)	0.3348
5	Goodwin Creek	ARMA (1, 1)	0.3877
6	Sioux Fall	ARMA (1, 2)	0.3812

Tableau 3.22 : Récapitulatif des expressions analytiques des processus ARMA.

Site 1	$x_{\mathrm{t}(Penn)} = 0.47 + 0.3142 * x_{t-1(Penn)} + rpenn_{\mathrm{t}}$
Site 2	$x_{t(Bond)} = 0.08 + 0.8963 * x_{t-1(Bond)} - 0.5709 * rbond_{t-1} - 0.2240 * rbond_{t-2} + rbond_t$
Site 3	$x_{t(Bould)} = 0.74 + 0.2477 * x_{t-1(Bould)} + rboul_t$
Site 4	$x_{t(Fort)} = 0.05 + 0.9444 * x_{t-1(Fort)} - 0.7030 * rfort_{t-1} - 0.1527 * rfort_{t-2} + rfort_t$
Site 5	$x_{t(Good)} = 0.62 + 0.2518 * x_{t-1(Good)} + 0.1350 * rgood_{t-1} + rgood_t$
Site 6	$x_{t(Sioux)} = 0.16 + 0.8089 * x_{t-1(Sioux)} - 0.5405 * rsiou_{t-1} - 0.1003 * rsiou_{t-2} + rsiou_t$

Dans l'ensemble, il y a deux processus ARMA (1, 0) correspondant respectivement aux sites de Penn-State et de Boulder, trois processus ARMA (1, 2) pour les sites de Bondville, Fort-Peck et Sioux-Fall et enfin un processus ARMA (1, 1) pour le site de Goodwin Creek.

Tableau 3.23 : Résumé des moyennes, des variances des résidus et le p-value du test de McLeod & Li (1983).

N°	Sites	Résidus	Moyenne	var	p-value
1	Penn State	*rpenn*	0	0.1465	0.11
2	Bondville	*rbond*	0	0.1543	0.29
3	Boulder	*rboul*	0	0.1584	0.47
4	Fort Peck	*rfort*	0	0.1121	0.41
5	Goodwin Creek	*rgood*	0	0.1503	0.59
6	Sioux Fall	*rsiou*	0	0.1453	0.49

3.4.2. Conclusion

L'analyse visuelle des modèles physiques et des données mesurées de rayonnement solaire pour chaque site montre une évolution qui suit une même allure. Précisons que les modèles physiques sont estimés en supposant les conditions idéales d'ensoleillement (ciel clair) et une opacité atmosphérique (transmittance = 1). Ces hypothèses donnent un modèle théorique qui a une amplitude moyennement supérieure à celle des mesures réelles d'énergie solaire. Mais les transformations effectuées sur ces données ont permis de réduire les effets saisonniers existant, d'éliminer la tendance captée par le modèle physique et d'avoir des données transformées qui semblent stationnaires. Pour estimer les processus ARMA (p, q) stationnaires, nous sommes partis de deux ensembles de données par site (données théoriques et mesures réelles), nous avons fait des transformations, choisi le nombre maximum de retard par les fonctions d'autocorrélation et autocorrélation partielle avant d'estimer les modèles par la fonction « *auto.arima* ». C'est une méthodologie différente et relativement simple

par rapport aux travaux déjà réalisés par Boland (2008), Azami et *coll.* (2009) et par Nowicka-Zagrajek et Weron (2002) qui sont mentionnés dans les paragraphes de la section 1.3.2 au chapitre 2.

L'analyse des innovations des six processus ARMA estimés suppose que les résidus suivent un processus « bruit blanc », c'est-à-dire qu'elles ont une moyenne nulle, que les valeurs de ces résidus ne sont pas corrélées entre elles et qu'elles sont stationnaires. Pour la suite des travaux, nous générons une matrice appelée « *allres* » dans lequel nous compilons les résidus de chaque site. Cette matrice comporte six (6) colonnes et chaque élément représente les innovations du processus ARMA de chacun des sites (tableau 3.22 et tableau 3.23). Les éléments de la matrice « *allres* » représente les données d'entrée que nous utilisons pour analyser la dépendance spatiale des sites au chapitre 4.

Chapitre 4 : Dépendance spatiale

Au chapitre 3, nous avons estimé six processus stochastiques stationnaires (ARMA) qui s'ajustent aux données de nos six sites (Penn-State, Bondville, Boulder, Fort-Peck, Goodwin-Creek et Sioux-Fall)[3]. Même si chaque site est modélisé correctement, leur comportement conjoint peut être lié. Pour modéliser la dépendance spatiale des données, nous ajustons un modèle multivarié aux innovations des séries chronologiques. Ainsi, chaque site a des innovations « bruit blanc » marginalement, mais leur distribution conjointe comporte une structure de dépendance. Pour simplifier, nous supposons que les innovations demeurent indépendantes d'une journée à l'autre, même pour les sites d'ouest en est. Même si la normale multivariée est le modèle le plus naturel pour représenter la dépendance géographique, nous évaluerons aussi une approche par copule. Dans ce chapitre, nous présenterons :

1. La matrice des corrélations linéaires entre les résidus et une analyse succincte de la dépendance géographique des différents sites exprimée par ces corrélations.

2. Une analyse de la structure de dépendance basée sur les copules bivariées. Cela permet d'identifier une famille de copules bivariées usuelles (Gaussienne, Student, Clayton, Frank et Gumbel) qui s'ajuste aux données (résidus). Pour cela nous utilisons comme données, les résidus des sites corrélés géographiquement (tableau 4.1 : matrice des corrélations). Notre analyse s'inspirera des travaux de Rémillard (2013), Genest et MacKay (1986a, b) et Genest et Rémillard (2004).

3. Une analyse de la structure de dépendance basée sur la méthode des copules en vigne. Nous utiliserons essentiellement la R-Vine qui est la méthode d'estimation régulière des copules en vigne « Vine-Copulas » (Kramer et Schepsmeier (2011),

[3] Figure 3.1 : Position géographique des sites sur la carte des États-Unis d'Amérique.

Brechmann (2013) et Dissmann & coll., (2013). Il sera question ici d'élaborer une structure de dépendance en considérant l'ensemble des innovations des six sites. Cette façon de faire permet une visualisation graphique des liens de dépendance de l'ensemble des sites et le regroupement des paires de copules bivariées pour déterminer une copule qui s'ajuste à un ensemble de données de dimension supérieure à deux (2).

Cette manière d'aborder la compréhension et l'analyse de la structure de dépendance multivariée d'un ensemble de sites de production d'énergie solaire géographiquement dispersés est nouvelle. Elle constitue donc notre contribution pour mettre en évidence les risques de production d'électricité des centrales solaires conjointement dépendantes, lors de la survenance de problèmes liés aux conditions atmosphériques ou climatiques tels que les couverts nuageux, des pluies ou même une période hivernale prolongée. Mais avant de présenter nos résultats et notre analyse, nous présentons brièvement quelques définitions théoriques des familles de copules utilisées dans notre étude, le concept de dépendance et la description théorique des tests d'ajustement pour le choix de la copule bivariée qui exprime le mieux la structure de dépendance recherchée.

4.1. Théorie des copules

Les copules sont des outils de modélisation de la structure de dépendance des variables aléatoires. La vulgarisation de cet outil probabiliste a été essentielle à la compréhension de certains domaines tels que la finance quantitative, la réplique de la performance des modèles statistiques, les mesures de risques multiples, etc. (Genest et MacKay, 1986), (Rémillard, 2013). Il est essentiel de distinguer les comportements des distributions marginales de la structure de dépendance. Les copules permettent d'extraire la structure de dépendance d'une distribution conjointe. Elles séparent donc la dépendance et le comportement marginal. En isolant cette structure de dépendance, il est alors possible de déterminer une loi multivariée originale (Genest & *coll.*, 2009).

4.1.1. Théorème de Sklar

Lorsque qu'une copule est définie à l'aide d'une loi bivariée préexistante, il est normal de se référer au théorème de Sklar du nom de celui qui a introduit le concept de copule en 1959. Ce théorème précise le lien défini par une copule *C* déterminée à partir d'une distribution jointe *F* et des fonctions de répartition marginales univariées F_1 et F_2 de la distribution complète bivariée *F*.

Théorème de Sklar : *Soit F une fonction de distribution bivariée ayant des marginales* F_1, F_2. *Il existe une copule C associée à F qui s'écrit :*
$C\ (F_1(x),\ F_2(y)) = F\ (x,\ y)$. *De plus, si les lois marginales* F_1 *et* F_2 *sont des lois continues alors la copule associée est unique.*

En schématisant le théorème de Sklar, nous pouvons dire que : si les marginales de chaque variable *x* et *y* sont données, il suffit de les joindre par une fonction copule ayant les propriétés de dépendance souhaitées de manière à obtenir une loi jointe.

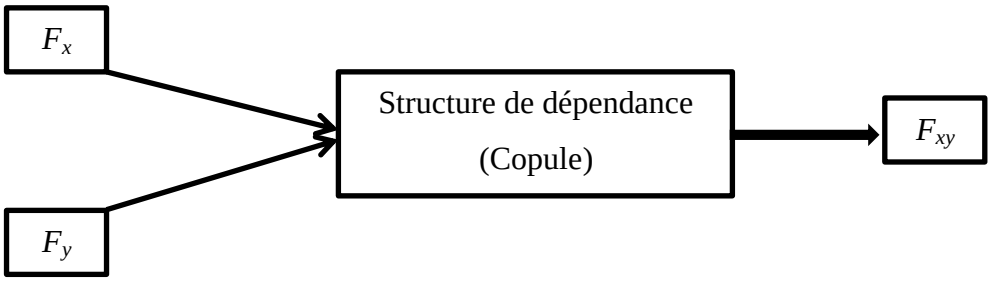

Figure 4.1 : Schéma simplifié du théorème de Sklar avec F_x et F_y les fonctions marginales de la fonction de distribution bivariée F_{xy} (x et y sont des variables).

Genest & *coll.* (2009) définissent également les copules en mettant en évidence la notion de bidimension. Ceci se traduit par la définition suivante :

Définition : *La copule bivariée C fonction de* $[0,1]^2 \rightarrow [0, 1]$ *est définie par les caractéristiques suivantes :*

(i) $C(u, 0) = C(0, u) = 0, \forall u \in [0, 1]$

(ii) $C(u, 1) = C(1, u) = u, \forall u \in [0, 1]$ *: les marges des distributions marginales sont des marges uniformes,*

(iii) La copule C est deux fois croissante :

$C(v_1, v_2) - C(v_1, u_2) - C(u_1, v_2) + C(u_1, u_2) \geq 0, \forall (u_1, u_2) \in [0, 1]^2, (v_1, v_2) \in [0, 1]^2$ *tel que* $0 \leq u_1 \leq v_1 \leq 1$ *et* $0 \leq u_2 \leq v_2 \leq 1$.

Comme une distribution de probabilité avec des marges uniformes, l'expression d'une copule serait : *Soient* U_1 *et* U_2 *deux variables aléatoires uniformes sur [0, 1], alors on a :* $C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2), \forall (u_1, u_2) \in [0, 1]^2$.

L'estimation des copules se fait en général avec les rangs, ce qui évite de travailler avec les valeurs brutes des variables. Ce qui se traduit dans la pratique à la transformation linéaire des réalisations ($x_1, x_2, \ldots, x_n$) en variables uniformes empiriques $u_1, u_2, \ldots, u_n$ où $u_i = \frac{Rang(x_i)}{n+1}$ pour tout $i \in [1, \ldots, n]$. Cette manière de procéder se justifie par le fait que les rangs ne dépendent pas des lois marginales (Genest et Rémillard, 2004).

Les deux grandes familles de copules auxquelles nous ferons référence dans ce mémoire sont respectivement la famille de copules elliptiques et la famille des copules archimédiennes. Il s'agira essentiellement des copules usuelles multivariées (Normale et Student) pour les copules elliptiques et des copules archimédiennes bivariées telles que la copule de Frank, celle de Gumbel et la copule de Clayton.

4.1.2. Famille des copules elliptiques

Genest & McKay, (1986a, b) proposent une définition pour chacune des deux copules souvent utilisées dans cette famille. Il s'agit des copules de Student et la Gaussienne que nous avons indiquées précédemment.

La copule Gaussienne (ou normale)

Définition : *La copule Gaussienne bivariée C fonction de* $[0,1]^2 \rightarrow [0, 1]$ *est définie par les caractéristiques suivantes :*

$c(u_1, u_2; \rho) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$. *$\rho$ est le coefficient de corrélation, Φ_ρ est la fonction de répartition de matrice de corrélation ρ et Φ est la fonction quantile de la loi normale standard.*

L'une des observations qui ressort dans l'étude de Brechmann et Schepsmeier (2013) est que la copule normale modélise passablement une dépendance non linéaire ou en présence d'évènements extrêmes (sauf pour une corrélation parfaite). Néanmoins, il est très facile de simuler des variables aléatoires avec une distribution conjointe Normale (ou Gaussienne).

La copule de Student

La copule de Student est estimée suivant le même principe que la copule Normale (ou Gaussienne) mais à partir de la distribution de Student bivariée.

Définition (Genest & McKay, 1986a, b) : La copule de Student bivariée est définie de la façon suivante : $C(u_1, u_2; \rho, k) = T_{\rho, k}(T_k^{-1}(u_1), T_k^{-1}(u_2))$. *Avec ρ comme le coefficient de corrélation, $T_{\rho,k}$ la distribution de Student bivariée standard de matrice de corrélation ρ et de degré de liberté k et T_k^{-1} est la fonction inverse de la distribution standard de Student à k degrés de liberté.*

Remarque : Les copules Normale et celle de Student sont des copules symétriques et relativement simples à utiliser parce que les distributions associées sont bien connues. Elles sont souvent appelées les copules implicites car elles n'ont pas de forme analytique explicite. Ces copules s'expriment, par conséquent, en fonction des distributions bivariées auxquelles elles sont associées (Théorème de Sklar).

4.1.3. Famille des copules archimédiennes

Définition Genest et MacKay (1986a, b) **:** Les copules Archimédiennes sont définies de la manière suivante :

$$c(u_1, u_2) = \begin{cases} \varphi^{-1}(\varphi(u_1) + \varphi(u_2)), & si\ \varphi(u_1) + \varphi(u_2) \leq \varphi(0) \\ 0, sinon \end{cases}.$$

Avec φ vérifiant $\varphi(1) = 0$, $\varphi'(u) < 0$ *et* $\varphi''(u) > 0$ *pour tout* $0 \leq u \leq 1$. *φ est la fonction génératrice de la copule.*

Ces copules regroupent entre autres la copule de Clayton, la copule de Frank, la copule de Gumbel, etc. Pour ces copules, la transformation appliquée aux marginales rend les valeurs indépendantes. Le tableau 4.1 contient les copules archimédiennes auxquelles nous ferons souvent référence dans notre analyse de la structure de dépendance entre les innovations représentant les sites.

Tableau 4.1 : Copules, générateurs et intervalle de définition du paramètre thêta pour les familles de Clayton, Gumbel et Frank.

Nom	Générateur	Copule bivariée
Clayton ($\theta > 0$)	$u^{-\theta} - 1$	$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$
Gumbel ($\theta \geq 0$)	$(-\ln u)^{\theta}$	$\exp\left(-\left(u_1^{\theta} + u_2^{\theta}\right)^{1/\theta}\right)$
Frank ($\theta \neq 0$)	$-ln\frac{e^{-\theta u} - 1}{e^{-\theta} - 1}$	$-\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1}\right)$

Particularité des copules archimédiennes (Genest & MacKay, 1986a, b) :

- La copule de Gumbel capte les dépendances positives et est plus accentuée sur la queue supérieure.

- La copule de Frank capte les dépendances aussi bien positives que négatives.

- La copule de Clayton quant à elle capte les dépendances positives et particulièrement entre les événements à faible intensité (graphique « a » Figure 4.2).

En comparaison, dans l'ajustement des copules Gaussienne (normale) et de Student il peut être pertinent d'utiliser le coefficient de corrélation usuel dans l'analyse de la structure de dépendance entre les variables. Les erreurs d'interprétation et limites sont discutées dans les travaux d'Embrechts & *coll.* (1999). Nous illustrons ici quelques distributions de ces copules à la Figure 4.2.

