



MODÉLISATION ET OPTIMISATION STOCHASTIQUE DES
SOURCES ET BESOINS D'UN RÉSEAU MULTI-ÉNERGIES DANS
UNE VILLE BAS CARBONE

TAÏEB Zinédine

Table des matières

1 Contexte du stage	1
1.1 Introduction générale	1
1.2 Contexte académique et industriel du stage	2
1.2.1 L'entreprise et son champ d'action	2
1.2.2 Le laboratoire SATIE et son expertise dans le domaine des smarts grids	4
1.3 Enjeux scientifiques du stage	4
1.3.1 Considérer le problème de gestion d'énergies comme un problème multi-modal	4
1.3.2 L'énergie solaire offre de très nombreuses applications et s'intègre parfaitement à un contexte d'utilisation dans un carrefour énergétique	7
1.3.3 Un stage s'inscrivant dans un travail de recherche	9
2 Etat de l'art et Méthodologie	11
2.1 État de l'art sur les méthodes de prédictions et les enjeux par rapport au contexte de recherche	11
2.1.1 L'importance des données	11
2.1.2 Un accès qualitatif aux données industrielles encore limité	11
2.1.3 Une revue très générale des méthodes de prédictions sur l'irradiance	12
2.2 Cas d'étude	14
2.2.1 Formulation du problème et du cadre du stage	14
2.2.2 Un choix de modèles de prédictions constraint entre temps de calculs limités et performance prédictive	15
2.2.3 Objectifs du stage	16
2.3 Présentation des modèles de prédictions utilisés	16
2.3.1 Modèle régressif SARIMA	16
2.3.2 Modèle Markov Caché multivariée à émission gaussienne :	17
2.4 Méthodologie	20
2.4.1 Pré-traitement des données	21
2.4.2 Discussion sur les critères de performance	21
2.4.3 Méthodes de prédictions adaptées pour les deux modèles, et calcul de l'energy score	22
2.4.4 Optimisation des paramètres	23
2.4.5 Une étude sur la taille optimale d'entraînement	24
3 Résultats des modèles de prédition sur des profils d'irradiance globale	26
3.1 Résultats des modèles SARIMA	26
3.1.1 Validation sur des données artificielles	26
3.1.2 Evaluation de la performance des modèles SARIMA sur les données d'irradiances	29
3.2 Résultats des modèles HMM	31

3.2.1	Validation sur les données artificielles	31
3.2.2	Evaluation de la performance des modèles HMM à émissions gaussiennes multivariées sur les données d'irradiance	35
4	Conclusions	42

1

Contexte du stage

1.1 Introduction générale

La communauté scientifique ne cesse d'alerter sur le désastre environnemental qui s'accélère et s'aggrave au fil du temps. Une prise de conscience mondiale se fait ressentir ces dernières années poussées par plusieurs gouvernements, mais les contributions sont encore loin d'être suffisantes. En France, l'objectif est d'arriver à réduire les émissions de carbone de 40% d'ici 10 ans et finalement d'atteindre la neutralité carbone d'ici 2050. Le gouvernement a lancé une programmation pluriannuelle de l'énergie avec des objectifs sur l'horizon 2019-2023 et 2024-2028. Faire baisser la consommation d'énergie, réduire l'usage des énergies fossiles et nucléaires et diversifier le mix énergétique par le développement des énergies renouvelables, sont les maîtres-mots de la politique gouvernementale pour assurer la stratégie bas carbone d'ici les prochaines décennies. A ce titre, accélérer le développement des énergies renouvelables et leur déploiement en fonction des ressources locales sur le territoire national est un moyen fort de diminuer la part d'utilisation des énergies fossiles et nucléaire. L'énergie d'origine fossile ou nucléaire à ce jour représente encore aujourd'hui 70 % de la production d'énergie primaire en France.

Il est certain que la transition énergétique va modifier drastiquement l'actuelle gestion des ressources énergétiques. Nos modes de consommation et de production devront s'adapter aux nouvelles contraintes sur différents plans afin de préserver la continuité du développement de la société. Anticiper les possibles problématiques futures liées à la transition énergétique est un moyen efficace d'assurer l'accomplissement de grands objectifs sur le long terme tels que la neutralité carbone d'ici 2050 et l'optimalité énergétique en équilibre avec l'environnement pour le futur qui poserait les bases d'une civilisation écologique. C'est à travers ces questionnements et thématiques de gestion durable des énergies, que j'ai voulu orienter mes stages de 4e et 5e année. J'avais à cœur de pouvoir mettre en application les compétences acquises durant tout mon cursus en mathématiques appliquées sur des thématiques qui ont un sens concret pour moi. Ce stage porte ainsi sur la modélisation stochastique des sources et besoins d'un réseau multi-énergie dans une ville bas carbone, en intégrant des stockages flexibles et des productions d'énergies renouvelables. Au cours de l'année précédente, j'ai eu l'occasion d'étudier la prédiction de production d'énergie fatale industrielle. Mon stage actuel s'effectue en soutien de la thèse de mon tuteur, M. Ibrahim Al Asmi, actuellement en troisième année de doctorat. Ses travaux de thèse gravitent essentiellement autour des problématiques sur la mise en place d'un mix énergétique à travers un réseau multi-énergies smart-grid, de sa faisabilité à sa gestion. La thèse en question est sous une convention CIFRE, entre l'entreprise Eco-Tech Ceram, start-up spécialisée dans la revalorisation d'énergie thermique industrielle, et le laboratoire SATIE de l'ENS Rennes, se consacrant aux systèmes et applications des technologies

de l'information et de l'énergie. Les missions de cette année furent de donner des premiers éléments de réponses quant aux bonnes familles de modèles de prédiction d'irradiance solaire, notion importante dans le contexte de contrôle de réseau multi-énergies intégrant les énergies renouvelables dont l'énergie solaire.

1.2 Contexte académique et industriel du stage

Comme introduit plus tôt, ce stage est effectué en connexion avec le laboratoire SATIE et l'entreprise Eco-Tech CERAM. Il nécessite un travail de recherche pour répondre aux problématiques industrielles.

1.2.1 L'entreprise et son champ d'action

Eco-Tech Ceram (ETC) est une société d'ingénierie écologique industrielle qui propose des solutions de récupération d'énergie pour améliorer l'efficacité énergétique des entreprises industrielles et la rentabilité des producteurs d'énergies renouvelables. ETC développe et commercialise son produit phare "Eco-Stock", une solution clé en main innovante de stockage d'énergie thermique.

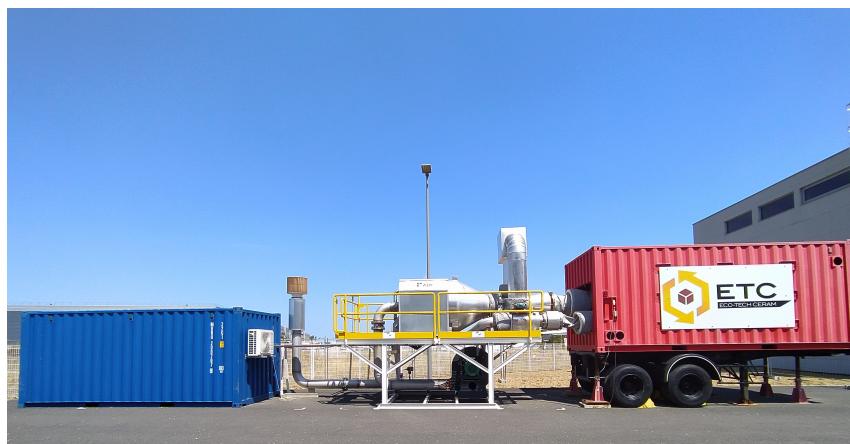


FIGURE 1.1 – Un dispositif Eco-stock (container rouge) commercialisé par l'entreprise, connecté à un convertisseur électrique (container bleu) comme source de chaleur d'entrée

Ce stockage de chaleur fournit une énergie décarbonée à un prix inférieur (jusqu'à 80%) à l'énergie fossile. Cette énergie disponible peut facilement être réutilisée directement sous forme de chaleur, ou convertie en électricité, en air froid ou en vapeur. Comme les sources de chaleur et les besoins énergétiques sont divers et uniques, ETC tient à faire progresser notre connaissance du carrefour de l'énergie. De ce fait, les outils développés peuvent être utilisés pour optimiser la gestion et le dimensionnement des intersections (stockage et conversion), réduisant respectivement les OPEX (coûts d'exploitation) et les CAPEX (dépenses d'investissement).

La solution Eco-stock, une batterie de stockage thermique pour valoriser la chaleur fatale

Ces batteries thermiques Eco-Stock remplies de billes de céramique, peuvent être utilisées dans un contexte de réseau multi-énergies en revalorisant la chaleur émise par les usines thermiques. Ainsi, on peut convertir cette dernière en énergie électrique via des résistances électriques ("Heat To Power"). En effet, au lieu de délester cette énergie dite fatale dans

l'atmosphère, ces usines peuvent la stocker, sous forme thermique, captée dans les batteries afin de pouvoir la réutiliser ultérieurement, en fonction des besoins itinérants. Les possibilités d'applications sont alors multiples comme montré dans la Figure 1.2.