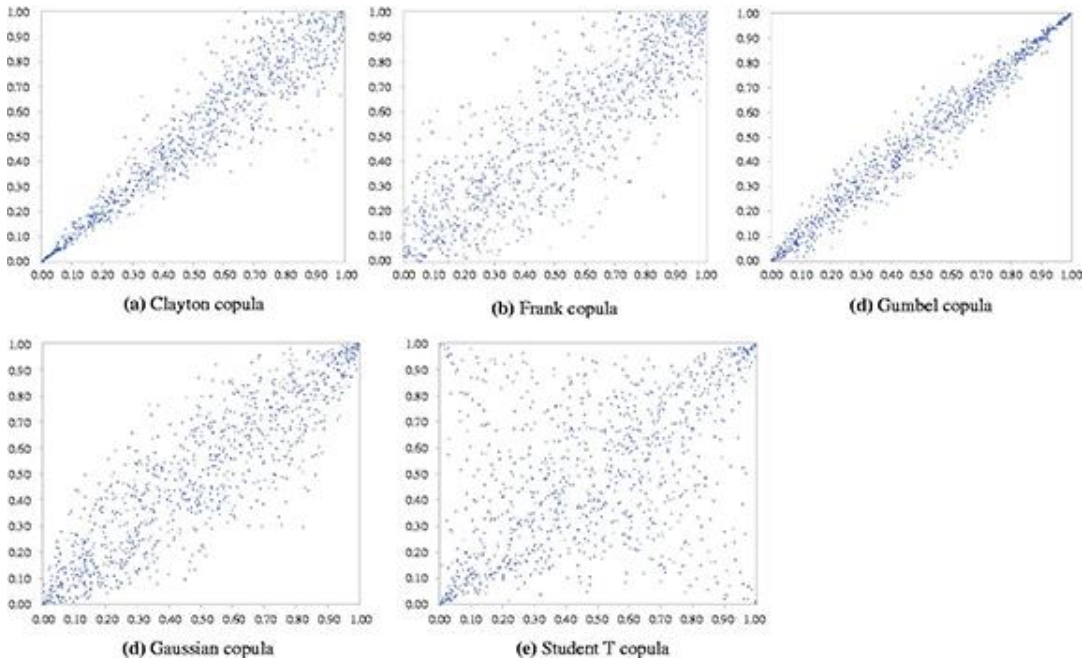

Figure 4.2 : Illustrations des distributions des copules de Clayton, Frank, Gumbel, Gaussienne et de Student. Source : Larsen & *coll.* (2013).

4.1.4. Dépendance caudale

La dépendance caudale fournit une description de la dépendance au niveau des queues de distribution, cette notion est utile pour étudier la survenance simultanée de valeurs extrêmes. C'est une mesure locale contrairement au tau de Kendall (voir section 4.2.1) qui mesure la dépendance sur l'ensemble de la distribution. Ce concept est essentiel dans l'analyse de la dépendance asymptotique entre deux variables aléatoires (Genest et Rémillard, 2004). Lorsque nous aurons besoin de comparer les modèles au chapitre 5, La notion de valeurs extrêmes nous sera utile.

4.2. Tests de la qualité de l'ajustement (goodness-of-fit)

Les tests de la qualité de l'ajustement (ou goodness-of-fit) permettent d'identifier la copule qui s'ajuste le mieux aux données. Nous rappelons à cette étape que les données utilisées sont les résidus de sites géographiquement corrélés. Nous calculons les corrélations entre les résidus des sites afin d'identifier les sites corrélés deux à deux. Ensuite nous appliquerons les tests d'ajustement des copules aux innovations corrélées. Le test utilise dans son exécution la pseudo-vraisemblance des données (*Package développé par Bruno Rémillard*) Rémillard (2013). Nous ajustons les copules en appliquant deux tests :

1. Le test de Cramér-von Mises et de Kolmogorov-Smirnov décrit par Rémillard (2013, pages : 67, 95, 360). Ce test est basé sur la fonction empirique de Kendall.

2. Le test de Cramer-von Mises basé sur la transformée de Rosenblatt (Rémillard, 2013, page 430). Ce deuxième test permettra d'identifier de façon raisonnable, la copule qui décrit la structure de dépendance des innovations. « Avec la transformée Rosenblatt nous pouvons tester la qualité d'ajustement sur les variables avec une distribution donnée de sorte que si nous générons une pseudo-vraisemblance de ces variables aléatoires indépendantes et distribuées de manière uniforme sur

l'intervalle de 0 à 1, ces données ont la même distribution que les variables » (Rémillard, 2013, page 430, Traduction libre).

4.2.1. Algorithme de Cramer-von Mises

Rémillard (2013) a proposé des algorithmes pour réaliser les ajustements de copules selon la statistique de Cramer-Von Mises (notée S_n). Cette statistique s'applique aussi bien dans les cas univariés que dans les cas bivariés. Dans notre étude nous nous intéresserons essentiellement au cas bivarié. Selon Rémillard (2013), pour des variables indépendantes et identiquement distribuées dans l'intervalle [0, 1] sous l'hypothèse nulle, de grandes valeurs de S_n devraient entraîner le rejet de l'hypothèse nulle. Mais aussi, S_n dépend du paramètre thêta (θ) de la distribution. Le calcul du p-value de S_n se fait par la technique de rééchantillonnage paramétrique (appelé aussi *parametric bootstrap*). Ainsi l'ajustement des copules que proposent Rémillard (2013, page 325) est fait suivant la statistique de Cramer-von Mises et Kolmogorov-Smirnov basée sur le coefficient empirique de Kendall. Nous allons donc nous inspirer de cette application pour ajuster les copules à nos innovations.

4.2.2. Coefficient de corrélation de Kendall

Comme nous l'avons mentionné plus haut, le coefficient de corrélation de Kendall (ou le tau de Kendall) est l'un des paramètres dans l'analyse de la dépendance entre des variables aléatoires selon la statistique de Cramér-von Mises. Elles donnent une mesure de la corrélation entre les rangs des observations, à la différence du coefficient de corrélation linéaire qui lui apprécie la corrélation entre les valeurs des observations. Elles offrent par ailleurs l'avantage de s'exprimer simplement en fonction de la copule associée au couple de variables aléatoires. Selon Genest et MacKay (1986) pour une copule archimédienne de générateur φ, le tau de Kendall s'écrit comme suit :

$$\tau = 1 + 4\int_0^1 \frac{\varphi(u)}{\varphi'(u)}du.$$

Le Tableau 4.2 résume les expressions analytiques du coefficient de corrélation de Kendall pour différentes familles de copules bivariées usuelles. Pour la copule de Franck, le coefficient de corrélation de Kendall se définit selon la formule suivante :

$$\tau = 1 - \frac{4}{\theta} + \frac{4}{\theta^2}\int_0^{\theta} \frac{u}{e^u - 1} du.$$

L'expression $D_1(\theta)$, dans le tableau 4.2, est connue sous le nom de la fonction Debye. Elle s'exprime de la façon suivante : $D_1(\theta) = \frac{1}{\theta}\int_0^{\theta} \frac{u}{e^u - 1} du.$

Tableau 4.2 : Coefficients de corrélation théorique de Kendall (ou tau de Kendall) associés aux copules (Gaussienne, Student, Clayton, Gumbel et Frank).

Copule	**Tau de Kendall**
Gaussienne ($\rho \in [-1, 1]$)	$\frac{2}{\pi}\arcsin(\rho)$
Student-t ($\rho \in [-1, 1]$, $\nu > 2$)	$\frac{2}{\pi}\arcsin(\rho)$
Clayton ($\theta > 0$)	$\frac{\theta}{\theta + 2}$
Gumbel ($\theta \geq 0$)	$1 - \frac{1}{\theta}$
Frank ($\theta \neq 0$)	$1 - \frac{4}{\theta} + 4\frac{D_1(\theta)}{\theta}$

4.2.3. Application des copules

L'application des tests de la qualité d'ajustement des copules sera réalisée sur les résidus générés à partir des processus ARMA représentant chacun des sites étudiés. Mais avant, nous estimons les corrélations entre les sites à partir des résidus des séries chronologiques estimées. Cette étape présentera les sites qui ont un lien géographique selon l'expression du coefficient de corrélation entre les résidus. Ensuite des tests d'ajustement des copules seront appliqués sur certains sites corrélés. L'objectif principal de la section est d'identifier la copule qui décrit la structure de dépendance

entre les sites corrélés. Les copules testées dans cette section sont celles préalablement spécifiées au chapitre 3, c'est-à-dire Clayton, Gumbel, Frank, la Gaussienne et Student. Les méthodes d'ajustement des copules appliquées s'inspirent des algorithmes proposés par Rémillard (2013). Il s'agit de tester en premier, les copules suivant la statistique de Cramer-von Mises (S_n) et de Kolmogorov-Smirnov (T_n) basée sur le coefficient de Kendall (Rémillard, 2013, page 95). En deuxième lieu, la statistique de Cramer-von Mises calculée avec la transformée de Rosenblatt ($S_n^{(B)}$) sera utilisée (Rémillard, 2013, pages 321 à 323). Dans l'exécution de ces tests, la fonction utilise la pseudo-vraisemblance des résidus (variables) pour estimer les paramètres des copules, les valeurs de (S_n) et de (T_n). L'estimation des p-values est effectuée avec un *bootstrap* paramétrique (*N=1000*).

Corrélations entre les résidus des processus séries chronologiques (ARMA)

Tableau 4.3 : Coefficients de corrélation de Spearman entre les résidus des processus ARMA estimés.

Résidus	rbond	rgood	rboul	rfort	rsiou	rpenn
rbond	1	**0.283**	-0.023	-0.021	0.117	0.090
rgood	**0.283**	1	-0.021	-0.047	0.029	0.080
rboul	-0.023	-0.021	1	0.034	**0.190**	-0.037
rfort	-0.021	-0.047	0.034	1	**0.193**	-0.033
rsiou	0.117	0.029	**0.190**	**0.193**	1	-0.091
rpenn	0.090	0.080	-0.037	-0.033	-0.091	1

Les coefficients de corrélations estimés dans le Tableau 4.3 ne sont pas élevés dans l'ensemble. Les sites qui ont des corrélations positives et relativement significatives sont entre autres, Bondville et Goodwin Creek (0.283), ensuite Fort Peck et Sioux Fall (0.193) et enfin Boulder et Sioux Fall (0.190). Nous dirons que ces sites ont une certaine dépendance géographique bien qu'ils soient situés dans des régions géographiques différentes aux États-Unis d'Amérique (Figure 3.1). Les tests

d'ajustement des copules seront donc appliqués aux innovations conjointes des sites dont les corrélations conjointes sont jugées intéressantes et significatives.

Test d'adéquation des copules avec Cramer-Von Mises et Kolmogorov-Smirnov

Les Tableaux 4.4, 4.5, et 4.6 présentent les tests d'adéquation pour trois groupes de deux sites sélectionnés. Il s'agit des sites de : Bondville et Goodwin Creek (corrélation = 0.283), Fort Peck et Sioux Fall (corrélation = 0.193), Boulder et Sioux Fall (corrélation = 0.190).

Tableau 4.4 : Résultats du test d'ajustement des copules de Clayton, Gumbel et Frank suivant Cramer-von Mises (S_n) et Kolmogorov-Smirnov (T_n) avec un seuil de rejet du p-value de 5% : sites de Bondville et Goodwin Creek.

Modèle	Thêta (estimé)	S_n T_n	p-value (%)
Clayton	0.3700	0.4096 1.1592	0 0.9
Frank	1.7129	0.1189 0.7123	7.9 28.4
Gumbel	1.2504	0.2530 1.0795	0.7 1.6

Tableau 4.5 : Résultats du test d'ajustement des copules de Clayton, Gumbel et Frank suivant Cramer-von Mises (S_n) et Kolmogorov-Smirnov (T_n) avec un seuil de rejet du p-value de 5% : sites de Fort Peck et Sioux Fall.

Modèle	Thêta (estimé)	S_n T_n	p-value (%)
Clayton	0.2238	0.1720 0.8825	4.2 11.2
Frank	1.1181	0.1560 0.8994	2.5 5.7
Gumbel	1.1721	0.2363 0.9373	1.5 7.8

Tableau 4.6 : Résultats du test d'ajustement des copules de Clayton, Gumbel et Frank suivant Cramer-von Mises (S_n) et Kolmogorov-Smirnov (T_n) avec un seuil de rejet du p-value de 5% : sites de Boulder et Sioux Fall.

Modèle	Thêta (estimé)	S_n T_n	P-value (%)
Clayton	0.1810	0.2914 1.3195	0.3 0.1
Frank	1.0575	0.0931 0.8402	15.4 9.7
Gumbel	1.1687	0.0770 0.6850	38.8 47.2

Au Tableau 4.4, nous observons que seule la copule de Frank s'ajuste aux innovations de Bondville et Goodwin Creek suivant S_n et T_n. Les tests rejettent les ajustements de Gumbel et Clayton suivant Cramer-von Mises. Les résultats du Tableau 4.5 montrent que le test rejette les trois copules pour les sites de Fort Peck et Sioux Fall. Le test de Kolmogorov-Smirnov ne rejette aucune des trois, mais leur p-value demeure faible. Quant au tableau 4.6, les copules de Frank et de Gumbel semblent présenter des résultats encourageant suivant S_n et T_n, mais la copule de Gumbel semble être la meilleure structure de dépendance qui s'ajuste aux innovations de Boulder et Sioux Fall en observant son p-value. La copule de Clayton est rejetée.

Test d'adéquation basés sur la transformée de Rosenblatt

Le test utilise le coefficient de Cramer-Von Mises avec la transformée de Rosenblatt ($S_n^{(B)}$) (Rémillard, 2013, page 325). Ce test utilise également la pseudo-vraisemblance des résidus pour estimer les paramètres des copules ajustées et détermine des p-values avec un *bootstrap* paramétrique (*N=1000*). Compte tenu du fait que le premier test basé sur la fonction empirique de Kendall semble rejeter la copule de Clayton, le deuxième test se fera avec les copules de Student et la Gaussienne. Les mêmes variables (résidus des sites corrélés) seront utilisées pour le test.

Tableau 4.7 : Résultats du test d'ajustement des copules Gaussienne et Student basé sur la transformation de Rosenblatt ($S_n^{(B)}$) avec un seuil de rejet du p-value à 5% : sites de Bondville et Goodwin Creek.

Modèle	Paramètres estimés	$S_n^{(B)}$	p-value (%)
Gaussien	$\rho = 0.2870$	0.0811	1.5
Student	$(\rho, \nu) = (0.2983, 6.5483)$	0.0326	45.2

Tableau 4.8 : Résultats du test d'ajustement des copules Gaussienne et Student basé sur la transformation de Rosenblatt ($S_n^{(B)}$) avec un seuil de rejet du p-value à 5% : sites de Fort Peck et Sioux Fall.

Modèle	Paramètres estimés	$S_n^{(B)}$	p-value (%)
Gaussien	$\rho = 0.1893$	0.0679	5.1
Student	$(\rho, \nu) = (0.2044, 6.8317)$	0.0510	7.9

Tableau 4.9 : Résultats du test d'ajustement des copules Gaussienne et Student basé sur la transformation de Rosenblatt ($S_n^{(B)}$) avec un seuil de rejet du p-value à 5% : sites de Boulder et Sioux Fall.

Modèle	Paramètres estimés	$S_n^{(B)}$	p-value (%)
Gaussien	$\rho = 0.1975$	0.0471	19.2
Student	$(\rho, \nu) = (0.17, 14.826)$	0.0359	40.8

4.2.4. Analyse des résultats

Le premier test d'ajustement des copules de Clayton, de Gumbel et de Frank suivant Cramer-Von Mises (S_n) et Kolmogorov-Smirnov (T_n) basé sur la fonction empirique de Kendall n'offre pas de famille de copule s'adaptant à tous les sites

(Tableaux 4.4, 4.5, et 4.6). Pour le même ensemble de données, le deuxième test avec la transformée de Rosenblatt ($S_n^{(B)}$), suggère que la copule de Student s'ajuste aux données des trois paires de sites (Tableaux 4.7, 4.8 et 4.9). Ce constat, nous le faisons en observant les p-values de l'ajustement de Student qui sont supérieurs au seuil de 5%.

Pour nous assurer que la copule de Student définit bien la structure de dépendance de tout l'ensemble des sites corrélés (corrélation élevée, faible ou même négative), il faudrait appliquer la même méthodologie que les précédents tests d'ajustement à la copule en six dimensions, ce qui s'avérerait très long. L'approche que nous proposons consiste plutôt à adopter un modèle de copule en vigne (Vine-Copula) pour décrire la structure de dépendance de l'ensemble des innovations d'une dimension égale à six. Ce modèle offre une grande flexibilité, et permet entre autres d'adopter des marges bivariées de loi Student.

4.3. La méthode des copules en vigne (ou Vine-Copulas)

4.3.1. Généralités

Allen & *coll.* (2013) rappellent dans leur étude que les copules sont des distributions multivariées avec des marges normalisées qui captent les structures de dépendance des variables aléatoires. Selon ces chercheurs, les copules offrent une grande flexibilité dans la construction de modèles stochastiques multivariées. Mais, la remarque qu'ils formulent est que la modélisation avec un ensemble de plusieurs variables, comme les nôtres par exemple (ensemble de six variables), pose quelques limitations. Ces limitations peuvent être : la restriction de paramètres et la précision du modèle estimé. Les copules bidimensionnelles standard peuvent donc être non flexibles en grandes dimensions (dimension > 2). Allen & *coll.* (2013) proposent une alternative qui est l'utilisation les copules en vigne (Vine-Copulas) pour ces types de données de plus grandes dimensions.

4.3.2. Particularités des copules en vigne

Allen & *coll.* (2013) décrivent une vigne comme un outil graphique pour identifier des contraintes dans les distributions de grande dimension. Mais en parcourant la littérature, nous constatons que les copules en vigne sont également flexibles à construire et elles apportent une simplification dans la mise en œuvre de modèles stochastiques de dépendance multivariée. C'est Joe (1996) qui a proposé initialement la méthode de copules en vigne mais Bedford et Cooke (2001, 2002) ainsi que Kurowicka et Cooke (2006) ont vulgarisé la méthode en apportant plus de détails. Ainsi les copules en vigne, décrites par Bedford et Cooke (2001, 2002) sont des modèles graphiques flexibles pour décrire des copules multivariées construites en utilisant une cascade de copules bivariées ou des copules par paires. Les travaux d'Allen & *coll.* (2013) démontrent aussi que la percée statistique des copules en vigne est due aux travaux des chercheurs Aas & *coll.* (2009) qui décrivent dans l'un de leurs articles paru en 2009, les techniques d'inférence statistique pour deux classes de construction de dépendance par les copules en vigne. Ces deux classes sont les « C-Vines » et « D-Vines ». Ces chercheurs ont montré que les deux classes appartiennent à une autre classe plus générale qui est la vigne régulière (ou R-Vine). La R-Vine peut être utilisée pour un modèle théorique graphique afin de déterminer les paires de copules incluses dans la décomposition. C'est une méthode pour laquelle toutes les contraintes sont bidimensionnelles (Allen & *coll.*, 2013).

4.3.3. Quelques rappels théoriques

Densité d'une copule bivariée

Si nous revenons au théorème de Sklar définit à la section 4.1.1, nous pouvons définir la densité d'une copule bidimensionnelle de la façon suivante :

$$c_{12}(x_1, x_2) = \frac{\partial^2 C_{12}(x_1,x_2)}{\partial x_1 \partial x_2}.$$

Ceci implique, respectivement selon la littérature, une densité jointe et une densité conditionnelle suivante :

$f(x_1, x_2) = c_{12}\big(F_1(x_1), F_2(x_2)\big) f_1(x_1) f_2(x_2).$

$f(x_2|x_1) = c_{12}\big(F_1(x_1), F_2(x_2)\big) f_2(x_2).$

Modèle R-Vine

Selon Aas & coll. (2009) la densité conjointe d'un système multivarié de n-variables $x_1, \ldots, x_n$ se factorise en densités conditionnelles successives, comme :

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_1, x_2) \ldots f(x_n|x_1, \ldots, x_{n-1}).$$

Chaque facteur de ce produit peut encore être décomposé en une fonction de densité bivariée et une densité marginale conditionnelle (Aas & coll., 2009). Mais la construction de paires de copules est itérative et plusieurs schémas de décomposition peuvent être obtenus à partir de différentes factorisations de la densité conjointe. Aas & coll. (2009) précisent que pour une distribution de n dimensions, il aurait $n! = 1*2*3*\ldots*(n-1)*n$ différentes combinaisons de densités d'une « R-Vine ». Ces chercheurs ont montré dans leur étude que la densité d'une vigne régulière pour une telle distribution est composée de densités des classes « D-Vines » et « C-Vine ». Aas & coll. (2009) ont illustré ces densités d'une « R-Vine » en s'appuyant sur les travaux de Bedford et Cooke (2001b) et de Kurowicka et Cooke (2006) qui résument toutes les combinaisons de paires de copules permettant de construire la densité multivariée de la « R-Vine ». Aas & *coll.* (2009) présentent la forme algébrique de la densité $f(x_1, \ldots, x_n)$ de n-dimensions (n>2) d'une « R-Vine » comme le produit de densités des paires de copules conditionnelles, définit précédemment, et des densités marginales pour les classes « D-Vine » et « C-Vine ».

Densité d'une « D-Vine » :

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_k(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,(i+j)|(i+1),\ldots,(i+j-1)} \left\{F\left(x_i \middle| x_{i+1}, \ldots, x_{i+j-1}\right), F(x_{i+j}|x_{i+1}, \ldots, x_{i+j-1})\right\}.$$

Densité d'une « C-Vine » :

$$f(x_{1,}\dots,x_n) = \prod_{k=1}^{n} f_k(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,(i+j)|1,\dots,(j-1)} \{F(x_j|x_1,\dots,x_{j-1}), F(x_{j+i}|x_1,\dots,x_{j-1})\}.$$

La première partie de ces équations représente le produit des densités marginales et la seconde partie représente le produit des densités des paires de copules conditionnelles. Bedford et Cooke (2001) ont introduit une forme graphique de la vigne régulière « R-Vine » pour structurer la présentation de l'expression analytique du modèle. Pour Bedford et Cooke (2001), une « R-Vine » de dimension *d* est une construction en séquences de *d-1* arbres. Le premier arbre possède *d* nœuds et *d-1* densités de paire de copules entre les nœuds. Au nœud *j* quelconque, il aura *d+1-j* nœuds et *d-j* densités conditionnelles de paires de copules. Pour illustrer ces notions de densité et d'arbre, nous prenons premièrement une distribution d'un ensemble de variables à 3-dimension (*d=3*).

Exemple *(Aas & coll., 2009 et Kramer & Schepsmeier, 2011)* ***:*** *Nous prenons d = 3 dimensions. En nous référant à Aas & coll. (2009), il aurait 3! (c'est-à-dire 6) combinaisons possibles de la densité de cette distribution. Une décomposition possible de la fonction de densité* $f(x_{1,}x_{2,}x_3)$ *est :*

$f(x_{1,}x_{2,}x_3) = f(x_3|x_1,x_2)f(x_2|x_1)f_1(x_1).$

$f(x_2|x_1) = c_{12}(F_1(x_1),F_2(x_2))f_2(x_2).$

$f(x_3|x_1,x_2) = c_{13|2}(F(x_1|x_2),F(x_3|x_2))f(x_3|x_2).$

$f(x_3|x_2) = c_{23}(F_2(x_2),F_3(x_3))f_3(x_3).$

En réécrivant l'expression de la densité conjointe, avec respectivement les densités marginales, les densités des paires non conditionnelles et les densités des paires conditionnelles, nous arrivons à la forme suivante :

$f(x_{1,}x_{2,}x_3) = f_3(x_3)f_2(x_2)f_1(x_1)$ *(Marginales).*

$* \; c_{12}(F_1(x_1),F_2(x_2))c_{23}(F_2(x_2),F_3(x_3))$ *(Non conditionnelles).*

$* \ c_{13|2}\big(F(x_1|\ x_2), F(x_3|x_2)\big)$ *(Conditionnelles).*

Une représentation graphique pour une « R-Vine » de l'exemple précédent peut s'illustrer comme le graphe de la figure 4.3. Nous observons ici deux arbres. Le premier arbre est composé de trois nœuds (1, 2 et 3) qui représentent les variables (x_1, x_2, x_3). Les nombres sur les arcs sont les paires de copules c_{12} et c_{23}. Le deuxième arbre est composé de deux nœuds et un arc. La paire de copules conditionnelles est représentée sur l'arc et les nœuds sont les pseudo-vraisemblances des variables conditionnelles (Kramer et Schepsmeier, 2011).

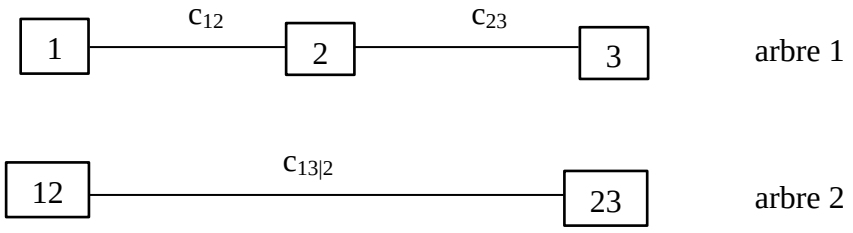

Figure 4.3 : Arbres d'ajustement de copules sur les variables de dimension égale à trois (d = 3) avec une « R-Vine » selon Kramer & Schepsmeier (2011).

Si nous considérons une distribution à six dimensions, nous aurons selon 6! (c'est-à-dire 720) combinaisons possibles pour exprimer la densité de cette distribution avec une « R-Vine ». Dans ce qui suit, nous illustrons une combinaison de la densité d'une distribution à six dimensions, en nous inspirant des travaux des chercheurs Aas & coll. (2009). L'illustration de la densité d'un tel ensemble de six dimensions correspondrait à celle de notre ensemble de variables (c'est-à-dire les six marginales des sites faisant l'objet de notre étude).

La densité conjointe *f(x_1, x_2, x_3, x_4, x_5, x_6) peut être de la forme :*

$$\boldsymbol{f(x_1, x_2, x_3, x_4, x_5, x_6)} = f_6(x_6) * f_5(x_5) * f_4(x_4) * f_3(x_3) * f_2(x_2) * f_1(x_1) *$$
$$c_{12}\{F_1(x_1), F_2(x_2)\} * c_{23}\{F_2(x_2), F_3(x_3)\} * c_{34}\{F_3(x_3), F_4(x_4)\} *$$
$$c_{35}\{F_3(x_3), F_5(x_5)\} * c_{36}\{F_3(x_3), F_6(x_6)\} * c_{13|2}\{F(x_1|x_2), F(x_3|x_2)\} *$$
$$c_{24|3}\{F(x_2|x_3), F(x_4|x_3)\} * c_{25|3}\{F(x_2|x_3), F(x_5|x_3)\} * c_{26|3}\{F(x_2|x_3), F(x_6|x_3)\} *$$

$c_{14|23}\{F(x_1|x_2,x_3),F(x_4|x_2,x_3)\} * c_{15|23}\{F(x_1|x_2,x_3),F(x_5|x_2,x_3)\} *$
$c_{16|23}\{F(x_1|x_2,x_3),F(x_6|x_2,x_3)\} * c_{45|123}\{F(x_4|x_1,x_2,x_3),F(x_5|x_1,x_2,x_3)\} *$
$c_{56|123}\{F(x_5|x_1,x_2,x_3),F(x_6|x_1,x_2,x_3)\} *$
$c_{46|1235}\{F(x_4|x_1,x_2,x_3,x_5),F(x_6|x_1,x_2,x_3,x_5)\}.$

Cette expression est l'une des 720 combinaisons possibles de la densité conjointe d'une distribution à 6-dimensions. Nous verrons un graphique des arbres en appliquant la méthode à nos variables.

4.3.4. Méthode d'ajustement d'une vigne régulière (ou R-Vine)

La méthodologie que nous proposons pour ajuster une méthode « R-Vine » aux résidus des processus ARMA estimés se présente suivant trois points :

- Les résidus des séries estimées seront transformées avec la fonction de répartition de la loi normale pour obtenir des marginales uniformes dans l'intervalle [0, 1]. Notre matrice de résidus est constituée de six colonnes (correspondant au nombre de sites) et de N lignes (avec N = 1800 dans notre étude).

- Nous choisissons ensuite un ensemble de familles de copules bivariées. Nous avons choisi comme familles de copules : la Gaussienne (1), la Student (2), Clayton (3), Gumbel (4), Frank (5) et la copule de Joe (6). Les chiffres indiqués entre parenthèse en face de chaque copule représente son identifiant dans la fonction R du « package : *VineCopula* ».

- Ensuite, nous utilisons deux fonctions R dans notre programme pour estimer, respectivement, le modèle « R-Vine » et le graphique des arbres. Ces fonctions R ont été développées par Dissmann & *coll.* (2013) et Brechmann (2013). Une description brève de ces fonctions est donnée à la page suivante.

La fonction « *RVineStructureSelect () »* procède à la sélection de la structure des paires de copules en fonction des familles de copules prédéfinies. Par exemple, *library (VineCopula): {MODEL <- RVineStructureSelect (data, familyset, type=0, selectioncrit="AIC", indeptest=FALSE, level=0.05, trunclevel, progress=TRUE, weights)}.*

La fonction *RVineTreePlot* produit des graphiques des arbres et des paramètres souhaités des copules. Par exemple, *library (VineCopula): {RVineTreePlot (data, MODEL=MODEL, method="mle", tree="ALL", edge.labels)}*

Tableau 4.10 : Légende des paramètres de quelques fonctions de la bibliothèque de fonctions *VineCopula*.

data	Matrice de données avec les marginales uniformes
familyset	Vecteur d'entiers pour les familles de copule ou paires de copules
type	Type de copule en vigne (0 pour R-Vine par défaut)
selectioncrit	Critère de sélection de la copule « AIC » par défaut
indeptest	Test d'indépendance. « FALSE » par défaut
level	Niveau d'importance du test d'indépendance (5%)
trunclevel	Niveau de troncature dabs la progression des arbresProgression
progress	normale des arbres si égale à « TRUE »
weights	Poids affectés aux observations
method	Méthode d'estimation « mle » (maximum de vraisemblance)
tree	Affichage des arbres si l'expression égale à « ALL »
edge.labels	Vecteur de paramètres selon la famille de copule estimée

Le tableau 4.10 présente une légende des paramètres des fonctions illustrées ci-haut.

4.3.5. Application de la méthode « R-Vine »

Les familles de copules choisies pour l'application de la méthode des copules en vigne (ou Vine-Copulas) dans ce mémoire sont : la Gaussienne, la copule de

Student, la copule de Clayton, la copule de Frank, la copule de Gumbel et celle de Joe. Notre matrice de données contient les résidus de six sites, une légende des acronymes utilisés pour identifier les sites se trouve au tableau 3.23. La dimension détermine le nombre total d'arbres à construire. Pour six variables, nous aurons cinq arbres à construire (section 4.3.2). Les arbres sont illustrés sur les figures 4.4, 4.5, 4.6, 4.7 et 4.8. On y retrouve également le nom des copules et les coefficients empiriques de Kendall associé aux copules sur les arcs des arbres. Nous avons également les paires de variables formées dans les nœuds rectangulaires des arbres. Les paramètres des copules sont compilés dans le tableau associé à chaque arbre. Nous observons sur l'arbre 1 à la figure 4.4, cinq densités de copule exprimées entre les résidus. Nous avons trois densités de la copule de Student, une Gaussienne et une de Joe.

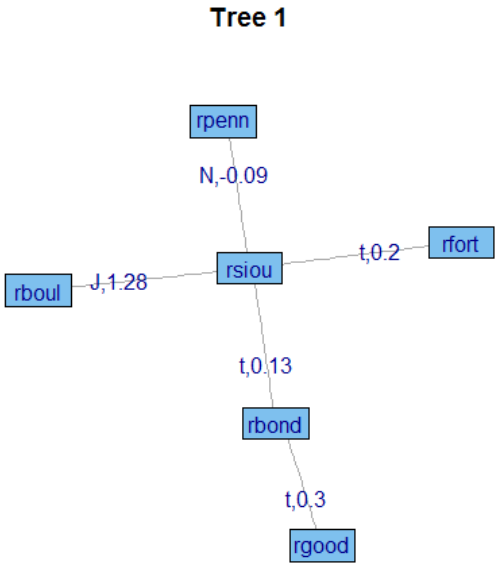

Figure 4.4 : Arbre 1 du modèle « R-Vine » estimé. Les variables sont dans les nœuds et sur les arcs (ou branches), sont représentées les copules et leur paramètre.

Tableau 4.11 : Paramètres estimés de l'arbre 1. Les paires de variables sont formées, on y trouve les copules et les coefficients théoriques de Kendall.