Les possibilités d'utilisation d'Eco-stock dans un contexte smart-grid

L'accumulateur Eco-Stock est un mode de stockage flexible aux différents contextes de production et de consommation. Cette flexibilité ouvre le champ des possibilités dans un contexte de gestion de ressources, dans l'optique de s'adapter aux contraintes du réseau, et aux demandes en temps réel. Avec un stockage thermique de l'ordre de quelques jours, il est possible de valoriser l'énergie fatale à l'aide de convertisseurs énergétiques :

- 1) dans un contexte de conversion “Heat to Heat” : stockage thermique et valorisation de chaleur fatale pour les besoins en chaleur dans le secteur industriel et domestique (exemple : chauffage, séchage de matières, préchauffage d'air de combustion...).
- 2) dans un contexte de conversion “Heat to Power” : stockage thermique et valorisation de chaleur fatale en électricité dans le secteur industriel et domestique.
- 3) dans un contexte de conversion “Power to Heat” : stockage thermique pour l'effacement et l'engagement de consommation électrique dans le secteur industriel et domestique pour un substitut aux équipements de chauffage, de cuisson fossile...
- 4) dans un contexte de conversion “Power to Power” : stockage massif d'électricité sous forme thermique pour répondre aux besoins “énergivores” insatisfaits par les systèmes conventionnels de stockage d'électricité, (exemple : résistance électrique couplée avec stockage thermique et une turbine à vapeur “Heat to Power”)

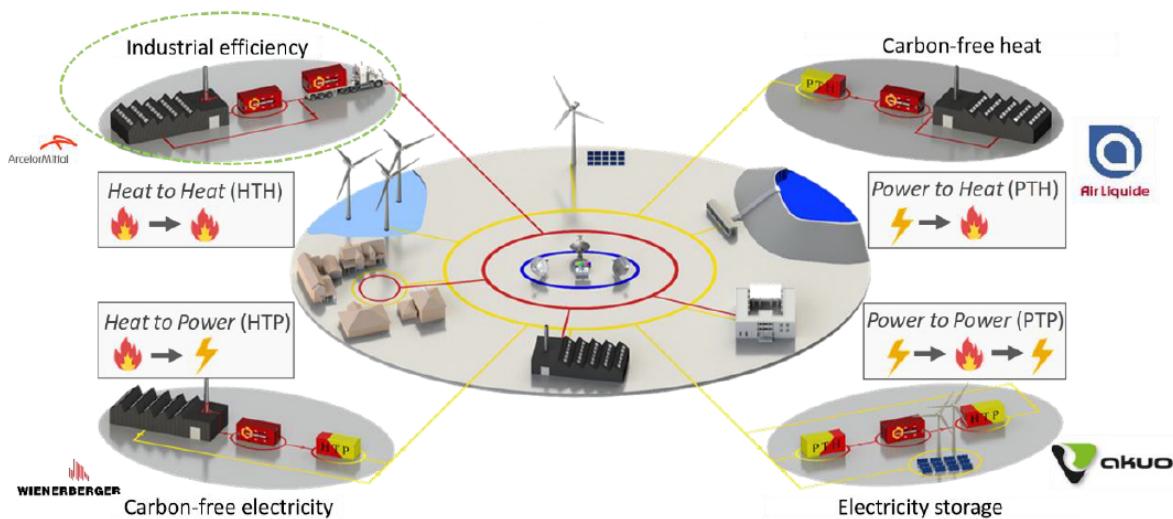


FIGURE 1.2 – Illustration des champs d'applications de la solution Eco-Stock

Ces applications rendent de ce fait possible le couplage entre le stockage thermique et électrique. Des contraintes sur la température minimale de stockage sont par ailleurs fixées. La technologie actuelle du stockage thermique ne permet pas une conservation prolongée de l'énergie issue de la chaleur fatale, et doit être réutilisée dans une période de quelques jours sous risque de diminuer fortement la qualité de l'énergie fournie.

1.2.2 Le laboratoire SATIE et son expertise dans le domaine des smarts grids

Le SATIE (Systèmes et Applications des Technologies de l'Information et de l'Energie) est un laboratoire de recherche en sciences appliquées. C'est une unité mixte de recherche du CNRS comptant 200 personnes. Plus spécifiquement, l'équipe avec laquelle j'ai collaboré est spécialisée dans les systèmes d'énergies pour les transports et l'environnement sous la direction de Hamid BEN AHMED.

Concernant les réseaux électriques dit intelligents "Smart Grid", le SATIE a développé depuis un certain nombre d'années, un savoir-faire reconnu dans le domaine des méthodes de la co-optimisation - dimensionnement et gestion - d'un réseau électrique avec stockage. Les principaux apports méthodologiques de l'équipe sont la modélisation des composantes aléatoires des productions et des consommations d'électricité, l'optimisation stochastique des stratégies de gestion prenant en compte le vieillissement des composants ainsi que le couplage dimensionnement-gestion sur cycle de vie de ces réseaux. Le tout en intégrant à la fois les aspects stochastiques mais aussi les approches de cycle de vie (vieillissement du stockage électrochimique).

1.3 Enjeux scientifiques du stage

1.3.1 Considérer le problème de gestion d'énergies comme un problème multi-modal

Les besoins énergétiques, tant des consommateurs particuliers que des industriels, sont essentiellement multimodaux. Il faut pouvoir offrir un panel d'utilisation très large, car l'amplitude et le type d'énergie peuvent différer considérablement en fonction de l'usage visé : chauffage intérieur, mobilité, équipements électriques, procédés industriels, etc. De plus, des convertisseurs - en production et en consommation - permettent la conversion d'une forme à une autre : conversion de l'électricité en chaleur, de la chaleur en électricité... Chercher à construire le meilleur système d'approvisionnement en énergie nécessite par conséquent de considérer ce problème dans son ensemble multimodal. Découpler ce problème en plusieurs réseaux mono vecteurs, rendrait le problème technique certes plus simple, mais se paiera en contre-partie par une sous optimalité de la solution apportée, due à une formulation du problème non adaptée à la réalité. Conserver la nature couplée du problème se heurte néanmoins à une grande complexité technologique, qui va de la conception de chacun des composants nécessaires à un tel réseau hybride jusqu'à leur modélisation et à la gestion globale de l'ensemble.

Les difficultés d'exploitation des énergies renouvelables : le cas de l'énergie thermique

L'effort pour surmonter les difficultés techniques d'un réseau hybride est justifié car de nombreuses ressources pourraient alors être exploitées beaucoup plus largement qu'aujourd'hui. En effet, de nombreuses sources sont disponibles mais ne sont actuellement utilisées que de manière secondaire. C'est le cas de l'énergie thermique : en ne tenant compte que de la chaleur fatale, celle-ci est dégagée par les industries de manière incontrôlée en marge de leur process. Celle-ci représente 3500 TWh par an et 2450 Mt de CO₂ par an en terme de rejets de gaz carbonique (11).

Plusieurs sources renouvelables se manifestent également par d'importants gisements de chaleur. C'est le cas également de la géothermie, qui dégage de la chaleur en permanence et est facilement accessible, mais difficile à extraire en raison des basses températures.

L'utilisation thermique du rayonnement solaire est également une source de chaleur basse température (chauffe-eau solaire) facile à utiliser qui peut aussi se présenter sous forme de très hautes températures grâce aux champs héliostatiques qui concentrent la chaleur ; on peut citer le célèbre exemple de four solaire d'Odeillo dans le sud de la France géré par le CNRS-PROMES. L'installation peut produire une puissance thermique d'un mégawatt, et atteindre une température les 3 300°C . Cette chaleur peut ensuite être convertie en électricité grâce à des turbines. Il s'agit alors d'un couplage intrinsèque entre la chaleur et l'électricité dans le processus de conversion d'énergie. De plus, ce lien entre chaleur et électricité est également présent dans les centrales électriques qui utilisent la co-génération, une combinaison de formes d'énergie thermique et électrique.

Diverses sources de chaleur sont utilisées depuis longtemps, par exemple à travers les réseaux thermiques. Ceux-ci profitent de la proximité entre les sources et les consommateurs dans des situations - relativement rares à l'heure actuelle - où la production de chaleur peut à tout moment dépasser la demande, faute d'élément chauffant de réserve. L'utilisation de toutes ces sources de chaleur est vraiment faible car leur exploitation est difficile.

Par ailleurs, l'utilisation de cette énergie se voit entravée par divers limites physiques. D'une part les basses températures sont difficilement exploitables ; et d'autre part, la chaleur est une forme d'énergie locale qui ne peut être exploitée efficacement que dans son voisinage.

Des moyens de flexibilités pour atténuer l'intermittence et la faible densité des énergies renouvelables

L'utilisation de sources de chaleur nécessite des composants spécifiques pour la flexibilité. Tout d'abord, la capacité à stocker la chaleur multiplie les scénarios d'utilisation. Sans moyen de stockage, la création d'un réseau de chaleur ne serait possible que si à chaque instant la puissance produite est égale à la consommation. Le système de stockage introduit alors un élément de flexibilité important, permettant de répondre à tout moment aux besoins du consommateur en utilisant des sources de chaleur fluctuantes et irrégulières. Une telle flexibilité peut être mentionnée dans les centrales solaires à concentration incluant le stockage thermique à haute température (20). Par conséquent, il n'est pas nécessaire de convertir immédiatement la chaleur en électricité.

De plus, l'utilisation de sources de chaleur nécessite également des convertisseurs capables de convertir l'énergie thermique en électricité - et vice versa. On peut alors citer des solutions telles qu'un système ORC, une turbine à gaz ou un moteur Stirling.

Symétriquement, les réseaux électriques ont le même besoin de flexibilité pour répondre à la demande des consommateurs en utilisant un cluster d'usines de fabrication. En particulier, les sources d'énergie renouvelables variables telles que l'énergie photovoltaïque ou éolienne ont l'inconvénient de manquer de contrôle et de prévisibilité.

Le développement des réseaux multi-énergies

Les réseaux multi-énergie sont devenus un sujet de recherche majeur ces dernières années comme moyen innovant de gérer et consommer les énergies (18). Ce type de réseau contient un carrefour énergétique intégrant des systèmes énergétiques de sources variées - chimique, thermique, électrique - et des systèmes de conversions respectifs. Leur nature intrinsèquement souple suscite l'intérêt dans de nombreuses applications. Ils peuvent s'adapter très facilement aux fluctuations de productions et de demandes, à des échelles plus localisées, et répondent en ce sens à la nature variable des échanges énergétiques dans un réseau. En

outre, il est maintenant admis que l'utilisation de sources d'énergies renouvelables contribue de manière significative à la réduction des émissions polluantes et à l'amélioration de la qualité du cadre de vie dans les villes de demain (14). Par conséquent, la conception de tels systèmes énergétiques basés sur des critères propres et renouvelables permettrait de pénétrer ces ressources aisément tout en gommant leurs limites - en particulier l'intermittence et la faible densité énergétique - via des moyens de flexibilités intégrés, comme les stockages électrochimique et thermique susmentionnés.

Le couplage du réseau thermique et du réseau électrique, chacun avec des moyens de flexibilité adaptés, augmente ainsi fortement la qualité de dimensionnement du système énergétique global. Pour ce faire, la solution optimale du problème doit donc considérer l'approvisionnement énergétique dans son intégralité, et déterminer la stratégie optimale de gestion, en d'autres termes : quelle est la passerelle optimale entre différents vecteurs en fonction de sa production et de sa consommation en temps réel ?