Paires constituées	Copule	θ	ν	K_{nt}
(rsiou, rfort)	Student	0.20	4.31	0.13
(rsiou, rboul)	Joe	1.28	-	0.14
(rsiou, rpenn)	Normale	-0.09	-	-0.05
(rsiou, rbond)	Student	0.13	4.95	0.08
(rbond, rgood)	Student	0.30	3.65	0.19

L'arbre 2 de la figure 4.5 se caractérise par deux densités de copules de Student, une densité de la copule de Clayton et une Gaussienne.

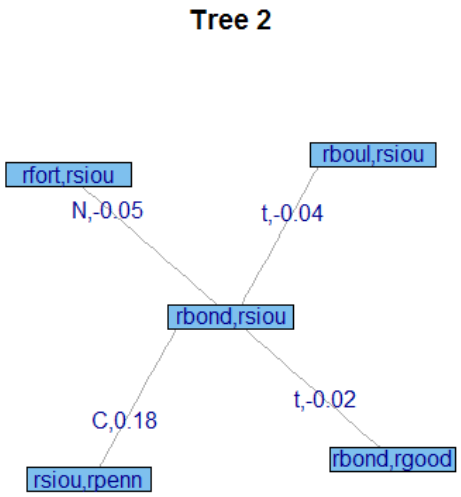

Figure 4.5 : Arbre 2 du modèle « R-Vine » estimé. Les paires de variables sont dans les nœuds et sur les arcs (ou branches), sont représentées les copules et leur paramètre.

Tableau 4.12 : Paramètres estimés de l'arbre 2. Les paires de variables sont combinées, on y trouve les copules et les coefficients théoriques de Kendall.

Paires constituées	Copule	θ	ν	K_{nt}
(rbond, rsiou) – (rsiou, rpenn)	Clayton	0.18	-	0.08
(rbond, rsiou) – (rfort, rsiou)	Normale	-0.05	-	-0.03
(rbond, rsiou) – (rboul, rsiou)	Student	-0.04	10.49	-0.02
(rbond, rsiou) – (rbond, rgood)	Student	-0.02	9.36	-0.01

La figure 4.6 représente l'arbre 3 avec trois copules de densité conditionnelle. Il s'agit des copules de Clayton, Student et la Gaussienne.

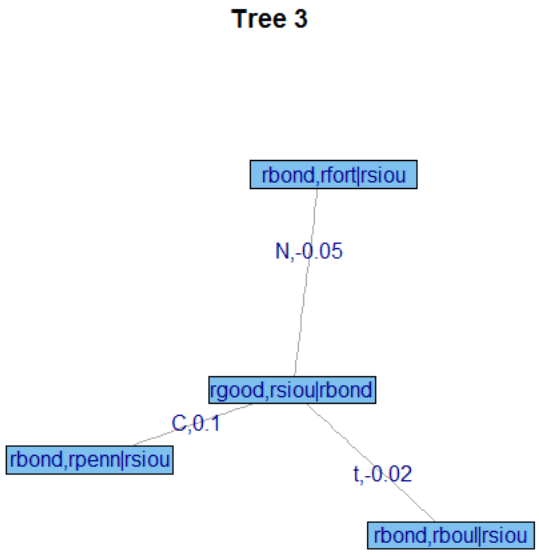

Figure 4.6 : Arbre 3 du modèle « R-Vine » estimé. Les paires de variables conditionnelles sont dans les nœuds et sur les arcs (ou branches), sont représentées les copules et les coefficients empiriques de Kendall associés.

Tableau 4.13 : Paramètres estimés de l'arbre 3. Les paires de variables conditionnelles sont combinées, on y trouve les copules et les coefficients théoriques de Kendall.

Paires constituées	Copule	θ	ν	K_{nt}
(rgood, rsiou|rbond) – (rbond, rboul|rsiou)	Student	-0.02	23.12	-0.01
(rgood, rsiou|rbond) – (rbond, rfort|rsiou)	Normale	-0.05	-	-0.03
(rgood, rsiou|rbond) – (rbond, rpenn|rsiou)	Clayton	0.1	-	0.05

L'arbre 4 de la figure 4.7, se caractérise par deux copules de Frank qui lient les paires conditionnelles de la pseudo-vraisemblance des résidus.

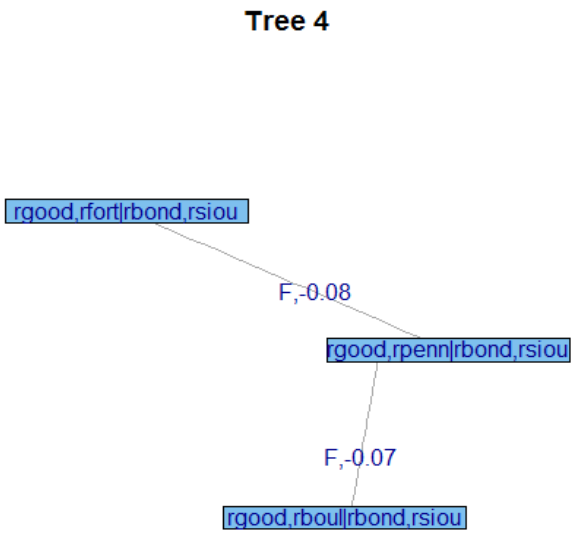

Figure 4.7 : Arbre 4 du modèle « R-Vine » estimé. Les paires de variables conditionnelles sont dans les nœuds et sur les arcs (ou branches), sont représentées les copules et leur paramètre.

Tableau 4.14 : Paramètres estimés de l'arbre 4. Les paires de variables conditionnelles sont formées, on a les copules et les coefficients théoriques de Kendall.

Paires constituées	Copule	θ	ν	K_{nt}
(rgood, rfort|rbond, rsiou) – (rgood, rpenn|rbond, rsiou)	Frank	-0.08	-	-0.01
(rgood, rpenn|rbond, rsiou) – (rgood, rboul|rbond, rsiou)	Frank	-0.07	-	-0.01

Le dernier arbre à la figure 4.8, présente la paire conditionnelle des pseudo-vraisemblances des résidus. La copule liante est celle de Student.

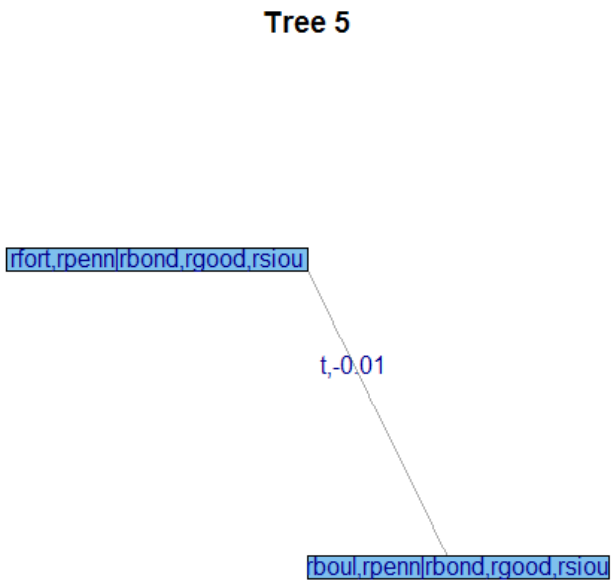

Figure 4.8 : Arbre 5 du modèle « R-Vine » estimé. Les paires de variables conditionnelles sont dans les nœuds et sur l'arc (ou branche), on y trouve la copule « R-Vine » et son paramètre thêta.

Tableau 4.15 : Paramètres estimés de l'arbre 5. La paire de variables conditionnelles formée, on y trouve la copule « R-Vine » et le coefficient théorique de Kendall.

Paire constituée	Copule	θ	ν	K_{nt}
(rboul, rpenn|rbond, rgood, rsiou) – (rfort, rpenn|rbond, rgood, rsiou)	Student	-0.01	8.32	0

4.3.6. Discussion des résultats de modèles R-vine

Il est à observer que la méthode « R-Vine » appliquée aux résidus des processus ARMA, combine les densités des copules prédéfinies pour construire des structures de dépendance conditionnelle, et cela jusqu'au dernier arbre. Ainsi, sur le dernier arbre (arbre 5), nous avons la copule de Student qui combine les structures de dépendance conditionnelle du modèle. Mais à partir dès l'arbre 1, nous observons différentes densités de copules entre les paires simples et les paires conditionnelles de pseudo-vraisemblance des variables. Nous remarquons que la méthode établie dès le premier arbre une structure de dépendance entre les résidus des sites corrélés qui correspond à celle établie suivant la statistique de Cramer-von Mises basée sur la transformation de Rosenblatt ($S_n^{(B)}$) (section 4.2.3), ce qui renforce la justesse de ce test.

Il est aussi possible de déterminer des sous-modèles « R-Vine » en procédant à des troncatures dans l'estimation du modèle. Les troncatures peuvent se faire en spécifiant dans la fonction R le nombre d'arbres que nous souhaitons construire. Nous avons ainsi estimé deux sous-modèles en faisant deux troncatures, une troncature à l'arbre 2 et l'autre à l'arbre 3 où nous avons une certaine variété de paires de copules. Le modèle de dépendance spatiale ainsi estimé par la « R-Vine », nous allons simuler, au chapitre 5, une production d'énergie solaire à partir de la normale multivariée d'une part et des modèles « R-Vine » estimés d'autre part afin de comparer leur niveau d'ajustement.

Chapitre 5 : Simulation et comparaison de modèles

Nous avons établi la structure de dépendance spatiale des sites en ajustant, aux résidus des sites, la « R-Vine ». Nous simulons maintenant les productions d'énergie solaire à partir des modèles de série chronologique développés au chapitre 3 couplés par une « R-Vine » ou une normale multivariée afin de comparer les résultats. Dans ce chapitre 5, nous cherchons à savoir si les copules permettent une meilleure estimation du risque comparativement à la normale multivariée.

5.1. Méthodologie

Nous générons des données de la production d'énergie solaire (kWh/m^2) pour une année choisie dans la période 2007 à 2011. Pour générer ces données, des simulations sont réalisées à partir de la normale multivariée et des modèles « R-Vine » pour produire des variables pseudo-aléatoires qui représenteront les résidus des sites étudiés. Ensuite des calculs itératifs sont faits à partir des expressions analytiques des processus stochastique ARMA pour déterminer la production d'énergie pour chaque site. Des conditions initiales sont fixées, bien évidemment, pour initialiser le calcul d'énergie. À l'aide de la fonction de répartition empirique et d'un graphique « boxplot », il sera ainsi simple de comparer les modèles. Les moyennes, les variances, les écarts type et l'intervalle de prévision (à 75% et 95%) des valeurs générées sont estimés.

L'année 2009 est choisie pour simuler la production d'énergie. Nous estimons que cette période qui se situe à deux années de 2007 et de 2011 serait une bonne période. Deux fonctions R « *RVineSim* » et « *mvrnorm* » développées respectivement par Dissmann *& coll.* (2013) et Ripley (1987) servent à simuler les variables pseudo-aléatoires des résidus. Les « packages » du logiciel R associes à ces fonctions sont : « VineCopula » et « MASS ».

5.2. Simulation de la production d'énergie solaire.

Comme indiqué dans la méthodologie, nous simulons une production d'énergie solaire par site à partir des modèles « R-Vine » estimés et de la normale multivariée. Les premières illustrations présentent la production d'énergie simulée pour une année et comparée aux données mesurées et à la courbe du modèle théorique. Nous considérons différents scénarios de déploiement et simulons la production quotidienne totale. Les premiers scénarios comportent deux sites choisis parmi les six. Une analyse est faite pour comparer les modèles « R-Vine » et la normale multivariée autant pour une production journalière qu'une production mensuelle. À titre comparatif, le scénario où la corrélation géographique est nulle a aussi été généré. Nous portons ensuite un regard sur la production du minimum et du maximum obtenue par un déploiement de puissance égale en chacun des six sites. En plus de la production totale, nous étudions aussi la distribution des sites qui présentent une corrélation significative. Le tableau 5.1 comporte ces sites concernés (extrait du tableau 4.3, section 4.2.3).

Une comparaison graphique à l'aide du « boxplot » et de la fonction de répartition empirique des données simulées permet de visualiser les différences apparentes entre les modèles. Lorsque les courbes sont rapprochées ou se superposent, il devient très difficile de déterminer les écarts. Un tableau contenant la valeur de la distribution en différents points dans les queues est présenté afin de comparer les différences subtiles entre les ajustements.

Tableau 5.1 : Corrélations significatives entre les sites.

Sites	Bondville	Boulder	Fort Peck
Goodwin Creek	**0.283**	*-0.021*	*-0.047*
Sioux Fall	*0.117*	**0.190**	**0.193**

5.2.1. Production d'énergie solaire annuelle

Nous présentons la production annuelle d'énergie simulée à partir de la « R-Vine » et de la normale multivariée pour deux sites choisis : le site de Bondville dans l'Illinois et le site de Boulder dans le Colorado. L'année de simulation est 2009. Sur la figure 5.1 et la figure 5.2, nous comparons en a) la courbe théorique et les données mesurées, en b) la courbe théorique et la simulation à partir de la « R-Vine », en c) la courbe théorique et la simulation à partir de la normale et en d) la superposition des courbes théorique, « R-Vine » et la normale. La figure 5.3 quant à elle, présente la moyenne des productions annuelles simulées (moyenne de *10000* données simulées).

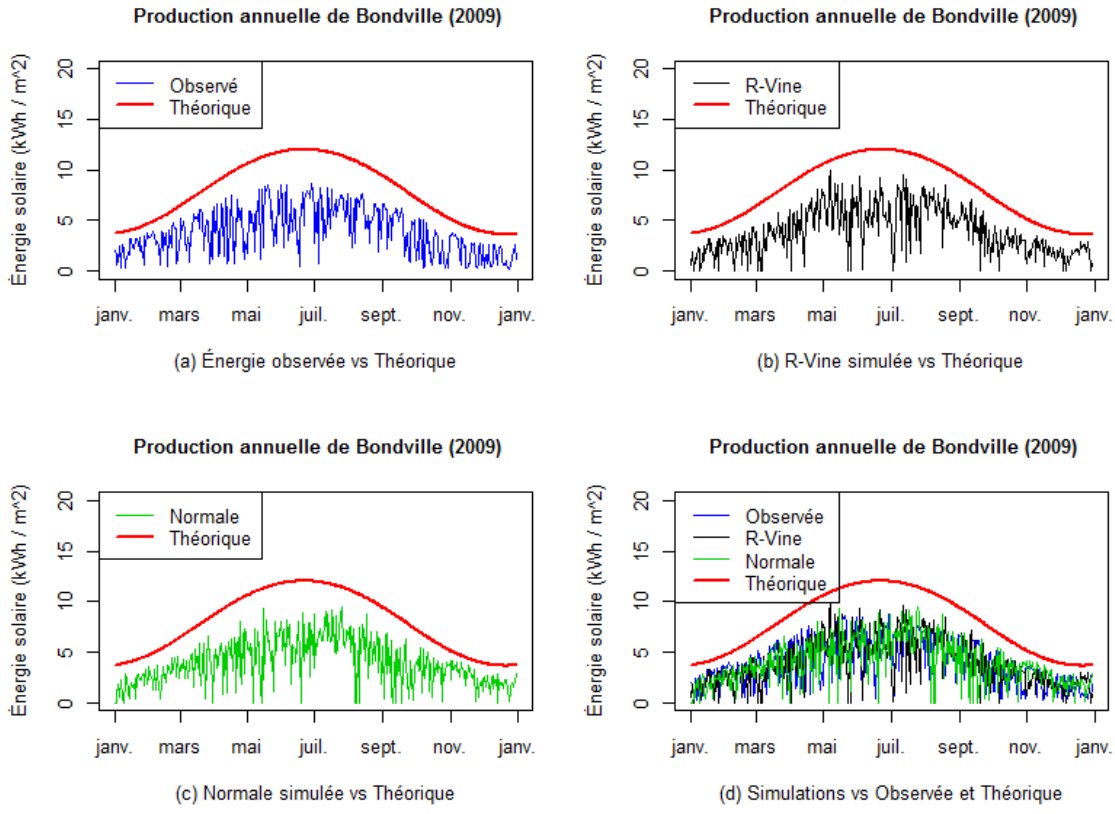

Figure 5.1 : Site de Bondville dans l'Illinois; comparaison de la courbe théorique, des données mesurées et des productions annuelles simulées à partir de la « R-Vine » et de la normale multivariée.