Plusieurs travaux abordent ce couplage, de manière cependant incomplète. L'étude d'un réseau multi-énergies se décompose en plusieurs problèmes allant du contrôle en temps réel à la planification de la construction d'infrastructures.

Les études actuelles ne considèrent pas le problème d'optimisation de gestion couplée au dimensionnement du réseau

L'étude des réseaux multi-énergies est tout d'abord considérée par la littérature comme un problème de contrôle qui associe deux marchés distincts. Le couplage entre ces deux marchés passe par un carrefour énergétique – energy hub – qui permet la conversion entre les formes diverses d'énergie comme illustré dans la Figure 1.3 . Bien que ces travaux permettent des développements très avancés quant à l'optimisation de la performance globale, plusieurs aspects nécessitent d'y être développés. En effet le problème considéré, le plus souvent déterministe, concerne un seul pas de temps, ce qui exclut de fait l'optimisation de la stratégie de gestion d'un stockage, qu'il soit électrique ou thermique. De surcroît ce type d'étude est fait à notre connaissance systématiquement sur une infrastructure connue.

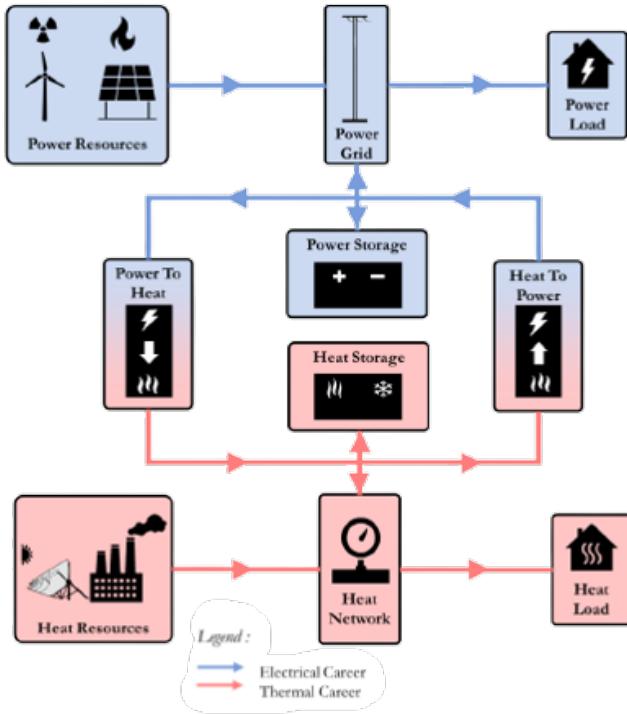


FIGURE 1.3 – Schéma de principe du carrefour énergétique étudié

Les travaux qui s'intéressent à l'optimisation de ce dimensionnement considèrent quant à eux des lois de gestion à chaque instant le plus souvent simplifiées. L'analyse de l'état de l'art montre ainsi que le couplage co-optimisation gestion de réseau et dimensionnement du réseau reste encore très faible. Or l'optimisation globale d'un système requiert de prendre en compte simultanément les deux horizons temporels. Cette prise en compte des décisions à chaque instant dès l'étape d'optimisation des infrastructures est fondamentale dans une approche de co-optimisation.

1.3.2 L'énergie solaire offre de très nombreuses applications et s'intègre parfaitement à un contexte d'utilisation dans un carrefour énergétique

L'énergie électrique reste le vecteur le plus utilisé actuellement et les thématiques allant de la génération, la transmission par un réseau, jusqu'à la gestion et le stockage, ont donc été traitées très largement dans la littérature. Les énergies renouvelables, et l'énergie solaire en particulier, ont connu un retard dans l'appréciation des avantages qu'elles apportent. Mais aujourd'hui, le développement de technologies telles que les systèmes énergétiques efficaces permet d'exploiter leur potentiel, en les intégrant dans des réseaux d'électricité, de gaz naturel et de chauffage urbain (18).

Une énergie non exploitée au maximum de son potentiel en France

Au vu du contexte énergétique, l'énergie solaire est une énergie qui peut offrir un rendement très intéressant dans des applications diverses, tant électriques que thermiques. En ce sens, cet aspect multi-modal colle parfaitement avec notre cadre multi-énergétique étant donné que l'utilisation de l'énergie peut se faire facilement autant par le vecteur électrique que thermique.

De plus, la pénétration de l'énergie solaire est une thématique actuelle prometteuse dans la communauté scientifique et de plus en plus d'articles sont recensés. Son abondance lui permet également d'être utilisée sur une grande superficie du globe. Elle devient la ressource renouvelable la plus populaire dans la communauté scientifique. Ceci se voit notamment par un nombre d'articles scientifiques grandissant (figure 1.4) ; traitant de ses possibles applications dans le domaine de la gestion énergétique.

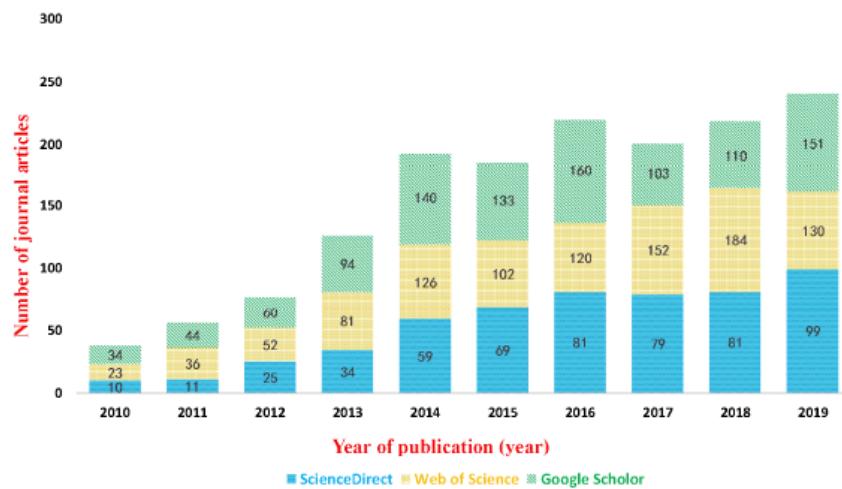


FIGURE 1.4 – Nombre d'articles scientifiques sur les applications de l'énergie solaire dans un contexte de gestion de réseaux d'énergies, en fonction des années

La volonté de développer l'énergie solaire vient également de décisions gouvernementales jouant un rôle prépondérant dans l'investissement de cette ressource. Pour répondre aux objectifs gouvernementaux fixés dans le plan de neutralité carbone avant 2030, RTE prévoit que la demande en énergie d'origine soutenable aura une tendance très forte ces prochaines années. Dans un rapport public, RTE chiffre un investissement entre 20 – 25 milliards d'euros par an sur 40 ans afin d'envisager d'alimenter le pays en électricité tout en atteignant les objectifs environnementaux. En particulier, cela revient à multiplier la production d'électricité d'origine solaire entre 7 et 22 fois, en fonction des scénarios possibles étudiés (10).

Son grand intérêt pose alors des questions d'intégration dans les réseaux électriques mais aussi thermiques, à travers respectivement des centrales photovoltaïques ou à concentrations solaires. Il est donc important de pouvoir prédire cette ressource pour des contraintes de réseaux, telles que la production, de stockage éventuellement, mais aussi des questions de transmissions dans le réseau. En l'occurrence, des variabilités de puissance peuvent induire de grandes instabilités dans le réseau ce qui peut diminuer la robustesse de l'approvisionnement voire détériorer le réseau. La plupart des gestionnaires de réseaux se basent sur le rayonnement solaire - précisément l'irradiance solaire- quand il s'agit de contrôler les ressources solaires en temps réel. L'irradiance solaire correspond au flux d'énergie solaire reçu par unité de surface, et se mesure en W/m^2 . Les modèles de prédiction d'énergie solaire figurant dans la littérature l'utilisent comme donnée d'entrée principale aux algorithmes (9; 28), et peuvent la combiner avec d'autres variables de l'ordre météorologique (température, pression, humidité, taux de couverture nuageuse...).

Or, l'irradiance solaire est une grandeur complexe à mesurer par les interactions entre l'atmosphère et le rayonnement solaire direct. Par conséquent, la ressource solaire au niveau du sol est très variable, principalement en raison de la variabilité de la couverture nuageuse, des niveaux d'aérosols atmosphériques et, indirectement et dans une moindre mesure, des gaz participants dans l'atmosphère (15). La prédiction d'irradiance solaire

est donc un axe de recherche à fort potentiel scientifique qui pourrait déboucher sur de vastes applications importantes dans la gestion de la ressource de réseaux énergétiques. En l'occurrence, en fonction des applications- parcs de panneaux photovoltaïques ou centrale à panneaux solaires - la variable d'intérêt peut encore se décliner sous deux formes : l'irradiance globale et directe. L'irradiance globale comprend à la fois l'irradiance directe issue du "flux" solaire, et l'irradiance diffuse, variant selon la topologie du terrain. Elle est d'avantage utilisée dans des contextes de génération d'électricité par panneau photovoltaïque, car la diffusion est une composante qui reste importante. Alors que dans des contextes de production par centrale solaire, on préfère prédire uniquement le rayonnement qui contient la grande majorité de l'énergie solaire convertible.

Malgré un besoin réel de la gestion de l'énergie solaire, des études portant sur l'élaboration de bases de données de références d'irradiance solaire nécessitent d'être encore faites, contrairement au domaine de l'éolien déjà largement documenté (13; 25). Ce manque de données clé en main s'explique par une expansion du marché de l'énergie solaire et photovoltaïque plus tardive que l'éolien ou l'électrique (21). En effet, la performance des modèles sera toujours bornée par la qualité des données, et indépendamment de l'application finale (1; 19; 30) : il est donc important de pouvoir s'appuyer sur des données corrigées et de bonne qualité en vue de diminuer la propagation d'éventuelles erreurs. La qualité des données peut se présenter comme une réelle limite dans certains cas, notamment dans les milieux industriels où l'acquisition des données peut être un défi en soi.

Cette gestion de l'énergie solaire dans un réseau multi-énergie nécessite des modèles de prédiction plus ou moins fins pour répondre à l'objectif de co-optimisation (gestion et dimensionnement couplés) d'un tel réseau dans un contexte stochastique (qui présente des incertitudes). C'est précisément l'une des problématiques abordée dans la thèse de mon tuteur, dont les objectifs sont détaillés ci-dessous.

1.3.3 Un stage s'inscrivant dans un travail de recherche

Une thèse CIFRE est menée depuis 2018 par Ibrahim Al Asmi, dans le voeu d'élaborer des outils et des méthodes de co-optimisation d'un réseau hybride thermique et électrique.

A ce jour, les algorithmes commandant les réseaux ne proposent qu'une approche déterministe. La thèse a pour objectif d'étudier la pertinence de considérer ce problème d'optimisation par des méthodes stochastiques. Les travaux réalisés serviront à l'élaboration de modèles et d'outils nécessaires à une co-optimisation sur cycle de vie d'un réseau hybride multi-énergie, le but ultime étant d'analyser les conditions de pertinence d'un réseau hybride avec stockage décrit dans Figure 1.3. Pour la résolution de ce problème d'optimisation, un algorithme MPC (Model Predictive Control) a été développé et discuté dans les articles (3; 4; 16).

Résolution de la co-optimisation par MPC

Il est possible de résoudre le problème de co-optimisation via des méthodes de contrôle prédictif, dite MPC (Model Predictive Control).

Le MPC est basé sur l'optimisation itérative à horizon fini d'un système donné, ici l'energy hub décrit. À l'instant t , l'état actuel de l'installation est échantillonné. De plus, une stratégie de contrôle est mise en place, minimisant la fonction coût du système, pour un horizon temporel dans le futur. En effet, un calcul en temps réel est mis en oeuvre afin de trouver la stratégie de contrôle optimale, au sens où elle minimise la fonction coût du système dans l'horizon temporel de prédiction. Seule la valeur optimale au premier pas de temps futur $t + 1$ est conservée, puis l'état du système est à nouveau échantillonné et les calculs sont répétés à partir du nouvel état actuel, ce qui donne un nouveau contrôle et

un nouveau chemin d'état prédit, et l'horizon de prédition continue à être déplacé vers l'avant de manière itérative.

En fonction des entrées prédites que l'on anticipe via des modèles de prédition, par exemple des profils de consommations et de productions d'énergies renouvelables, le contrôleur devra donc fournir la commande optimale sur les flexibilités relative à leur stockage, ce qui revient à donner les commandes de puissance à stocker ou déstocker sur les stockages électriques et thermiques en fonction de l'état échéant du système.

Cette thèse a pour but de caractériser toutes les briques composant le réseau hybride 1.3.

Un stage autour de l'impact des modèles de prévision sur la co-optimisation

Pour l'instant, seules des études de co-optimisation déterministes ont été menées, au sens où les modèles de prédition implémentés dans le contrôleur via la résolution MPC ne fournissent que des prédictions "ponctuelles", sans intervalle de confiance. Aux modèles de prédictions déterministes, on préfère les modèles probabilistes pour la prise de décision par le gestionnaire de réseau (17). En effet ces modèles offrent des informations précieuses quant à la confiance que l'on peut avoir dans les trajectoires prédites, et peuvent fournir un nombre important de scenarii différents, là où des méthodes déterministes même performantes ne donneront qu'une sortie pour un ensemble de données d'entrée seulement. Ces modèles à l'appui, des méthodes de MPC stochastiques basées sur plusieurs scenarii de production et/ou de consommation pourront ensuite être étudiées.

Une fois de tels modèles de prédition implémentés, une problématique clé est d'étudier l'impact de ces modèles sur la co-optimisation d'un réseau hybride. Autrement dit, il s'agirait de quantifier comment la solution optimale de co-optimisation du problème se dégrade lorsque les observations futures sont tout ou partie non observées réellement, et remplacées par des observations prédites par de tels modèles. Cette étude est primordiale afin de choisir les meilleurs modèles présentant le compromis parfait entre précision et temps de calcul de prédition pour un contrôleur qui soit le plus performant possible. Cette problématique et les verrous techniques associés figurent comme une question qui n'a pas encore été traitée, ce qui naturellement motive le sujet de mon stage actuel. Ainsi, ma mission principale fut d'élaborer un générateur des scenarii stochastiques de production-consommation thermique dans un contexte de réseau hybride. Pour différentes familles de modèles stochastiques que l'on se donne, il a fallu trouver le meilleur modèle et fournir des critères d'évaluation justifiés. Cette contribution s'intégrera par la suite dans des études plus globales dans l'impact de ces modèles dans la co-optimisation.

La prévision solaire précise est d'une grande importance pour les carrefours énergétiques. Pour les raisons mentionnées, mon stage s'est concentré sur la prédictions d'irradiance solaire afin d'apporter des éléments de réponses quant à la bonne famille de modèles adaptée à ce type d'application.

La section suivante explique la routine de recherche de modèles de prévision qui a été développée au cours de ce stage. De plus, deux familles de modèles - SARIMA et Markov caché (HMM pour Hidden Markov Model) - ont été étudiées et évaluées via la routine développée. Enfin, les résultats de comparaison entre les différents modèles obtenus seront explicités et discutés. Les perspectives d'utilisation de cette contribution pourront s'intégrer dans des études plus élaborées, en commençant par une étude sur l'impact de ces modèles dans un réseau thermique mono-vecteur, puis hybride électrique-thermique par exemple.

Dans les chapitres suivants, je vais détailler mes contributions dans le cadre de la co-optimisation de réseaux hybrides.

2

Etat de l'art et Méthodologie

2.1 État de l'art sur les méthodes de prédictions et les enjeux par rapport au contexte de recherche

2.1.1 L'importance des données

Avant de rentrer dans les détails du contrôle optimal de ressources énergétiques, il est capital de cibler quelle type de données voudrions-nous piloter. Sa nature va ainsi influencer le type d'application, la formulation du problème mathématique et requiert éventuellement du pré-traitement voire du post-traitement, pour pouvoir manipuler efficacement cette donnée. La qualité et la quantité de la donnée va inéluctablement tracer un cadre de méthodologies potentiellement réalisable ou non en fonction des outils en main.

On peut distinguer les en fonction du contexte d'utilisation : sont-elles issues de production industrielle (puissance électrique de parcs photovoltaïques, puissance thermique d'un four solaire etc...), ou de consommation(réseau de chaleur etc...) ?

2.1.2 Un accès qualitatif aux données industrielles encore limité

Néanmoins le manque de recul et des problèmes d'acquisition aux données rendent l'accès à une base de donnée de référence encore difficile. Nous avons peu de résultat sur la pré-diction de série temporelle de chaleur fatale ; et encore moins sur des modèles généralisant le comportement de plusieurs d'industries à la fois. Un travail axé sur des données de production de chaleur fatale industrielle se montre donc peu réalisable aujourd'hui.

Une possible solution est donc de s'appuyer sur des données opensource fiables, issues d'études gouvernementale ou universitaire. Les données gouvernementales peuvent requérir du pré-traitement et sont souvent à l'échelle nationale, ce qui peut être intéressant lorsqu'on mène des études plus génériques.

Cependant, de telles données en quantité mais plus localisées - par exemple à l'échelle d'une centrale - sont plus rares mais nous intéresseraient d'avantage, au vue du champ d'action de notre étude. De telles données existent pour différents contextes tant pour la consommation thermique (23), que pour la production thermique issue d'énergie solaire (21). Ces bases de données permettent d'utiliser ces données telles quelle comme point de référence.

2.1.3 Une revue très générale des méthodes de prédictions sur l'irradiance

Les méthodes d'irradiance (9) utilisent communément les données d'irradiance issues de mesures satellites, comprenant l'irradiance globale horizontale (GHI), qui peut se décomposer en deux parties : une irradiance directe normale (DNI) issue du flux solaire, et une irradiance diffuse (DIF) issue des interactions avec l'environnement dans l'air et au sol (nuages, bâtiments ...). En fonction du type d'application de l'irradiance, on préférera utiliser l'irradiance globale ou l'irradiance directe. A cette variable d'intérêt primaire, d'autres données caractéristiques complémentaires telles que météorologique (couverture nuageuse, humidité dans l'air, température ...) peuvent être des entrées aux algorithmes.

Ces données d'entrées peuvent en premier lieu être pré-traitées pour les ajuster en fonction des contraintes et hypothèses des modèles. On dénote 7 méthodes de pré-traitements principales dans la littérature (12) :

Méthodes	Année	Applicabilité	Avantages/inconvénients
Stationnarisation (24)	2000	données de tailles faible avec petites fluctuations	+ Adaptatif - Ne convient pas à des données trop complexes
Séries sans tendances (5)	2007	déterminer la tendance du rayonnement solaire	+ Insensible à la taille des données + Implémentation simple
Modèle ciel clair (6)	2009	données de taille faible avec fluctuations relativement moyenne	- Peu robuste + Temps de calcul rapide - Présence d'information redondante
Normalisation (26)	2012	données de taille moyenne avec haute redondance	+ Adaptatif + Réduit l'information redondante - Ne conserve pas les variations temporelles des données
Classification par apprentissage (29; 2)	2014	données de taille moyen voire grande avec des fluctuations nettes	+ Adaptée aux données non-linéaires + Simple et temps de calcul relativement rapide
Réduction de dimension par transformée d'ondelette (22)	2015	convertit un grand ensemble de données en un ensemble de métadonnées plus petit	- Peu robuste et précision variable + Réduction temps de calcul + Augmente considérablement la précision des modèles - Implémentation plus complexe

Une étude récente (28) a contribué à une précieuse revue de toutes les familles de méthodes de prédictions d'irradiance solaire à ce jour. 128 méthodes de prévision solaire sont résumées et comparées de manière exhaustive en fonction de leurs entrées, de leur résolution temporelle, de leur résolution spatiale, de leurs variables de prévision, de leurs paramètres et de leurs caractéristiques. Cette étude poussée pourra aider les lecteurs à utiliser ces méthodes de manière plus efficace dans le cadre de futures recherches approfondies. 30 critères d'évaluation sont également résumés pour une évaluation juste et pertinente des prévisions solaires.