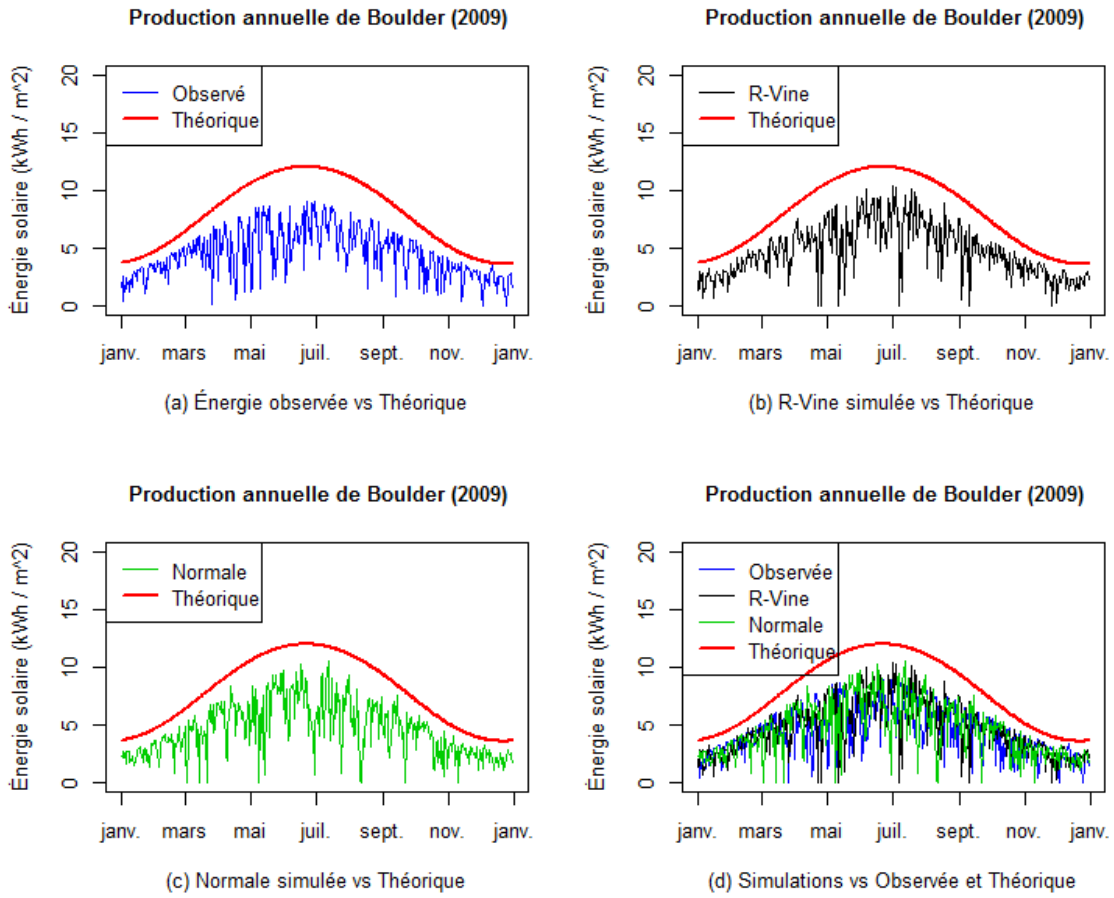

Figure 5.2 : Site de Boulder au Colorado; comparaison de la courbe théorique, des données mesurées et des productions annuelles simulées à partir de la « R-Vine » et de la normale multivariée.

La figure 5.3 montre par ailleurs que la moyenne des productions annuelles simulées à partir de la normale multivariée et du modèle de copules en vigne régulière « R-Vine » a une tendance similaire à celle des données mesurées sur les sites. Elle ne s'éloigne pas des mesures réelles comparativement à la courbe du modèle physique qui suit la même tendance que les deux autres courbes mais avec un certain écart. De façon générale les simulations de l'énergie à partir des deux modèles, c'est-à-dire la « R-Vine » et la normale multivariée semblent être cohérentes avec les mesures d'énergie réalisées sur les sites. En annexe B.3 à la page 111, nous avons simulé également la production annuelle de Sioux Fall et Penn State (voir la figure B.3.1 et la figure B.3.2).

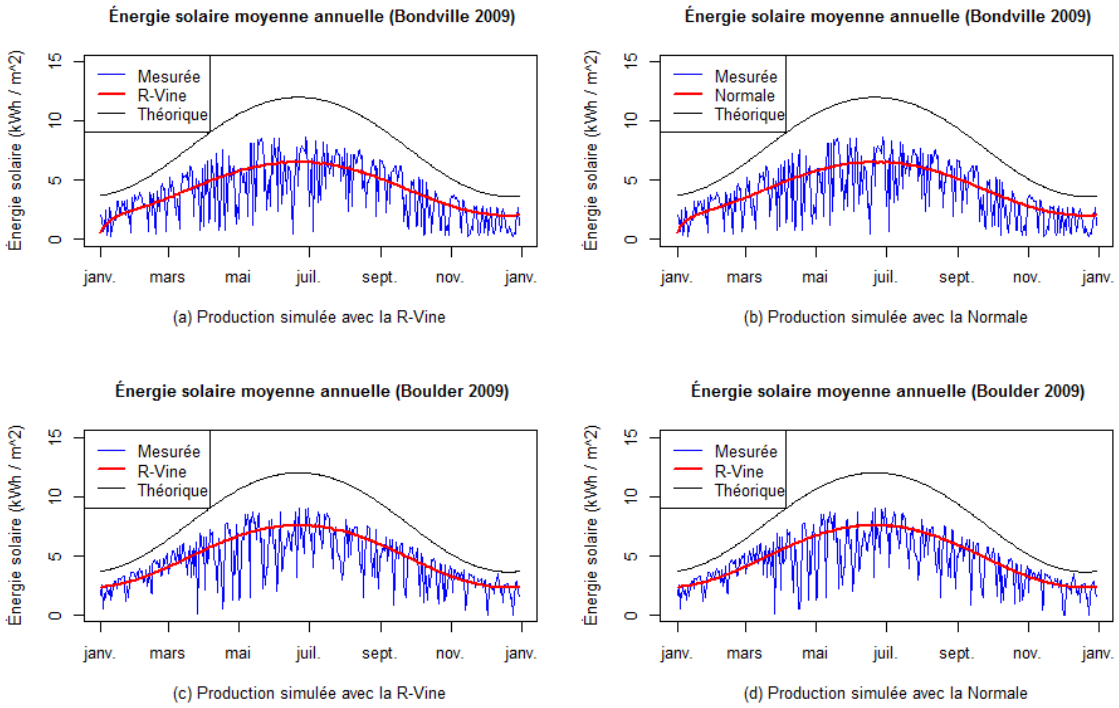

Figure 5.3 : Moyenne de *10000* données de production annuelle d'énergie solaire simulées à partir de la normale multivariée et de « R-Vine ».

5.2.2. Production d'énergie solaire pour une journée

Nous analysons dans cette section, trois groupes de deux sites corrélés. Ensuite, nous considérons l'ensemble des sites avec le minimum et le maximum des productions simulées à partir de la « R-Vine » et de la normale multivariée. La journée du 18 juin 2009 a été choisie pour simuler les données solaires. Les valeurs initiales d'énergie mesurée sur les sites à cette période de 2009 sont contenues dans le tableau 5.2. Rappelons que ce sont 10000 données d'énergie pour chaque mois de 2009 qui sont générées. Il est également possible de choisir une quelconque période pour établir les comparaisons.

Tableau 5.2 : Énergie mesurée sur les sites le 18 juin 2009 (données extraites de Surfrad).

18 juin 2009	**Bondville**	**Goodwin-Creek**	**Boulder**	**Fort-Peck**	**Sioux-Fall**	**Penn-State**
mesurée (kWh/m^2)	5.12	8.27	7.32	5.92	6.98	3.21

Données simulées des sites de Bondville et Goodwin-Creek au 18 juin

Tableau 5.3 : Moyenne, variance, et intervalle de prévision de la production du 18 juin 2009 basés sur 10000 répétitions pour les sites de Bondville et de Goodwin-Creek.

Modèle	**Moyenne (kWh/m^2)**	**Variance**	**Intervalle de prévision**		**Mesurée au 18 juin 2009 (kWh/m^2)**
			75%	**95%**	
R-Vine	6.19	3.33	4.08 – 8.18	1.81 – 9.09	
Normale	6.25	3.34	3.98 – 8.23	2.06 – 9.09	6.69
Indépendant	6.20	2.68	4.21 – 8.01	2.53 – 8.87	

La différence entre les modèles semble très subtile visuellement. La figure 5.4, montre clairement cette subtilité. Cependant la fonction de répartition du modèle indépendant semble différente de celle des deux autres modèles, c'est-à-dire les fonctions de répartition empirique de la vigne régulière « R-Vine » et de la normale multivariée. Cette différence est aussi observée pour les sites de Fort Peck et de Sioux-Fall ainsi que les sites de Boulder et de Sioux-Fall (figures 5.4, 5.5 et 5.6). Si l'on néglige de tenir compte de la corrélation géographique, il semble avoir un effet perceptible entre les modèles. La différence est moins visible pour les sites de Bondville et de Sioux Fall où la corrélation est faible (figure C, page 112).

Pour le modèle « R-Vine » et la normale multivariée, les différences sont difficilement perceptibles pour des valeurs d'énergie élevées. Mais pour de petites valeurs d'énergie, généralement < 2 kWh/m^2 (figures 5.4, 5.6 et 5.8), il apparaît une certaine différence. Un facteur d'écart est calculé entre les probabilités des observations pour les modèles « R-Vine » et la normale multivariée pour nous rendre compte de cette fine différence. Le facteur d'écart pour les petites valeurs peut se traduire par le rapport des probabilités *(RVine)/(Normale)*. Mais pour les grande valeurs, nous prenons le rapport *(1-RVine) / (1-Normale)*.

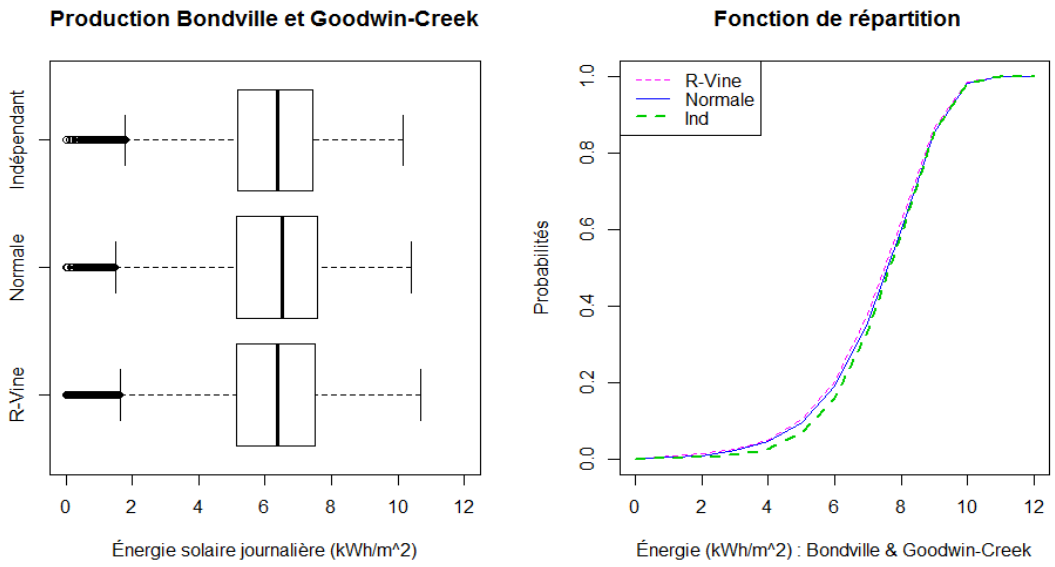

Figure 5.4 : Boxplot et fonction de répartition empirique de 10000 données simulées au 18 juin 2009 pour Bondville et Goodwin Creek.

Les données sont basées sur les mêmes modèles de série chronologique mais la corrélation spatiale est modélisée par la loi normale, la « R-Vine » ou l'indépendance.

Le tableau 5.4 qui suit, montre une différence entre la « R-Vine » et la normale multivariée pour des valeurs d'énergie inférieures à 2 kWh/m^2. Cela peut se traduire par le fait que la « R-Vine », comparativement à la normale, capte plus les petites valeurs d'énergie.

Tableau 5.4 : Probabilités empiriques de l'énergie produite un 18 juin pour les sites de Bondville et de Goodwin-Creek, basées sur 10000 répétitions selon trois modèles de dépendance spatiale (normale, R-Vine et indépendance).

Énergie solaire / Jour (kWh/m^2)	R-Vine	Normale multivariée	Indépendant	Facteur d'écart
Petites valeurs d'énergie				
< 1	0.0073	0.0004	0.0011	18.25
< 2	0.0147	0.0009	0.0039	16.33
< 3	0.0258	0.0230	0.0103	1.12
< 4	0.0500	0.0446	0.0265	1.12
Grandes valeurs d'énergie				
< 7	0.3764	0.3557	0.3335	0.97
< 8	0.6223	0.6037	0.5893	0.95
< 9	0.8677	0.8542	0.8544	0.91
< 10	0.9869	0.9838	0.9824	0.81

Données simulées des sites de Fort Peck et Sioux Fall au 18 juin

Tableau 5.5 : Moyenne, variance, et intervalle de prévision de la production du 18 juin 2009 basés sur 10000 répétitions pour les sites de Fort Peck et de Sioux Fall.

Modèle	Moyenne (kWh/m^2)	Variance	Intervalle de prévision		Mesurée au 18 juin 2009 (kWh/m^2)
			75%	95%	
R-Vine	6.54	2.52	4.72 – 8.25	2.87 – 9.12	
Normale	6.54	2.51	4.64 – 8.26	3.04 – 9.06	6.45
Indépendance	6.57	2.50	4.79 – 8.11	3.24 – 8.24	

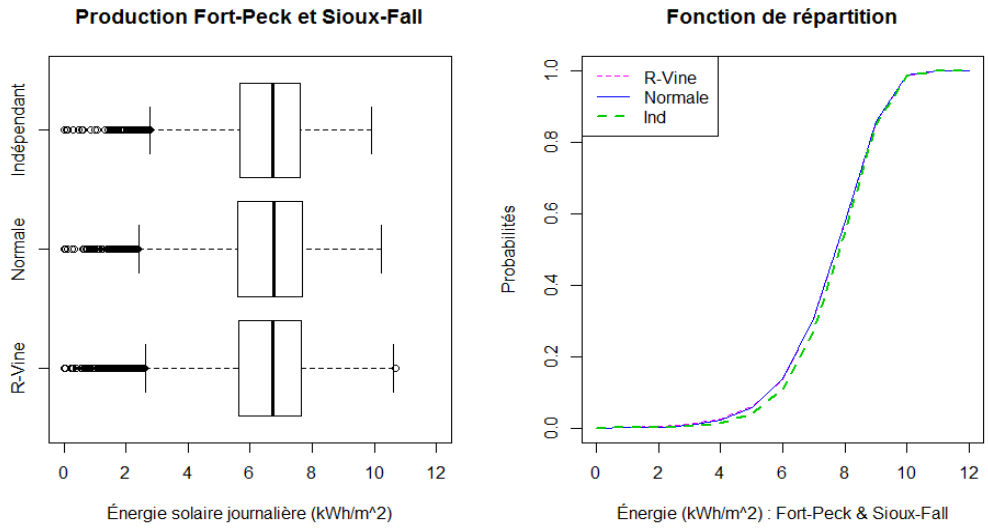

Figure 5.5 : Boxplot et fonction de répartition empirique de 10000 données simulées au 18 juin 2009 pour Fort Peck et Sioux Fall.

Tableau 5.6 : Probabilités empiriques de l'énergie produite un 18 juin pour les sites de Fort Peck et de Sioux Fall, basées sur 10000 répétitions selon trois modèles de dépendance spatiale (normale, « R-Vine » et indépendance).

Énergie solaire / Jour (kWh/m^2)	R-Vine	Normale multivariée	Indépendant	Facteur d'écart
Petites valeurs d'énergie				
< 1	0.0020	0.0009	0.0008	2.22
< 2	0.0053	0.0024	0.0015	2.20
< 3	0.0111	0.0071	0.0041	1.54
< 4	0.0263	0.0218	0.0127	1.21
Grandes valeurs d'énergie				
< 7	0.3046	0.3078	0.2691	1.00
< 8	0.5798	0.5785	0.5500	0.99
< 9	0.8548	0.8580	0.8496	1.02
< 10	0.9869	0.9883	0.9870	1.12

Données simulées des sites de Boulder et de Sioux Fall au 18 juin

Tableau 5.7 : Moyenne, variance, et intervalle de prévision de la production du 18 juin 2009 basés sur 10000 répétitions pour les sites de Boulder et de Sioux Fall.

Modèle	**Moyenne (kWh/m^2)**	**Variance**	**Intervalle de prévision**		**Mesure du 18 juin 2009 (kWh/m^2)**
			75%	**95%**	
R-Vine	6.83	2.55	4.93 – 8.60	3.28 – 9.51	
Normale	6.80	2.62	4.88 – 8.53	3.16 – 9.32	7.15
Indépendance	6.84	2.18	5.05 – 8.46	3.60 – 9.16	

Le constat est qu'il n'apparaît pas de différence évidente entre la « R-Vine » et la normale multivariée sauf pour les valeurs inférieures à 1 kWh/m^2 (voir la figure 5.6 et le tableau 5.8).

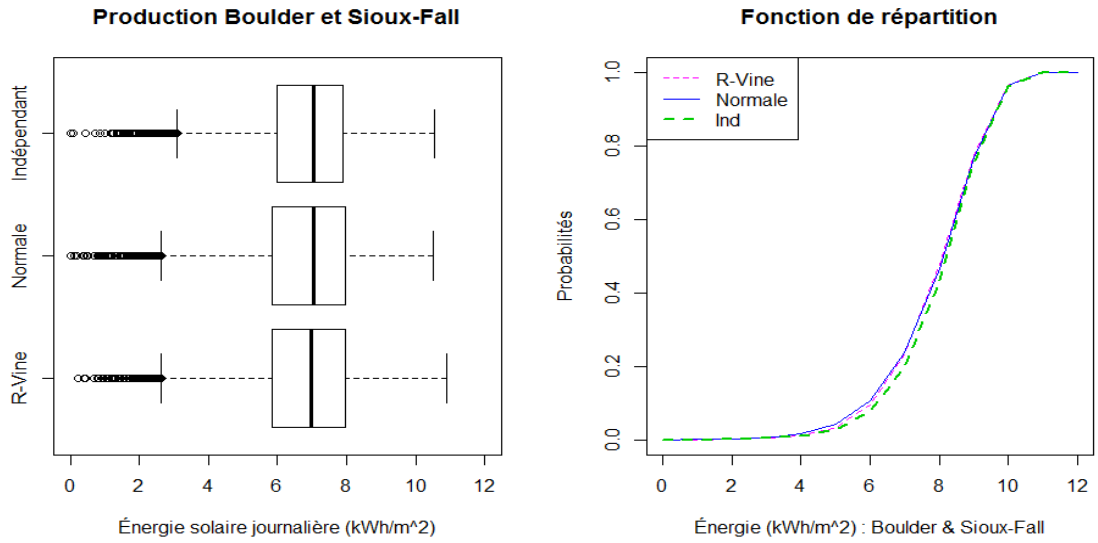

Figure 5.6 : Boxplot et fonction de répartition empirique de 10000 données simulées au 18 juin 2009 pour Boulder et Sioux Fall.

Tableau 5.8 : Probabilités empiriques de l'énergie produite un 18 juin pour les sites de Boulder et de Sioux Fall, basées sur 10000 répétitions selon trois modèles de dépendance spatiale (normale, R-Vine et indépendance).

Énergie solaire / Jour (kWh/m^2)	R-Vine	Normale multivariée	Indépendant	Facteur d'écart
Petites valeurs d'énergie				
< 1	0.0003	0.0010	0.0004	0.30 (3.33)
< 2	0.0014	0.0026	0.0013	0.54 (1.86)
< 3	0.0044	0.0063	0.0039	0.79 (1.43)
< 4	0.0122	0.0169	0.0104	0.73 (1.38)
Grandes valeurs d'énergie				
< 7	0.2325	0.2376	0.1966	1.00
< 8	0.4789	0.4677	0.4357	0.98
< 9	0.7733	0.7694	0.7523	0.98
< 10	0.9661	0.9652	0.9615	0.97

Pour la journée du 18 juin, 10000 données d'énergie sont générées par site à partir des modèles (Vigne régulière « R-Vine », Normale multivariée et l'indépendance). Pour chacune des six réalisations, la valeur maximum est retenue. La distribution du maximum peut s'avérer pertinente dans l'évaluation de certains risques, par exemple si le réseau de transport est limité. Aussi, la différence entre la normale et les « R-Vine » pourrait s'avérer plus grande pour un événement extrême.

Le maximum d'énergie mesurée à la période du 18 juin 2009 en considérant l'ensemble des sites correspond à l'énergie mesurée sur le site de Goodwin Creek avec 8.27 kWh/m^2 (voir le tableau 5.2). Mais il semble avoir moins de variabilité dans les grandes valeurs d'énergie simulées (voir le Tableau 5.9). Aussi, la distinction entre les trois modèles est difficilement perceptible, néanmoins, lorsqu'on regarde de près la figure 5.7, il semble avoir une légère différence entre le modèle indépendant et les deux autres (Normale et Vigne régulière ou « R-Vine »).

Tableau 5.9 : Moyenne, variance, et intervalle de prévision des maximums de productions du 18 juin 2009 basés sur 10000 répétitions pour l'ensemble des sites.

Modèle	Moyenne (kWh/m^2)	Variance	Intervalle de prévision		Mesurée au 18 juin 2009 (kWh/m^2)
			75%	95%	
R-Vine	8.69	0.75	7.70 – 9.66	6.82 – 10.22	
Normale	8.73	0.75	7.72 – 9.69	6.86 – 10.24	8.27
Indépendance	8.79	0.68	7.83 – 9.71	7.02 – 10.20	

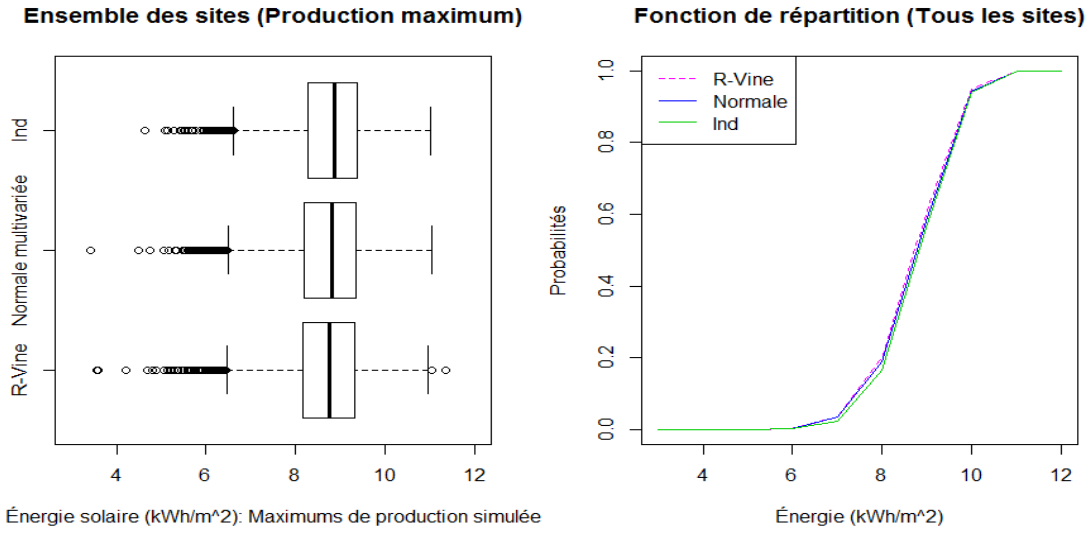

Figure 5.7 : Boxplot et fonction de répartition empirique de 10000 données simulées pour les maximums de la production au 18 juin 2009 de tous les sites.

Tableau 5.10 : Probabilités empiriques des maximums de l'énergie produite un 18 juin pour tous les sites, basées sur 10000 répétitions selon trois modèles de dépendance spatiale (normale, R-Vine et indépendance).

Énergie solaire / Jour (kWh/m^2)	R-Vine	Normale multivariée	Indépendant	Facteur d'écart
Petites valeurs d'énergie				
< 4	0.0002	0.0001	0.0000	2.00
< 5	0.0006	0.0003	0.0001	2.00
< 6	0.0042	0.0036	0.0018	1.17
< 7	0.0366	0.0286	0.0218	1.28
Grandes valeurs d'énergie				
< 8	0.2006	0.1897	0.1646	0.99
< 9	0.6135	0.5903	0.5741	0.94
< 10	0.9498	0.9433	0.9390	0.89
< 11	0.9998	0.9997	0.9999	0.67

Tableau 5.11 : Moyenne, variance, et intervalle de prévision des minimums de la production du 18 juin 2009 basés sur 10000 répétitions pour l'ensemble des sites.

Modèle	Moyenne (kWh/m^2)	Variance	Intervalle de prévision		Mesurée au 18 juin 2009 (kWh/m^2)
			75%	95%	
R-Vine	3.34	3.95	0.31 – 5.64	0 – 6.62	
Normale	3.27	4.02	0.11 – 5.61	0 – 6.69	3.21
Indépendance	3.15	3.80	0.11 – 5.48	0 – 6.50	

Contrairement aux grandes valeurs, le tableau 5.11 montre de grandes variabilités à produire de petites valeurs d'énergie. Ceci est vrai aussi bien pour la « R-

Vine » que pour la normale multivariée. Cependant, la figure 5.8 illustre une certaine différence entre les modèles même si cette différence semble passablement visible. Le 18 juin, la production minimum mesurées est celle du site de Penn State avec 3.21 kWh/m^2 (tableau 5.2).

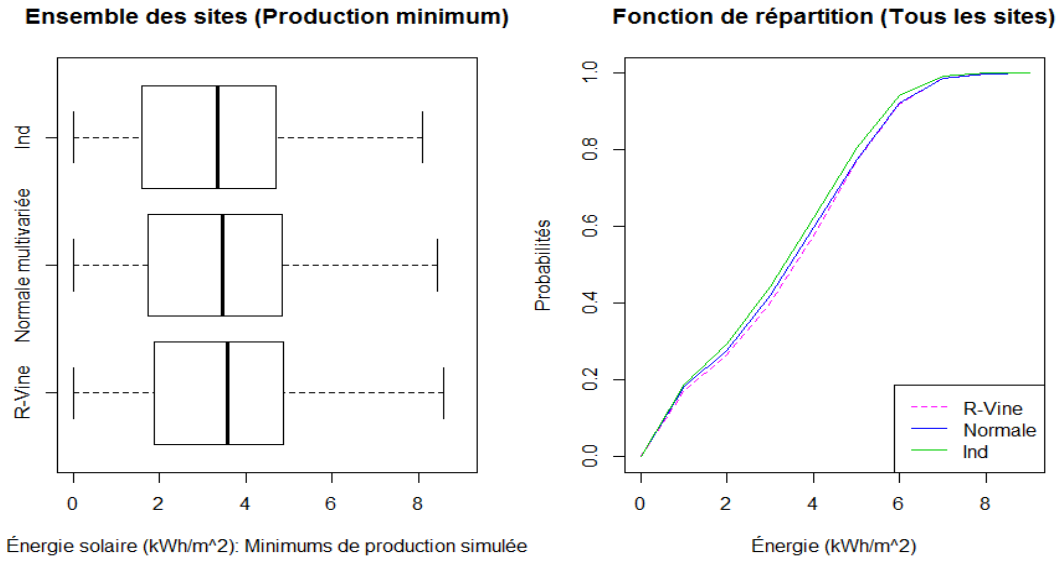

Figure 5.8 : Boxplot et fonction de répartition empirique de 10000 données simulées pour les minimums de la production au 18 juin 2009 de l'ensemble des sites.

5.2.3. Modèles « R-Vine » tronqués

Nous présentons deux résultats graphiques des valeurs d'énergie simulées à partir des modèles tronqués de la « R-Vine », c'est-à-dire le modèle tronqué à l'arbre 2 (R-Vine 1) et celui tronqué à l'arbre 3 (R-Vine 2). Les observations faites en comparant les modèles tronqués de la vigne régulière et la normale multivariée ne sont pas différentes de celles de la section 5.2.2. C'est à dire les différences entre les modèles restent minimes, néanmoins elles semblent exister et les commentaires sont identiques.

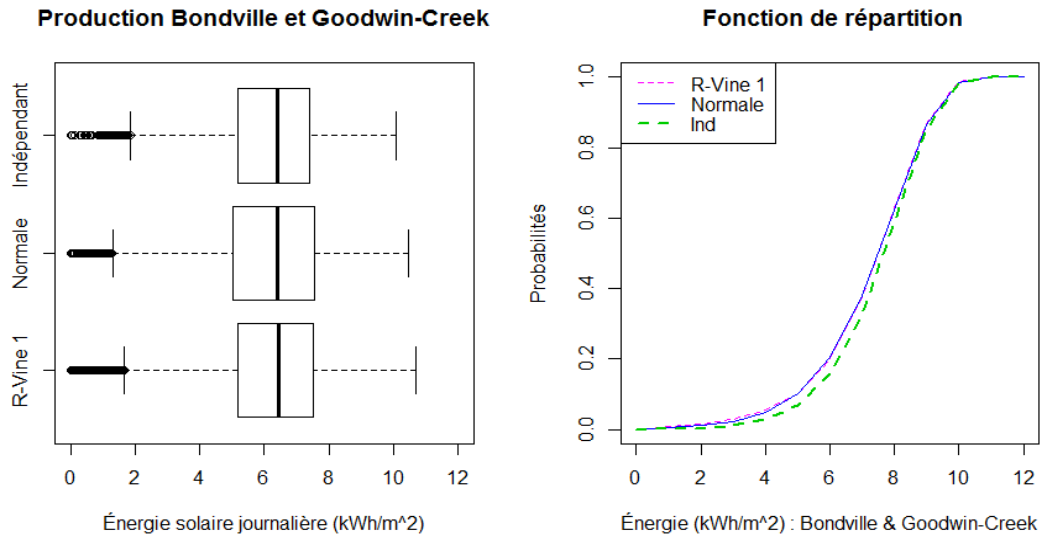

Figure 5.9 : Boxplot et fonction de répartition empirique de 10000 données simulées au 18 juin 2009 pour Bondville et Goodwin Creek (R-Vine 1, Normale et indépendance).

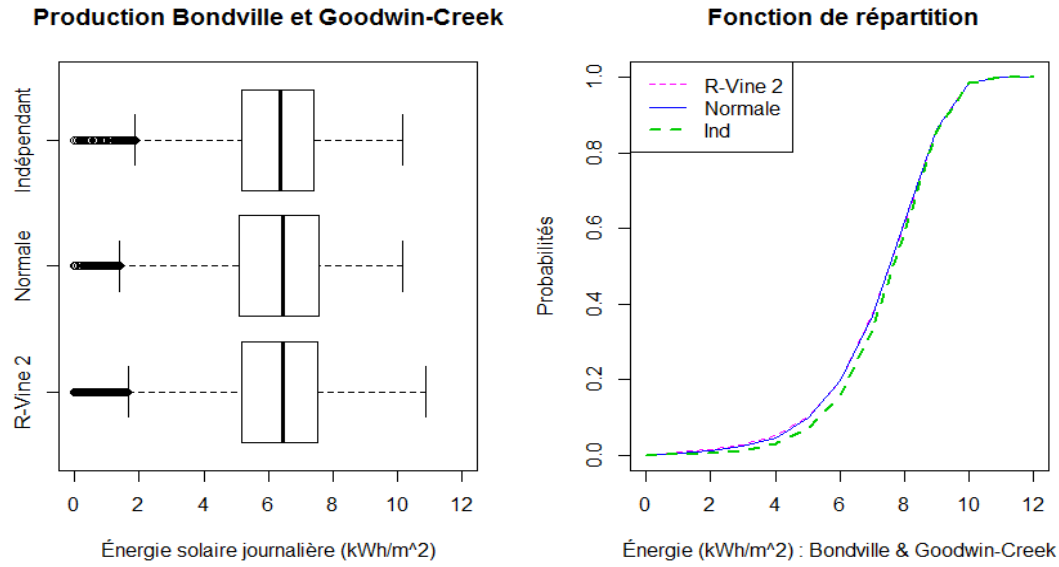

Figure 5.10 : Boxplot et fonction de répartition empirique de 10000 données simulées au 18 juin 2009 de Bondville et de Goodwin Creek (R-Vine 2, Normale et indépendance).