On peut classer ces méthodes autour de grandes familles de modèles 2.1, qui ont chacune une portée temporelle et spatiale propre :

1. Les modèles physiques :

Ces modèles collectent des données physiques via des modules de détection local, des imageurs, ou de télé-détection par satellite. Les méthodes de cette nature ont l'avantage d'avoir une portée spatiale très grande en moyenne par rapport aux autres classes de modèles. Toutefois, elles demandent des infrastructures spécifiques ce qui limite leur déploiement.

Les méthodes numériques de prévisions de temps (Numerical Weather Prediction NWP) résolvent les équations différentielles des lois de la dynamique sur la base des données physiques collectées. La simulation numérique permet de prédire des variables telles que l'irradiance solaire mais également la couverture nuageuse. Des imageurs locaux sont des caméras digitales qui photographient le ciel en temps réel, puis prédisent les données l'irradiance et l'évolution de la couverture nuageuse. via des modèles très simples de persistance Ils ont une résolution temporelle très fine,

ce qui permet de les utiliser dans des contexte de gestion de réseaux intra-horaire, avec une résolution de l'ordre de quelques minutes pour la plupart des cas. Des imageurs satellites sont similaires aux imageurs au sol.

2. Les modèles statistiques :

Le modèle statistique ne nécessite pas que le système fournisse des informations internes au modèle lui-même, qui sont sur la base d'un l'apprentissage du modèle de prévision avec des données variables. En outre, leur précision de prévision dépend de la longueur et de la qualité des données d'entrée historiques (1), qui sont subdivisées en deux groupes, à savoir les méthodes régressives et les méthodes d'intelligence artificielle (IA).

Concernant les méthodes régressives, on note la classe des méthodes auto-régressives et à moyenne mobile ARMA (Auto Regressive Moving Average), qui sont un type de modèles classique dans la prédiction de séries temporelles (voir paragraphe 2.3.1). Elles sont faciles à implémenter et peuvent présenter des résultats relativement bons, c'est pour cette raison qu'elles sont souvent utilisés comme modèle témoin à comparer avec d'autres modèles plus fins. Des variantes plus sophistiquées existent incluant des variables exogènes (ARMAX), ou adaptés au séries non-stationnaire avec une tendance (ARIMA), voire avec de la saisonnalité (SARMA). Ces méthodes requiert l'hypothèse de stationnarité ce qui demande du pré-traitement au préalable qui peut s'avérer relativement coûteux en fonction de la dynamique des données.

Les méthodes d'intelligence artificielle et d'apprentissage stochastique sont très utilisées ces dernières années au vu du développement croissant des ressources de calculs mis à disposition. Beaucoup d'articles ont contribué à l'implémentation de telles techniques dans un contexte de gestion d'énergies, allant des réseaux de neurones, aux machines à vecteur de support (Support Vector Machine) et des méthodes de logiques floues (fuzzy logic).

Les modèles statistiques présentent plusieurs avantages, tels que la tolérance au bruit, et la capacité à résoudre des problèmes non linéaires pour les méthodes d'IA.

Par conséquent, ces méthodes sont couramment utilisées dans les études de prévisions solaires, en particulier pour les prévisions à court terme (9) qui nécessitent une haute résolution temporelle et une vitesse de traitement rapide.

3. Les modèles hybrides :

On constate notamment que la plupart des chercheurs se sont concentrés sur des modèles uniques pour les prévisions solaires. Néanmoins, la performance d'un modèle unique n'est pas fiable dans la prévision solaire dans différents cas. L'une des motivations du développement de modèles hybrides est que la précision des prévisions peut souvent être améliorée en profitant des avantages de chaque méthode. Par conséquent, construire de tels modèles qui mélange des modèles et des infrastructures physiques avec des modèles par apprentissage peut être une solution.

Les modèles hybrides présentent une très grande diversité de méthodes permettant de couvrir tout le champ de résolution spatial et temporel des applications en questions. Néanmoins, l'étendue des méthodes étant vaste et pour certaines assez complexes à implémenter voire à déployer car nécessitant du matériel, elles peuvent être un pan de recherche à elles seules. C'est pour cette raison que dans le cadre de notre étude nous allons nous cantonner à appliquer certains modèles statistiques, à caractériser leur performance dans un contexte de prédiction d'irradiance solaire.

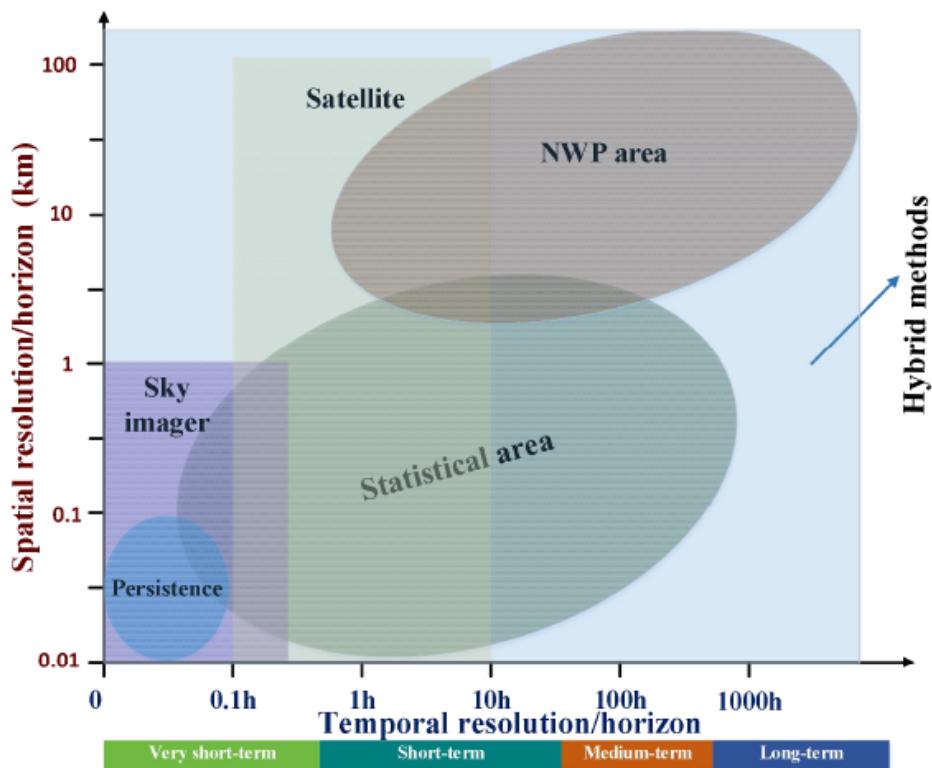


FIGURE 2.1 – Distribution des familles typiques de modèles, en fonction de leur résolution temporelle et spatiale (28)

2.2 Cas d'étude

2.2.1 Formulation du problème et du cadre du stage

Nous allons implémenter des modèles de prédictions d'irradiance solaire issus de la plateforme universitaire Renewable Ninjas¹, fournissant entre autre toutes les variables précises nécessaires, traitées et réanalysées dans un périmètre de 50km par 50m (21). Les variables d'intérêts sont la radiation globale à la surface du sol en $\frac{W}{m^2}$ ainsi que le rayonnement "top of atmosphere" utile pour pré-traiter et stationnariser la donnée. Nous utiliserons les profils d'irradiance solaire captées à Odeillo au sud de la France où est actuellement situé le four solaire installées par le laboratoire CNRS-PROMES. La topographie confère en effet une très grande période d'ensoleillement même en hiver.

1. <https://www.renewables.ninja/>

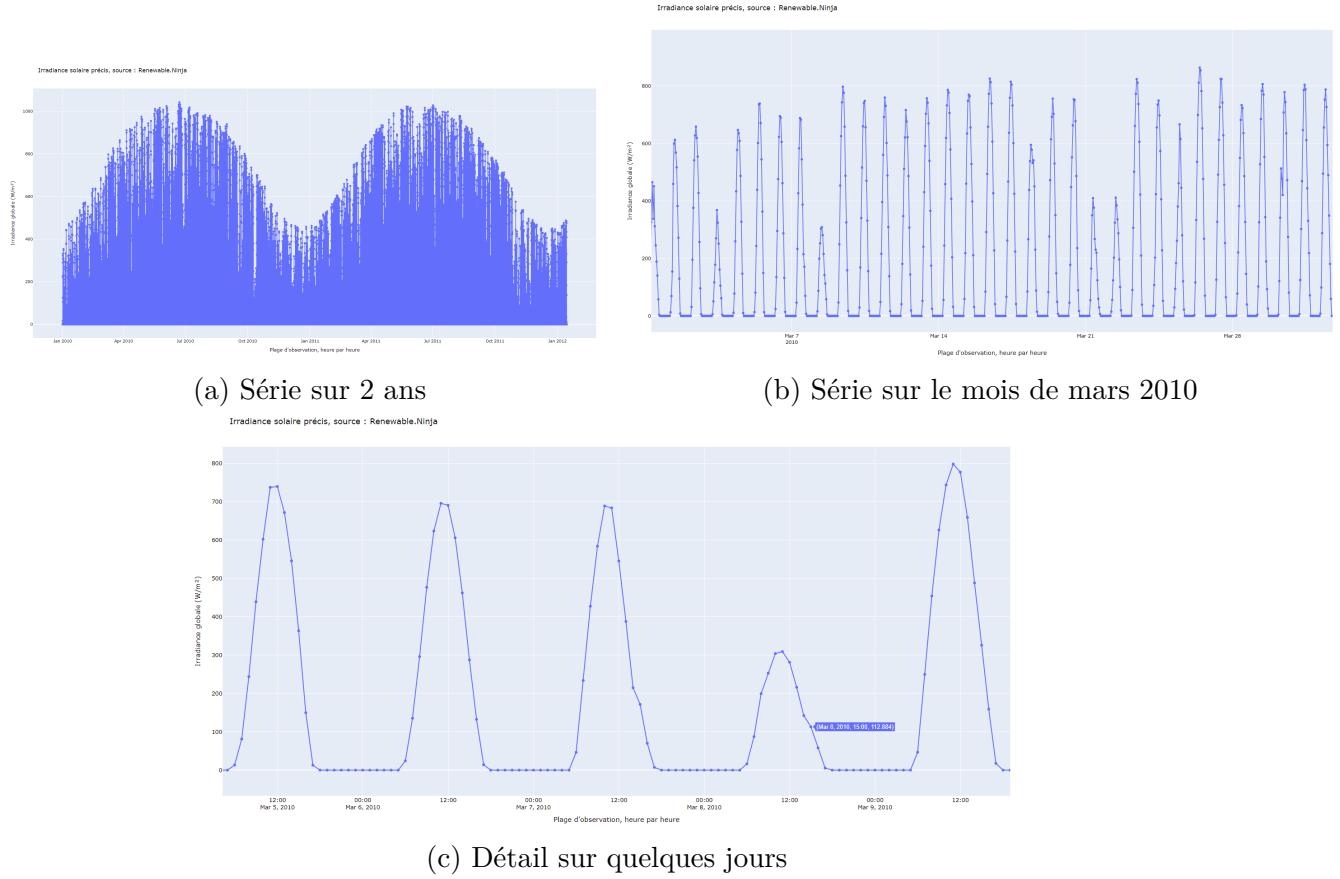


FIGURE 2.2 – Dynamique du profil d’irradiance solaire globale sur différentes périodes de temps issues du four solaire d’Odeillo CNRS-PROMES

Prédire l’irradiance solaire globale - directe et diffuse- est un choix motivé par le fait que cette variable est très utilisée dans des contextes de centrales photovoltaïques. On ne se restreignera pas pour autant aux applications éventuellement en perspectives, telles que les centrale à concentration solaires qui elles n’ont besoin que de l’irradiance directe. En effet, à partir de l’irradiance globale GHI, il est possible d’estimer facilement les autres composantes (directe et diffuse) à travers des modèles de radiométrie tel que le modèle de Boland-Ridley-Lauret (8).

L’irradiance reste une variable requise pour des réseaux divers avec précision. Parmi les variables d’entrées présentées 1.3.2, l’irradiance globale sera la grandeur prédite par les modèles que nous allons construire. L’irradiance prédite sera utilisée en entrée du contrôleur afin d’inférer la décision optimale. Il est possible par un modèle annexe de convertir l’irradiance en puissance à la sortie d’une centrale (31) - par un modèle de rendement d’une centrale photovoltaïque si on se situe dans un contexte électrique, ou celui d’un four à concentration solaire pour des applications thermiques.

2.2.2 Un choix de modèles de prédictions constraint entre temps de calculs limités et performance prédictive

La formulation telle quelle de notre problème engage une prédition des profils d’irradiance sur 24h (cf § 1.3.3), nommée dans la littérature comme "day-ahead forecast". Cette granularité temporelle reste assez fine comparé à des horizons de prédictions pour de la planification pouvant aller à plusieurs année de simulations. Cela nous oblige à rester dans un cadre somme toute exigeant sur les performances prédictives des modèles.

D’un autre côté nous recherchons des modèles de prédictions pouvant fournir les entrées

aux algorithmes de contrôles de réseaux tous les quinzaines de minutes. De modèles trop complexes bien que très performants existent dans la littérature - Markov Switching, réseaux de neurones classifieurs etc... - mais dépassent malheureusement notre cadre de travail. Des modèles légers en déploiement sont donc préférables pour assurer une gestion en temps réel. Parmi les modèles statistiques présents dans la littérature, on choisit d'implémenter deux modèles : une méthode de régression avec saisonnalité journalière SARIMA, et l'implémentation de modèle de Markov caché à émission gaussienne multivariée (7).

Ces modèles ont également l'avantage fort d'être des méthodes stochastiques et de fournir plusieurs prédictions de profils d'irradiance avec leur spécificité.

2.2.3 Objectifs du stage

La contribution principale de ce stage se résume en une étude et une discussion quant aux choix de familles de modèles sur un problème physique donné, ici la prédiction de données thermiques d'irradiance solaire.

En particulier, il s'agira de construire un générateur de scénario aléatoire pour des séries temporelles dans un contexte de ressource thermique. Ce générateur de scénario aléatoire devra fournir des prédictions de plusieurs modèles (HMM et SARIMA), et donner des critères de performance pour caractériser ces prédictions. Une recherche de meilleur modèle dans la famille donnée sera également menée dans le but de calibrer au mieux les hyper-paramètres associés à chacun des deux types de modèles.

Les études de ce stage pourront être réutilisées et approfondies dans l'étude de l'impact des modèles de prédictions dans la co-optimisation de réseau multi-énergie.

2.3 Présentation des modèles de prédictions utilisés

2.3.1 Modèle régressif SARIMA

En statistique, les modèles ARMA (AutoRegressive and Moving Average models), sont les principaux modèles de séries temporelles.

En se donnant une série temporelle X_t , le modèle ARMA est un outil pour comprendre et prédire, éventuellement, les valeurs futures de cette série. Le modèle est composé de deux parties : une partie autorégressive (AR) et une partie moyenne-mobile (MA). Le modèle est généralement noté ARMA(p,q), où p est l'ordre de la partie AR et q l'ordre de la partie MA. Un processus temporel X_t peut donc être modélisé par un processus ARMA(p,q) comme suit :

$$X_t = \varepsilon_t + \sum_{i=1}^p (\varphi_i \varepsilon_{t-i}) + \sum_{i=1}^q (\vartheta_i X_{t-i})$$

où les φ_i et ϑ_i sont les paramètres du modèle et les ε_i les termes d'erreur tel que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. La première étape est de déterminer les ordres p et q du modèle. Le modèle SARIMA(p,q,s) ("S" pour "Seasonal") similaire au modèle ARMA(p,q) est plus adapté pour les séries avec une saisonnalité s. Il ne prend en compte que les termes antérieurs, multiples de s.

$$X_t = \varepsilon_t + \sum_{i=1}^p (\varphi_i X_{t-si}) + \sum_{i=1}^q (\vartheta_i X_{t-si})$$

Une des grandes questions dans l'étude de séries temporelles est de savoir si celles-ci suivent un processus stationnaire. On entend par là le fait que la structure du processus sous-jacent supposé évolue ou non avec le temps. Si la structure reste la même, le processus est dit alors stationnaire. Si la série ne l'est pas, il est important de la stationnariser avant d'appliquer le modèle ARMA(p,q). Une manière de la stationnariser est d'estimer les différentes tendances non stationnaires, puis de les retrancher. Les modèles ARIMA(p,d,q) sont des variantes incluant un terme de différenciation d'ordre d entre les termes $X_t, X_{t-1}, \dots, X_{t-d}$, qui permet de diminuer la non stationnarité de la série. Ces processus sont donc englobés dans la famille des modèles $SARIMA(p, d, q)x(p_s, d_s, q_s, s)$, avec p, d, q : les termes d'autorégression, de moyenne mobile et de différenciation non saisonniers ; et p_s, d_s, q_s : les termes d'autorégression, de moyenne mobile et de différenciation saisonniers de saisonnalité s .

Par la suite, on stationarisera les données notamment les données d'irradiance.

Pour savoir si la série d'étude vérifie la propriété de stationnarité, on a recours au test augmenté de Dickey-Fuller. Ce test vérifie si l'hypothèse de stationnarité est valide, nous donnant une p-valeur p associée associé à la conviction que l'on peut porter sur l'acceptation de l'hypothèse. Plus la p valeur est petite, plus on peut accepter l'hypothèse avec conviction.

Un moyen pour déterminer approximativement les ordres p, q est de recourir aux graphes d'autocorrélation et d'autocorrélations partielles. Ces graphes permettent de développer une première intuition sur le modèle qui correspondrait a priori aux données mis à disposition, mais n'est en aucun cas une méthode rigoureuse.

Nous choisissons de construire différents modèles d'ordre différents et les comparer par un critère de vraisemblance comme un critère d'information d'Akaike (AIC) ou de Bayes (BIC). Cette manière permet comparer les modèles candidats parmi la famille SARIMA plus finement.

Par analogie, on peut voir les méthodes SARIMA comme des processus markovien d'ordre correspondant à l'ordre maximum des décalages.

2.3.2 Modèle Markov Caché multivariée à émission gaussienne :

Dans un modèle de Markov caché (Hidden Markov Model), on ne peut pas observer directement les états du processus, mais des symboles émis par les états selon une certaine loi de probabilité. On considère qu'une observation, ic un profil journalier d'irradiance, est un "symbole" émis par un état non accessible suivant une chaîne de Markov. Les états cachés suivent un processus markovien, défini par une matrice de transition entre les états, à chaque itération discrète du système, ici chaque jour étant une itération dans la chaîne de Markov des états cachés.

L'application d'un tel modèle revient à dire que chaque jour, un profil d'irradiance est émis par un état. Ces états-clusters cachés ne sont pas observables directement mais on peut retrouver la séquence d'états cachés la plus vraisemblable à partir des observations émises. Ils sont dénombrables et le nombre d'états cachés N un hyperparamètre à optimiser. On part d'un état caché initial X_0 , qui représente une famille de profil journalier ayant une cohérence, cet état "émet" un profil d'irradiance O_0 , qui fera office d'observation. Ce profil serait un vecteur aléatoire gaussien de 24 valeurs correspondant pour chaque heure à l'irradiance observée ou prédictive, en fonction du point de vue. A partir de la distribution courante des probabilités des états, et de la matrice de transition A , on met à jour ce vecteur des distributions à partir duquel on tire un état X_1 qui sera l'état caché du jour

suivant. Ensuite, on tire selon la loi d'émission de l'état caché un profil d'irradiance O_1 , et on répète le process autant de fois que nécessaire.

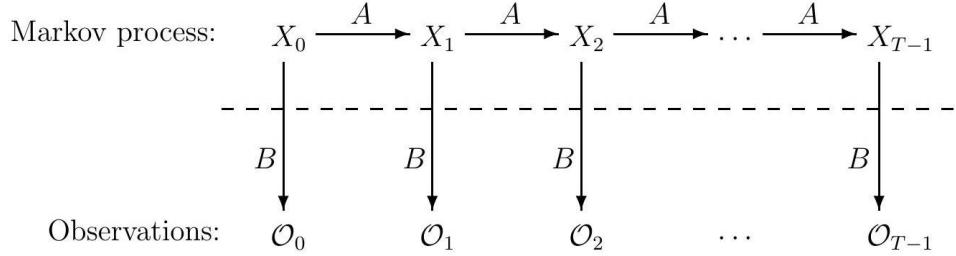


FIGURE 2.3 – Schéma de principe d'un modèle HMM, les états cachés x_i émettent à chaque transition de jour un profil selon la matrice A d'irradiance journalier o avec une probabilité B

Les modèles HMM prennent en compte la structure de la donnée et traduisent cette information en des clusters partageant des informations communes, qui peuvent avoir un sens physique ou non, avec des caractéristiques propres au cluster qui influe sur la forme du profil observé. On pose toutes les notations ci-dessous utiles pour la suite du rapport.