5.2.4. Production mensuelle d'énergie solaire

Les productions mensuelles simulées sont présentées dans cette section. Nous choisissons deux mois pour illustrer les simulations, il s'agit des mois de juillet et de décembre. Deux périodes de l'année où les productions sont vraisemblablement différentes. Il s'agit de présenter la production mensuelle de l'ensemble des sites, en choisissant les minimums et les maximums mensuels des productions simulées à partir de la « R-Vine » et de la normale multivariée.

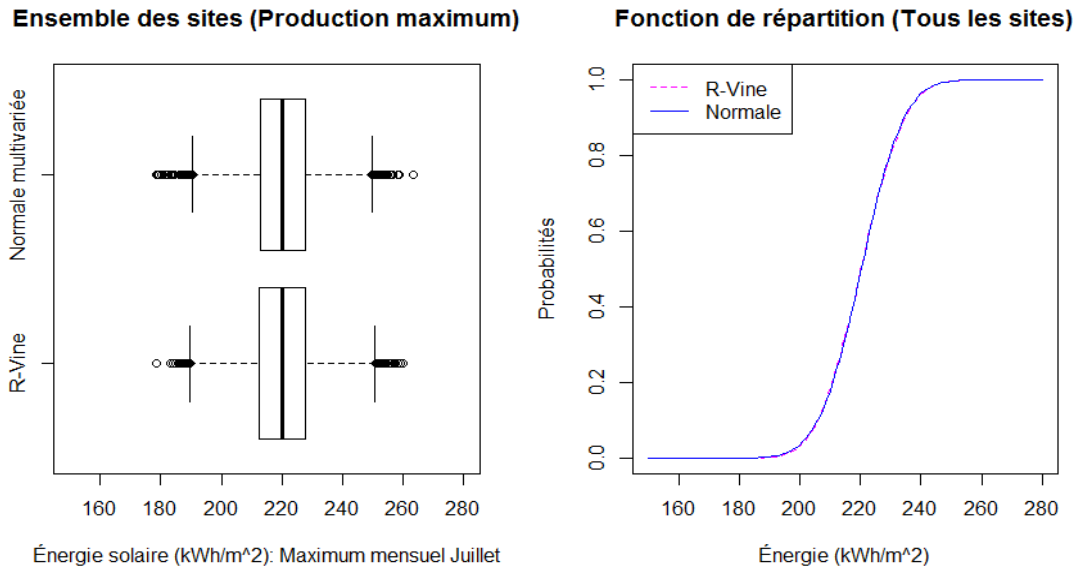

Figure 5.11 : Boxplot et fonction de répartition empirique de 10000 données simulées. Les maximums mensuels des productions de juillet 2009 pour l'ensemble des sites.

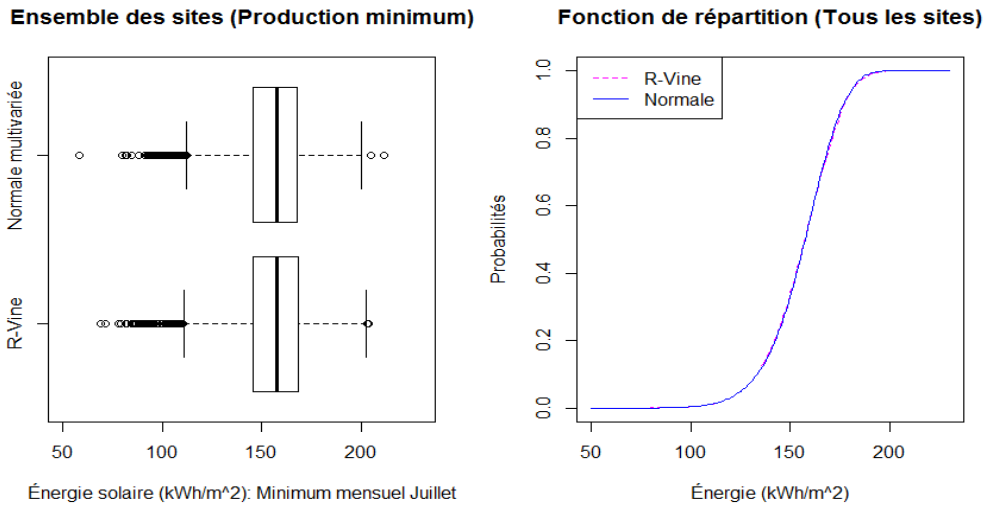

Figure 5.12 : Boxplot et fonction de répartition empirique de 10000 données simulées pour les minimums mensuels des productions de juillet 2009 pour l'ensemble des sites.

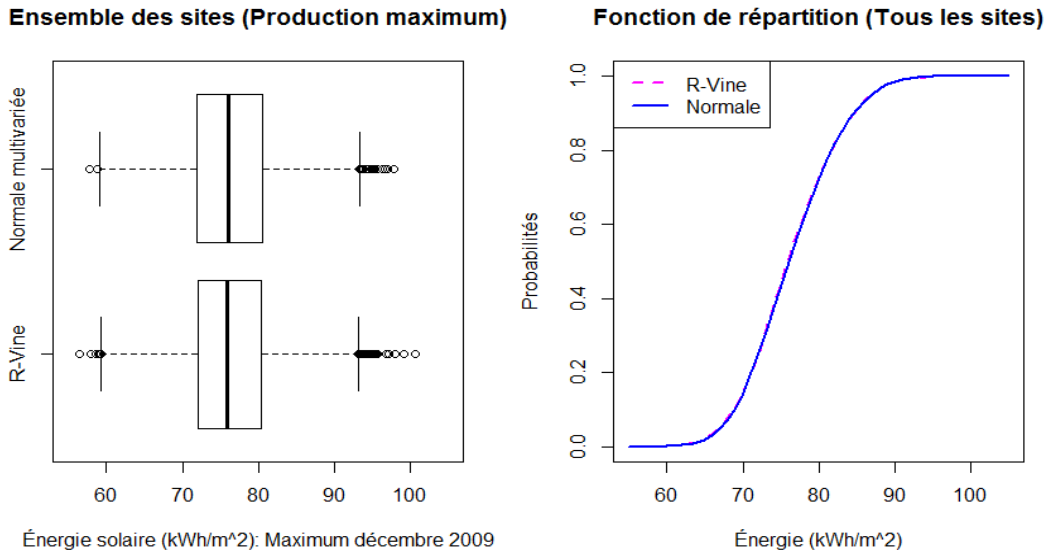

Figure 5.13 : Boxplot et fonction de répartition empirique de 10000 données simulées pour les maximums des productions de décembre 2009 pour l'ensemble des sites.

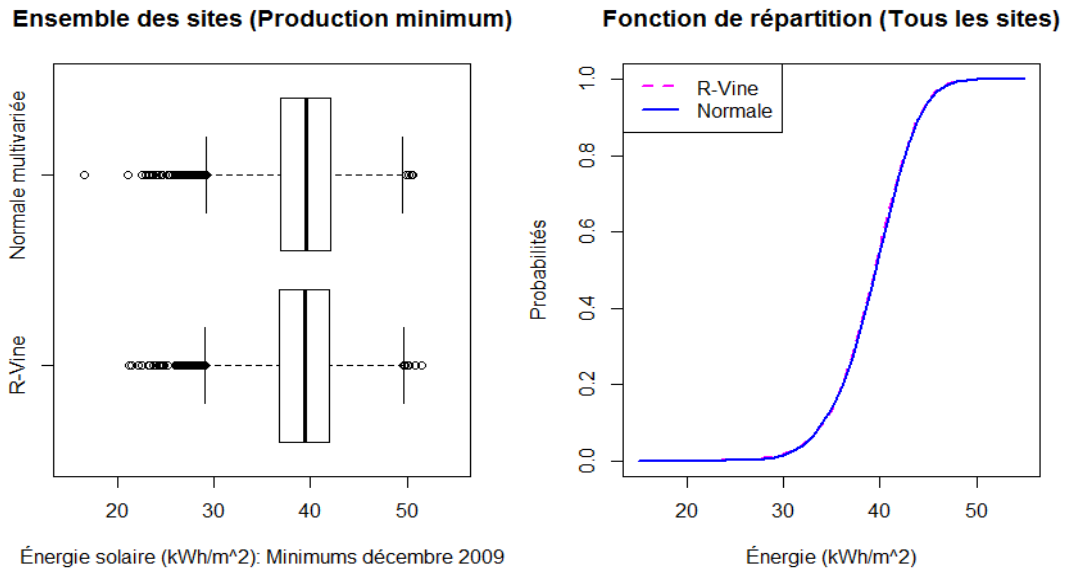

Figure 5.14 : Boxplot et fonction de répartition empirique de 10000 données simulées pour les minimums mensuels des productions de décembre 2009 pour l'ensemble des sites.

Sur les figures 5.12 et 5.14, aucune différence visuelle n'apparaît. Les deux courbes de répartition empiriques des ajustements de la « R-Vine » et de la normale multivariée sont nettement superposées.

5.3. Discussion des résultats

Les simulations de la production annuelle d'énergie, pour la période 2009, illustrant que les modèles de séries chronologiques estimés au chapitre 3 pour chaque site sont raisonnables. La figure 5.1 et la figure 5.2 relèvent l'allure similaire des données mesurées et celles simulées à la même période.

L'ajustement des données d'énergie, pour une journée, à partir des modèles « R-Vine » se différencie subtilement de celui de la normale multivariée. Mais il

convient toutefois de bien préciser ce que nous entendons par différence subtile. Lorsque que nous observons la figure 5.4 et les données du tableau 5.3 ainsi que la figure 5.5 et les données du tableau 5.5, la différence entre la « R-Vine » et la normale multivariée apparait pour de petites valeurs d'énergie (valeurs inférieures à 2 kWh/m^2). Signalons en passant que la différence d'ajustement entre les modèles est observée pour des sites qui ont une corrélation significative. Toutefois, pour des valeurs d'énergie supérieures à 4 kWh/m^2, les deux modèles semblent offrir le même ajustement. Cette remarque est d'autant justifiée qu'il apparait évident de la constater lorsqu'on considère la production mensuelle de l'ensemble des sites où les valeurs d'énergie sont souvent supérieures à 20 kWh/m^2, la figure 5.11, la figure 5.12, la figure 5.13 et la figure 5.14 de la section 5.2.4 mettent clairement en évidence cette absence de différence entre les deux ajustements.

D'un autre côté, si nous considérons deux sites dépendants de par leurs innovations et dont la structure de dépendance est définie par la copule de Student, l'ajustement des valeurs d'énergie est semblable car le modèle « R-Vine » estimé est définit par la loi Student (voir la figure 4.8 et le tableau 4.14). En admettant la flexibilité de la méthode « R-Vine », il ressort de toute évidence, dans cette étude que l'ajustement des données d'énergie à partir du modèle simple, c'est-à-dire la normale multivariée, n'est significativement pas différente de l'ajustement « R-Vine ».

Chapitre 6 : Conclusion

L'évaluation du risque par mesure de la variabilité dans la production d'énergie constitue un enjeu pour planifier et installer des centrales de production d'énergie solaire sur un site. Ce mémoire évalue cette variabilité en ajustant des productions d'énergie à partir d'un modèle simple, la normale multivariée et d'un modèle utilisant les copules (R-Vine). Mais pour y arriver, les variations à chaque site sont représentées par des modèles de séries chronologiques (de type ARMA), estimés à partir des données solaires extraites de Surfrad. La saisonnalité est par ailleurs capturée par un modèle physique de rayonnement solaire. Un modèle de dépendance spatiale entre les innovations des processus ARMA est déterminé par la copule en vigne régulière (R-Vine). La méthode est novatrice en ce sens qu'elle prend en compte un ensemble de variables de dimension plus grande pour établir la structure de dépendance conjointe. Cette méthode présente une approche attrayante dans la modélisation flexible des risques.

Mais si la « R-Vine » est flexible dans sa mise en œuvre pour des variables de plus grande dimension, nous constatons que l'ajustement des productions d'énergie solaire par la R-Vine, dans ce mémoire ne semble pas offrir une réelle différence par rapport à l'ajustement de la normale multivariée. Ces résultats nous amène à nous questionner, d'une part sur l'hypothèse de départ pour élaborer le modèle théorique de rayonnement solaire en supposant un ciel sans atmosphère (transmittance égale à 1), et d'autre part sur le fait que nous n'avions pas tenu compte des distances marginales sur l'espace géographique des sites de l'étude. Sur une échelle géographique plus locale, la différence très subtile entre les copules et la normale multivariée pourrait possiblement s'exprimer plus clairement. Il pourrait être mieux indiqué de considérer des sites situés dans un même espace géographique pour réduire les aléas hypothétiques météorologiques et également éliminer les effets de fuseau horaire car les conditions atmosphérique et météorologique peuvent être très variables d'un État à l'autre. Un

exemple pour illustrer cela est que l'évaluation de la variabilité dans la production pourrait se faire pour des sites situés uniquement soit dans l'État du Colorado, soit dans l'État de l'Illinois. En plus d'une corrélation plus élevée, on peut sans doute s'attendre à des comportements particuliers des extrêmes de la dépendance par exemple.

Il serait également possible d'évaluer autrement le risque de production d'énergie solaire en optant pour les lois stables, dont une somme de variances est dans la même famille de lois. La normale possède cette propriété, mais d'autres lois aux queues plus lourdes aussi : que nous pouvons retrouver dans Lévy et Walter (2002), ainsi que dans les travaux de Fama et Roll (1971). Il ressort des travaux de ces chercheurs que l'hypothèse du « bruit blanc » gaussien pour les résidus des processus autorégressifs à moyenne mobile (ARMA) est souvent trop restrictive et ne rend pas compte de la variabilité des données.

Annexes

A.1 Position des sites pour le modèle physique (Surfrad)

Tableau A1 : Coordonnées des sites du réseau Surfrad

Penn State:		
Tower:	40.72033° N	77.93100° W
Tracker:	40.72023° N	77.93090° W
Platform:	40.72012° N	77.93085° W
Fort Peck:		
Tower:	48.30798° N	105.10177° W
Tracker:	48.30780° N	105.10172° W
Platform:	48.30783° N	105.10170° W
Bondville:		
Tower:	40.05155° N	88.37325° W
Tracker:	40.05195° N	88.37310° W
Platform:	40.05192° N	88.37309° W
Table Mountain:		
Tower:	40.12557° N	105.23775° W
Tracker:	40.12493° N	105.23677° W
Platform:	40.12498° N	105.23680° W
Goodwin Creek:		
Tracker:	34.2547° N	89.8729° W
Sioux Falls:		
Tower:	43.73431° N	96.62334° W
Tracker:	43.73399° N	96.62328° W
Platform:	43.73403° N	96.62328° W

A.2 Distances estimées entre les sites (en km)

Tableau A2 : Distances estimées entre les sites du réseau Surfrad.

Sites	Penn-State	Bondville	Boulder	Fort-Peck	Goodwin-Creek	Sioux-Fall
Penn-State	-	1275	2806	3190	1790	2660
Bondville	1275	-	1645	1985	1210	1465
Boulder	2806	1645	-	885	1935	675
Fort-Peck	3190	1985	885	-	2775	515
Goodwin-Creek	1790	1210	1935	2775	-	2210
Sioux-Fall	2660	1465	675	515	2210	-

B.1 Calcul des paramètres a_0 et a_1 du modèle physique

Le calcul est réalisé pour le 160ième jour d'une année (365 jours), delta ($\delta = 2.29$). Les paramètres a_0 et a_1 sont variables selon l'angle delta (la déclinaison solaire).

Tableau B1 : Paramètres a_0, a_1 et delta (δ) du modèle physique calculés pour un jour julien (J = 160).

Sites	**Bondville**	**Goodwin**	**Boulder**	**Fort Peck**	**Sioux Fall**	**Penn-State**
L	40.05	34.25	40.13	48.31	43.73	40.72
a_0	0.25	0.22	0.25	0.29	0.27	0.25
a_1	0.70	0.76	0.71	0.61	0.67	0.69
delta	2.29					
J	160					

B.2 Carte de la répartition des mesures d'énergie solaire

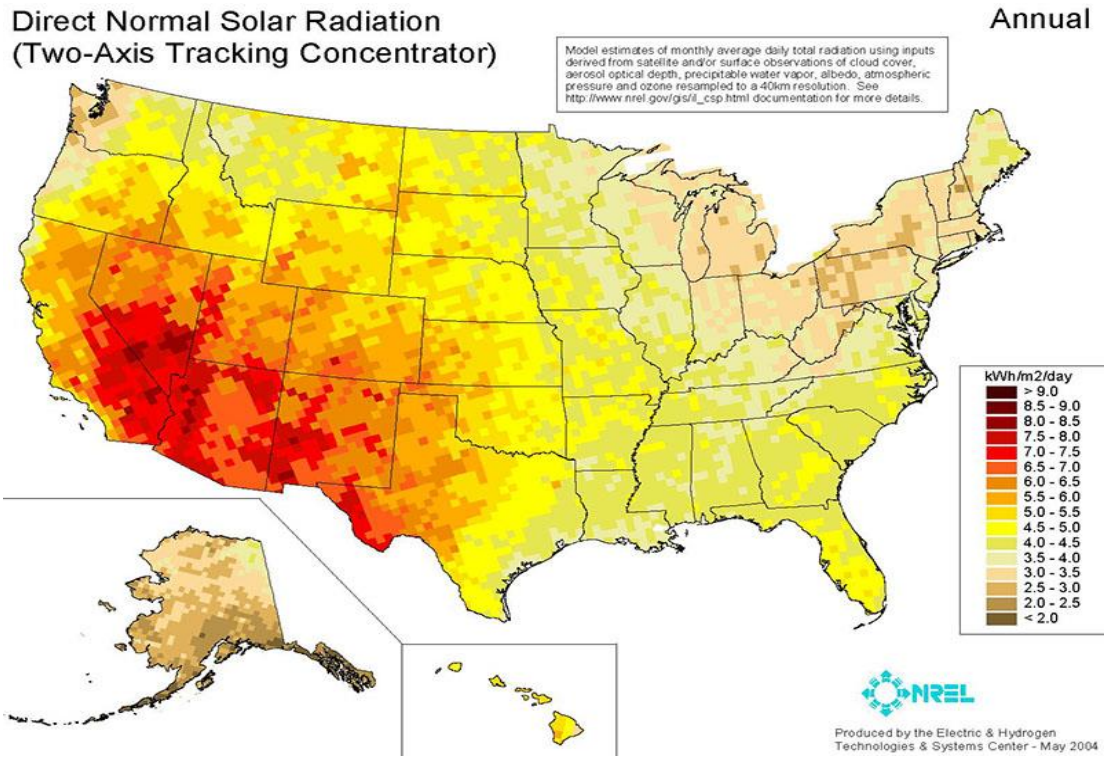

Figure B.2 : Répartition d'énergie solaire mesurée selon les zones géographiques aux États-Unis (source : www.pveducation.org).

Tableau B2 : Facteurs multiplicateurs pour la construction des modèles théoriques moyens par site.

Site	moyenne (« *log ratios* »)	Facteurs multiplicateurs : $(1\text{-}e^{(-\text{moyenne})})$
Penn State	0.687	0.496
Bondville	0.758	0.531
Boulder	0.782	0.543
Fort Peck	0.798	0.550
Goodwin Creek	0.746	0.526
Sioux Fall	0.797	0.549

B.3 Production annuelle d'énergie simulées

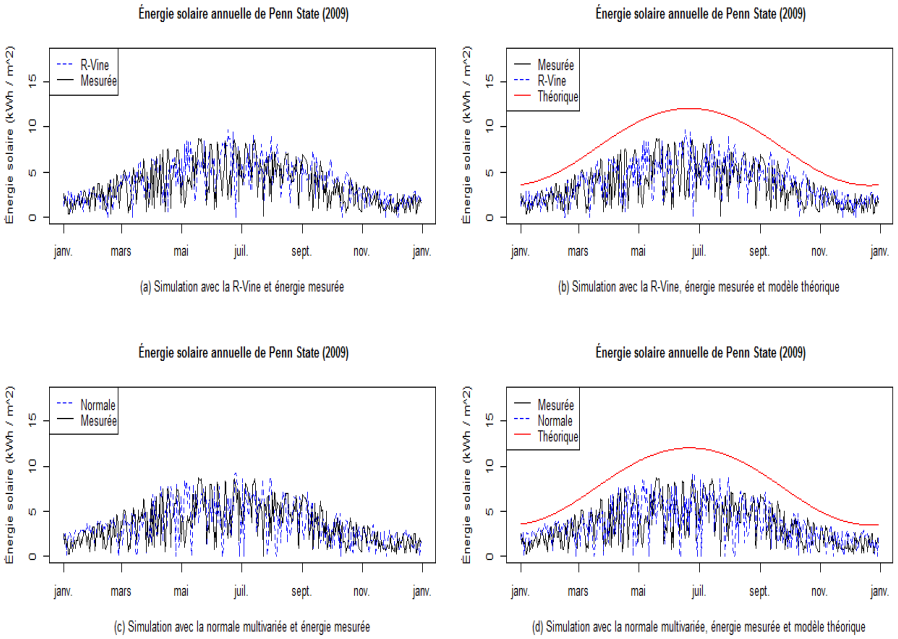

Figure B.3.1 : Production annuelle simulée du site de Penn State - 2009.

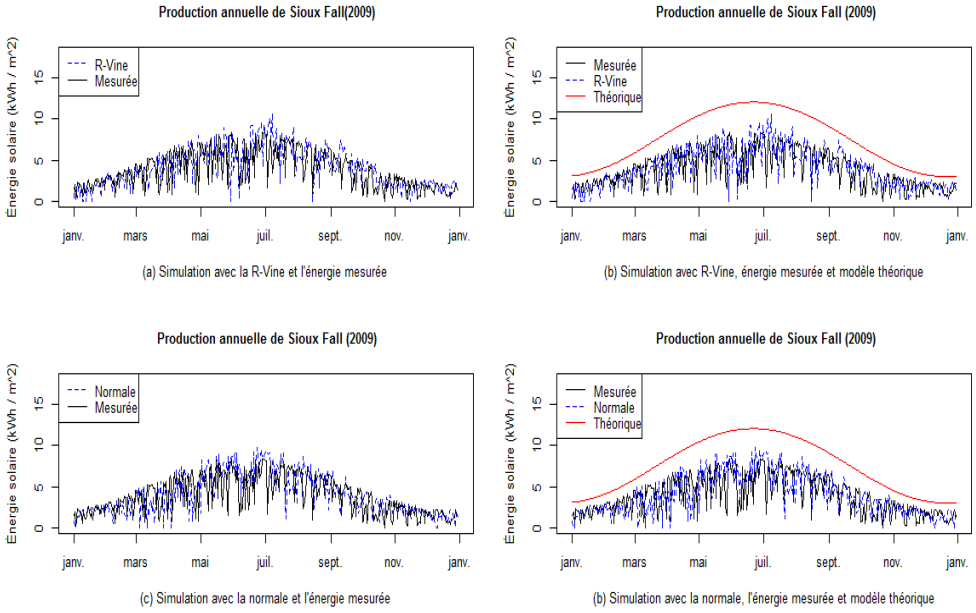

Figure B.3.2 : Production annuelle du site de Sioux Fall simulée - 2009.

C Production d'énergie solaire au 18 juin 2009 des sites de Bondville et Sioux

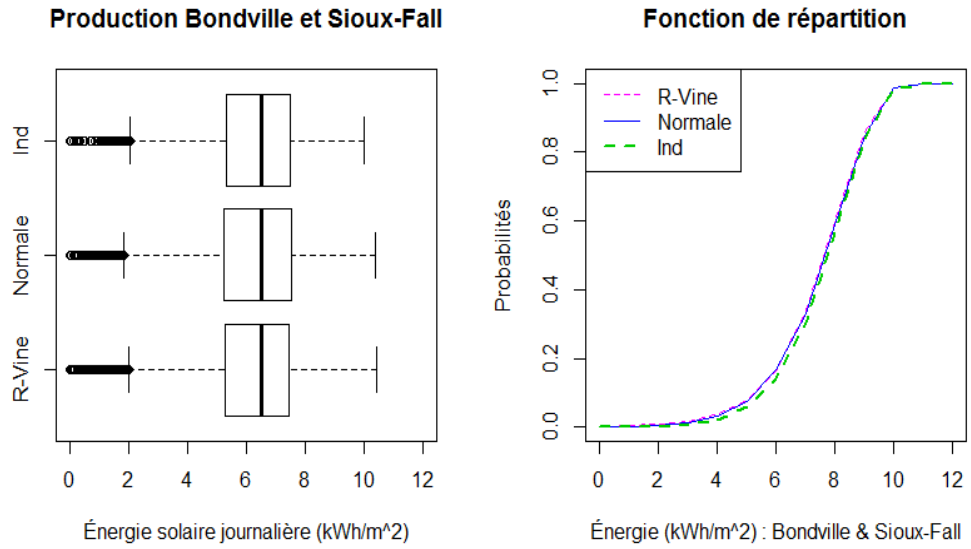

Figure C : Production d'énergie solaire au 18 juin 2009 des sites de Bondville et de Sioux Fall.

Bibliographie

Aas, K., Czado, C., Frigessi, A. & Bakken, H. (2009). *Pair-copula constructions of multiple dependence Insurance*, Mathematics and Economics, 44 (2), 182 – 198.

Allen, D. E., Ashraf, M. A., McAleer, M., Powell, R. J. & Singh, A. K. (2013). *Financial dependence analysis: applications of Vine Copulas*, Statistica Neerlandica, Vol. 67, nr 4, pp. 403 – 435.

Azami, Z., Ahmad M., R., Tee, P., G., & Kamaruzzaman, S. *(*2009*). Time Series Analysis of solar radiation data in the Tropics*. European Journal *of* Scientific Research, Vol.*25* No.4, pp.672-678.

Bastien, D. & Athienitis, A. (2011). *Le potentiel des énergies solaires au Québec,* Greenpeace Canada, ISBN 978-0-9877581-0-1.

Beckers, B. & Beckers, P. (2010). *Comment calculer la déclinaison du soleil*, Rapport Helio 7, http://www.heliodon.net/downloads/Beckers_2010_Helio_007_fr.pdf, consulté le 23 février 2013.

Beckers, B. & Beckers, P. (2011). *Calcul du rayonnement solaire atténué*, Rapport Helio 8, http://www.heliodon.net/downloads/Beckers_2010_Helio_008_fr.pdf, consulté le 23 février 2013.

Bedford T. & Cooke R. (2002). *Vines - a new graphical model for dependent random variables*, Annals of Statistics 30, 1031-1068.

Boland, J. (2008). *Time series and statistical modelling of solar radiation*, Recent Advances in Solar Radiation Modelling, Viorel Badescu (Ed.), Springer-Verlag, pp. 283-312.

Brechmann, E. & Schepsmeier, U. (2013). *Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine*, Journal of Statistical Software, 52 (3), 1-27.

Brockwell, P. & Davis, R. (1991). *Time Series: Theory and Methods*, Springer Series in Statistics, Springer, second edition.

Campbell, G., & Norman, J. (1998). *An introduction to environmental biophysics,* New York, Springer, second edition, ISBN 0-387-94937-2.

Canmet Energy (2012). *Profil du secteur de l'énergie solaire photovoltaïque au Canada*, Rapport 063 (RP – TEC).

Dickey, D. & Fuller W. (1979). *Distribution of the Estimators for Autoregressive Time Series with a Unit Root*, Journal of the American Statistical Association 74 (427-431).

Dissmann, J. & Brechmann, C**.** (2013). *Selecting and estimating regular vine copulas and application to financial returns*, Computational Statistics and Data Analysis, 59 (1), 52-69.

Drake, B. & Hubacek, K. (2007). *What to expect from a greater geographic dispersion of wind farms? A risk portfolio approach*, Energy Policy, Volume 35 (8), 3999 – 4008.

Embrechts, P., Resnick, S. I. & Samorodnitsky, G. (1999). *Extreme Value Theory as a Risk Management Tool*, North American Actuarial Journal, Volume 3.

Enders, W. (2010). *Applied Econometric Time Series*, John Wiley and Sons, 3 rd. edition.

Fama, E. & Roll, R. (1971). *Parameters Estimates for Symmetric Stable Distributions*, Journal of the American Statistical Association, 66, 331-339.

Gardner, G., Harvey, A. C. & Phillips, G. D. A. (1980). *Algorithm AS154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering*, Applied Statistics 29, 311–322.

Genest, C. & MacKay, R. J. (1986a). *The joy of copulas - bivariate distributions with uniform marginals*, The American Statistician, 40, 280-283.

Genest, C. & MacKay, R. **(**1986b). *Copules archimédienne et familles de lois bidimensionnelles dont les marges sont données*, The Canadian Journal of Statistics, 14 (2) 145-159.

Genest, C. & Rémillard, B. (2004). *Tests of independence and randomness based on the empirical copula process*, Test 13 (2), 335-369.

Genest, C. & Favre, A. C. (2007). *Everything you always wanted to know about copula modeling but were afraid to ask*, Journal of Hydrologic Engineering, 12, 347-368

Genest, C., Rémillard, B. & Beaudoin, D. (2009). *Omnibus goodness-of-fit tests for copulas: A review and a power study*, Insurance Mathematic Economic, 44, 199-213.

Hamilton, J. (1954). *Time Series Analysis*, Princeton University Press.

Hyndman, R. J. & Khandakar, Y. (2008). *Automatic time series forecasting: The forecast package for R*, Journal of Statistical Software, 26 (3).

Iqbal, M. (1983). *An introduction to solar radiation*, First Edition Academic Press New York.

Joe, H. (1996). *Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters*, Institute of Mathematical Statistics, pp. 120-141.

Kramer, N. & Schepsmeier, U. (2011). *Introduction to vine copulas*, Technische Universitat Munchen, www.statistics.ma.tum.de/fileadmin/w00bdb/www/veranstaltungen/Vines.pdf, consulté le 14 novembre 2013.

Kreith, F. & Kreider, J. (1978). *Solar Energy Handbook*, McGraw-Hill.

Kurowicka, D. & Cooke, R. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*, Chichester, John Wiley.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. & Shin, Y. (1992). *Testing the null hypothesis of stationary against the alternative of a unit root: How sure are we that economic time series have a unit root?* Journal of Econometrics, 54 (1992) 159-178.

Larsen, R., James, W., Mjelde, J. W., Klinefelter, D. & Wolfley, J. (2013). *The use of copulas in explaining crop yield dependence structures for use in geographic diversification*, Agricultural Finance Review Vol. 73 No. 3, pp. 469-492.

Lauret, P., Boland, J. & Ridley, B. (2010). *Derivation of a solar diffuse fraction model in a Bayesian framework*, Case Studies in Business, Industry and Government Statistics, 3(1), pp. 108-122.

Levy-Véhel, J. & Walter, C. (2002). *Les marchés fractals*, PUF, Paris.

Liu, B. & Jordan, R. (1960). *The interrelationship and characteristics distribution of direct, diffuse and total solar radiation*, Solar Energy, 4 (3), 1-19.

McLeod, A. I. & Li, W. K. (1983). *Diagnostic checking ARMA time series models using squared-residual autocorrelations*, Journal of Time Series Analysis 4, 269–273.

Nowicka-Zagrajek, J. & Weron, R. (2002*). Modeling electricity loads in California: ARMA models with hyperbolic noise*, Signal Processing 82, 1903 – 1915.

Piedallu, C. & Gégout, J. C. (2007). *Multi-scale computation of solar radiation for predictive vegetation modelling,* Annals of Forest Science, 64, 899–909.

Piedallu, C. & Gégout, J. C. (2008). *Efficient assessment of topographic solar radiation to improve plant distribution models*, Agricultural and forest meteorology, 148, 1696 – 1706.

Rémillard, B. (2013). *Statistical Methods for Financial Engineering*, CRC Press, Taylor et Francis Group.

Ripley, B. D. (1987). *Stochastic Simulation.* Wiley, page 98.

Solanki, C.S. & Pimpalkar, P. (2005). *Solar irradiation potential for PV concentrator systems in India,* Scottsdale, Arizona.

Solanki, C. S. & Sangani, S. (2008). *Estimation of monthly averaged direct normal solar radiation using elevation angle for any location,* Solar Energy Materials and Solar Cells 92 38-44.

Younes, S., Claywell, R. & Muneer, T. (2005). *Quality control of solar radiation data: Present status and proposed new approaches*, Energy, 30, 1533 – 1549.