Notation :

- T = longueur de la séquence d'observation
- N = nombre d'états dans le modèle
- M = nombre d'observations émises
- d = dimension des observations
- $Z = \{z_0, z_1, \dots, z_{N-1}\}$ = ensemble d'états cachés
- $\Theta = \{\theta_i\}_{i=1..N}$ = ensemble des paramètres d'émissions $\theta_i = (\mu_i, \Sigma_i)$ de chaque état caché
- $O = \{o_0, o_1, \dots, o_{M-1}\}$ = ensemble des observations possibles
- A = matrice des probabilités de transition d'état
- $b_i(o_t) = p(o_t | q = i)$ probabilité d'émission de l'observation par l'état caché
- π = distribution de l'état initial
- $\lambda = (A, \Theta, \pi)$ = ensemble des paramètres du modèle HMM

Les observations sont des profils d'irradiance de dimension $d = 24$, à pas de temps horaire. On fait l'hypothèse qu'une observation suit une loi multivariée de taille d gaussienne $\mathcal{N}_d(\mu_i, \Sigma_i)$ issue du i ème état caché parmi les N tel que :

$$b_i(o_t) = p(o_t | q = i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (o_t - \mu_i)^T \Sigma_i^{-1} (o_t - \mu_i) \right\}$$

On fait l'hypothèse que la matrice de covariance est diagonale, ce qui simplifie le déterminant $|\Sigma_i|$ en le produit des valeurs propres, et l'inverse Σ_i^{-1} la matrice diagonale des inverses de valeurs propres.

Les modèles HMM permettent de résoudre trois problèmes :

Problème n°1 : calcul de vraisemblance d'une observation

Connaissant totalement le modèle, λ totalement déterminé, on souhaite calculer la probabilité d'apparition d'une observation, autrement dit on cherche à estimer la "vraisemblance" : $P(o_t|\lambda)$

Ce premier type de problème peut se résoudre par un algorithme itératif nommé l'algorithme "forward".

Problème n°2 : décodage

Étant donnée $\lambda = (A, B, \pi)$ et une séquence d'observations \mathcal{O} , trouvez une séquence d'état optimale pour le processus de Markov sous-jacent. En d'autres termes, nous voulons remonter aux états cachés à partir du modèle de Markov caché. Un algorithme classique est l'algorithme de Viterbi ou "forward-backward". Ce problème peut être utile dans notre contexte de contrôle optimal. A partir des données observées, on peut remonter à l'état caché le plus vraisemblable qui aurait émis ces observations, et prédire les pas de temps futurs en tirant des observations issues de cet état caché.

Néanmoins, cela demande de connaître les paramètres du modèle, ce qui nous amène au dernier problème-type, l'estimation des paramètres d'un modèle avec pour seule information les observations du passé.

Problème n°3 : apprentissage du modèle

Ce problème permet d'identifier ou du moins d'estimer λ , à partir d'une séquence d'observations donnée \mathcal{O} .

$$\lambda^* = \arg \max_{\lambda} P(\mathcal{O} | \lambda)$$

Nous allons détailler la résolution mathématique du problème n°3, problème qui nous intéresse principalement.

Comme il s'inspire des algorithmes forward et backward, on laisse au lecteur, s'il le souhaite, la possibilité de regarder plus en détail les algorithmes forward-backward dans la référence en bas de page ¹.

L'algorithme de Baum-Welch est une adaptation de l'algorithme expectation-maximization (EM) pour les modèles HMM. L'algorithme EM est une approche itérative de la méthode de maximisation par vraisemblance de paramètres inconnus. On décrit très brièvement les deux principales étapes de l'algorithme adapté pour l'estimation des paramètres de modèle HMM multivarié. L'objectif premier de ce stage n'était pas d'implémenter une telle méthode, somme toute assez ardue, on laisse encore une fois au lecteur le soin de lire les détails ².

Sans rentrer dans le détail de la formule de la vraisemblance \mathcal{L} , elle dépend de deux paramètres : λ les paramètres du modèles, q la probabilité d'occurrence de la séquence d'états cachés z , conditionnélement aux observations o et à λ .

$$\sum_z q(z | \mathcal{O}) \log \frac{p(x, z | \lambda)}{q(z | x)} = \mathcal{L}(q, \lambda)$$

1. <https://web.stanford.edu/~jurafsky/slp3/4.pdf>

2. <https://people.eecs.berkeley.edu/~jordan/courses/281A-fall04/lectures/lec-10-26.pdf>

Nous pouvons maximiser ce produit en utilisant l'algorithmme EM. C'est un processus en deux étapes pour l'estimation du maximum de vraisemblance lorsque la fonction de vraisemblance ne peut pas être calculée directement.

La première étape "**Expectation**" consiste alors à calculer la valeur attendue de la fonction de log-vraisemblance par rapport à q la distribution conditionnelle des états cachés z étant donné la séquence \mathcal{O} et les paramètres λ . La deuxième étape "**Maximization**" consiste à trouver les paramètres qui maximisent cette fonction, en calculant le gradient de la log-vraisemblance estimée.

Ces estimations de paramètres sont ensuite utilisées pour réimplémenter l'algorithme forward-backward et le processus se répète de manière itérative jusqu'à la convergence ou un certain nombre d'itérations spécifié. Ici, on se donnera comme critère d'arrêt une tolérance à l'écart du gradient de la log-vraisemblance entre deux itérations de l'étape M. Par ailleurs, il est important de noter que l'étape de maximisation est une optimisation locale autour de la meilleure estimation actuelle de λ .

En d'autres termes, pour tout $q(z|o) = p(z|o, \lambda)$:

étape E :

$$q^{(t+1)}(z | \mathcal{O}) = \arg \max_q \mathcal{L}(q, \lambda^{(t)})$$

étape M :

$$\lambda^{(t+1)}(z | \mathcal{O}) = \arg \max_{\lambda} \mathcal{L}(q^{(t+1)}, \lambda)$$

Une autre façon d'écrire l'étape M :

$$\begin{aligned} \mathcal{L}(q^{(t+1)}, \lambda) &= \sum_z q^{(t+1)} \log q^{(t+1)} + \sum_z q^{(t+1)}(z | \mathcal{O}) \log p(x, z | \lambda) \\ \lambda^{(t+1)}(z | \mathcal{O}) &= \arg \max_{\lambda} \sum_z q^{(t+1)}(z | x) \log p(\mathcal{O}, z | \lambda) \\ &= \arg \max_{\lambda} E_{q^{(t+1)}}[\log p(\mathcal{O}, z | \lambda)] \end{aligned}$$

$E_{q^{(t+1)}}[\log p(\mathcal{O}, z | \lambda)]$ étant l'espérance de $\log p(\mathcal{O}, z | \lambda)$.

Les modèles HMM présentés, ainsi que les algorithmes EM utilisés, proviennent d'une librairie python¹, qui a été réutilisée en grande partie pour l'optimisation de tels modèles dans la suite du stage.

2.4 Méthodologie

Nous allons détailler la méthodologie générique pour les deux modèles de prédiction ; à savoir la recherche de meilleur modèle et l'évaluation des prédictions. Il a fallu rechercher des librairies, les prendre en main et comprendre les fonctionnalités, et adapter les méthodes existantes à notre contexte, voire les réécrire.

Une discussion sur les critères d'évaluation a donc été faite au vu des types des modèles et des critères de performances voulues. En premier lieu, les routines implémentées ont été testées sur des données artificielles générées à partir de modèle de référence pour valider rigoureusement la performance des méthodes et écarter toutes potentielles sources d'erreur. Finalement, une fois les routines mises en places, nous les avons appliquées sur les données d'irradiance. Les routines seront expliqués par la suite plus en détail.

1. <https://github.com/hmmlearn/hmmlearn>

2.4.1 Pré-traitement des données

Tout d'abord, il n'existe pas de meilleur modèle de prédiction dans l'absolu, cela dépend avant de la nature des données, et des critères d'évaluation que l'on se donne. Il est donc important de bien connaître les hypothèses d'implémentations des modèles pour pouvoir prétraiter les données comme il se doit. Ainsi, nous avons dû stationnariser les données d'irradiance pour pouvoir apprendre ensuite un modèle de régression SARIMA.

Concernant le modèle HMM, un tel pré-traitement n'est pas nécessaire, mais nous avons appliqué une transformation par quantile afin d'avoir une distribution se rapprochant le plus possible d'une normale. Le modèle HMM en question fournissant des prédictions suivant une loi gaussienne, on s'assure de transformer le support des données bruts d'irradiance initialement positives ou nuls à un support infini en théorie.

Enfin comme simple post-traitement, on appliquera la fonction inverse au pré-traitement en question pour pouvoir avoir stricto sensu la prédiction d'irradiance.

D'autres pré-traitements sont possibles comme il a été discuté auparavant 1.3.2 et pourraient être effectivement testés. On peut penser notamment à des méthodes de réduction de dimension comme une analyse par composante principale, ou bien factorielle, ou encore une transformée par ondelettes avant d'apprendre des modèles HMM.

2.4.2 Discussion sur les critères de performance

Energy Score

Nous avons décidé d'évaluer les modèles sur deux critères complémentaires : un critère de performance prédictive, et en appui un critère de vraisemblance. Le critère de performance prédictif discrimine les modèle sur les erreurs de prédictions commises selon une métrique donnée. Dans le cadre de prédictions probabilistes où nous sommes, nous utilisons l'Energy score(27). Cette métrique orientée négativement est propre (une prédiction exacte donnera le meilleur score). Cette métrique quantifie à la fois la précision, l'attache aux données par rapport à une série de test, et la dispersion, l'hétérogénéité des profils prédits :

$$ES = \sum_{j=1}^J \|z - \hat{z}_j\|_2 - \frac{1}{2} \sum_{i=1}^J \sum_{j=1}^J \|\hat{z}_i - \hat{z}_j\|_2$$

Pour une observation z , on génère un nombre J de simulations \hat{z} . On mesure à la fois l'écart entre l'observation et les simulations, mais aussi les simulations entre elles. Pour chaque observation, on obtient alors un energy score traduisant la performance du modèle utilisé. L'energy score peut donc être utilisé dans le choix des paramètres optimaux des modèles. Pour deux set de prédictions à précision égale, l'energy attribuera le meilleur score au modèle fournissant des scenarii les plus diversifiés.

L'avantage d'une telle métrique est sa souplesse d'application qui ne requiert aucune information supplémentaires contrairement au score de Brier ou le CRPS qui se basent sur les fonctions densités de distribution qui peut être difficile à connaître (27). L'Energy score permet donc de comparer un large panel de méthodes, à partir des informations issus d'un set de scénarii de prédiction.

Critères d'informations de vraisemblance, AIC et BIC

La performance des prédictions est un critère important, mais il n'est pas le seul. En effet, trouver un modèle vraisemblable au vu des données d'entraînements fournis avec si possible le moins de paramètres peut être recherché. Des limites en temps de calcul

et en complexité du modèle nous amène donc à privilégier des modèles avec le moins de paramètre. Or, l'estimation des paramètres détaillés plus tard, tant des coefficients des modèles SARIMA que les modèle HMM, reposent sur une estimation de vraisemblance. Par conséquent, des critères qui prennent en compte la complexité du modèle ou la taille d'apprentissage peuvent avoir un atout pour sélectionner un modèle de qualité tout en évitant des modèles trop complexes, sujets au sur-ajustement. On prendra donc comme critère à mettre en compétition avec l'Energy Score, un critère d'information d'Akaike (AIC) ou de Bayes (BIC)

$$AIC = 2k - 2\ln(\ell), BIC = k\ln(n) - 2\ln(\ell)$$

avec k : nombre de paramètre, ℓ :vraisemblance du modèle, n : taille des données d'apprentissage

Pour appliquer ces critères d'information dans la pratique, nous calculons la vraisemblance issue de l'optimisation des paramètres des modèles, pour un ensemble de modèles candidats. Il y aura presque toujours une perte d'information due à l'utilisation d'un modèle candidat pour représenter le "vrai modèle". De plus nous ne pouvons pas le choisir avec certitude, mais nous pouvons minimiser la perte d'information estimée. On sélectionnera, parmi les modèles candidats, le modèle qui minimise la perte d'information. La quantité $\exp(\frac{AIC_{min} - AIC_i}{2})$ est interprétée comme la vraisemblance relative aux données d'apprentissage du modèle i .

2.4.3 Méthodes de prédictions adaptées pour les deux modèles, et calcul de l'energy score

L'Energy Score a été calculé par une cross-validation "hold out" adaptée de manière dynamique sur une fenêtre glissante de 24h. La cross-validation "holdout" consiste à séparer les données en deux échantillons indépendants, d'apprentissage et d'évaluation (données "tests"). En effet : l'optimisation des paramètres s'est faite sur la série temporelle d'irradiante - avec les pré-traitements respectifs pour chaque famille de modèles - de taille de 3 ans (données de janvier 2016 à décembre 2018). Une fois les modèles appris par le biais des données d'apprentissage, les poids des paramètres ne sont plus touchés, et on procède au calculs des scores du modèles sur la base des données test.

La validation s'est ensuite basée sur l'année 2019, qui n'a pas été utilisée dans la partie d'apprentissage. Néanmoins les deux familles en question, HMM et SARIMA, ne sont pas adaptés à faire des prédictions long-terme. Cela n'a pas de sens de faire une prédition "long-terme" sur toute l'année. Pour cette raison, on préfère une prédition de taille fixée, ici 24h, avec un horizon "fuyant" sur toute la période des données de test. La donnée de test mise de côté va servir alors, à "mettre à jour" les informations pour les prédictions futurs, jour par jour.

Prédiction dynamique pour SARIMA

Pour le modèle SARIMA avec une périodicité s , les coefficients sont pondérés par les données d'apprentissages, et on applique la formule de régression en "mettant à jour" les valeurs X_t par les données de test observées aux instants précédents.

$$X_t = \varepsilon_t + \sum_{i=1}^p (\varphi_i X_{t-si} + \sum_{i=1}^q (\vartheta_i \varepsilon_{t-si}))$$

Les termes X_t seront ainsi mis à jour par les valeurs de la donnée test au fur et à mesure que l'on fasse glisser la fenêtre de 24h.

Prédiction dynamique HMM

De manière analogue pour les modèles HMM, les paramètres sont fittés par les données d'apprentissage, à savoir la matrice de transition des états cachés A , les moyennes d'émissions μ et les matrices de covariances Σ . Ensuite grâce aux données tests on "met à jour" l'information de l'état caché du jour précédent. Pour ce faire, on trouve l'état caché h le plus proche du profil d'irradiance observée au jour précédent y_{t-1} . On sélectionne l'état h^* parmi les états cachés s minimisant la norme entre l'observation y_{t-1} , et le vecteur de profil moyen $\mu^{(h)}$ issu des états cachés h .

$$h^* = \arg \min_h \|y_{t-1} - \mu^{(h)}\|$$

A partir de cette information mise à jour, on simule les trajectoires à partir de la matrice de transitions des états cachés. L'état caché le plus "proche" du profil observé au jour t est calculé de la manière suivante : parmi les états cachés on calcule la norme entre l'observation et le profil d'émission moyen. L'état caché qui minimise en ce sens l'écart est choisi comme état du jour précédent, et on pose X_t la distribution des états cachés avec le poids à 1 pour l'état caché calculé. Par hypothèse de la chaîne de Markov des états cachés calcule X_{t+1} qui est la distribution des probabilités au jour suivant que l'on souhaite, à l'aide de X_t et A :

$$X_{t+1} = A.X_t$$

On tire aléatoirement ensuite cette distribution X_{t+1} , autant de fois que l'on souhaite de scénario Chaque tirage nous donne un état caché. De cet état caché, on récupère les paramètres des émissions gaussiens associé à l'état et on tire une finalement trajectoire multivariée gaussienne, qui fera office de prédiction pour le jour $t + 1$

Sur la base de la série de test, on prédit un set de scénario, admettons 100 trajectoires par exemple, sur une fenêtre de 24h. On calcule l'energy score 2.4.2 entre les 100 trajectoires prédites, et la vraie observation, ce qui donne un score "journalier", et on fait glisser la fenêtre de prédiction, en remettant à jour les informations nécessaires pour la prédiction. Enfin sur 1 an on obtient 365 scores journaliers dans l'année, avec lesquels on peut déduire un score moyen dynamique.

La méthode hold-out est avantageuse en gain de temps car elle ne demande d'être lancée qu'une seule fois l'échantillonage. L'inconvénient est que pour des échantillons trop petits, les critères estimés peuvent être faussés si l'échantillon d'apprentissage est trop petit ou n'englobe pas des observations les plus génériques possibles, par exemple des informations qui n'ont pas été incorporé dans les données d'apprentissage seront mal prédits. Des méthodes de cross-validation plus fine comme les méthodes K-fold méritent d'être mises en places. Ces méthodes intègre la cross-validation holdout - en séparant aléatoirement les données en deux échantillons train-test, puis en répétant cette opération K fois : le score d'évaluation finale est la moyenne des K scores.

2.4.4 Optimisation des paramètres

On présente la routine de recherche de meilleur modèle dans des familles de modèles de prédictions selon les critères décrits ci-dessus. En l'occurrence les critères AIC-BIC ont été calculés à partir du score vraisemblance du modèle durant l'apprentissage.

Algorithm 1: Routine Performance

Require: $\text{data}_{\text{train}}$: échantillon d'apprentissage utilisé pour faire correspondre un modèle,
 \mathcal{M} : ensemble des candidats familles de modèles-candidat m , ici les candidats de la familles des modèles SARIMA, et ceux de la famille des modèles HMM à émission gaussienne.
 N_{random} : nombre d'initialisation de paramètres pour l'étape d'optimisation de la vraisemblance des paramètres des modèles candidats
 $\text{parameters}_{\text{set}}$: ensemble des paramètres Ω du modèle théorique m ,
 n_{simu} : nombre de séries simulées à partir du modèle théorique $\mathcal{M} = 100$ scenarii
 $\{p^*\}_{1..K}$: ensemble des K paramètres exacts
Découper les données en un set d'apprentissage $\text{data}_{\text{train}}$ et de test $\text{data}_{\text{test}}$

for m in \mathcal{M} **do**

- Pour chaque modèle-candidat :
- Optimisation($m, \text{data}_{\text{train}}$) : on estime les paramètres optimaux maximisant la vraisemblance
- for** $i = 1$ to N_{random} **do**

 - Pour le même modèle, on lance N_{random} optimisations avec des initialisations aléatoires des paramètres
 - On optimise les coefficients du modèle par maximum de vraisemblance \mathcal{L}
 - Générer n_{simu} prédictions := predictions_set
 - Calcul Energy_Score($\text{data}_{\text{test}}, \text{predictions_set}$)
 - Calcul AIC,BIC

- end for**
- Calcul d'une erreur standarde des critères ES et AIC-BIC moyens

end for

Pour chaque modèle m , on associe un couple de critère (ES,AIC) ou (ES,BIC) avec une erreur standarde

L'optimisation est basée pour les deux familles sur des algorithmes de vraisemblance. On lance plusieurs optimisations de paramètres avec des initialisations aléatoires lorsqu'on apprend sur un modèle, HMM ou SARIMA , pour augmenter les chances de trouver un l'optimum global. On prend simplement le modèle qui a convergé avec le meilleur Energy Score.

Concernant les modèles SARIMA, l'optimisation par maximum de vraisemblance s'est faite par la méthode de direction conjuguée Powell qui assure de trouver un minimum local.

Concernant l'algorithme de Baum-Welch pour les modèles HMM, on l'exécute plusieurs fois avec différentes estimations initiales des paramètres $\lambda = (\mu, \Sigma, \pi)$. On pourrait par exemple établir des initialisations plus fine, par un algorithme de clustering de k proches voisins,(avec k fixé au nombre d'états cachés).

2.4.5 Une étude sur la taille optimale d'entraînement

Une étude sur la taille minimale a été aussi entamée mais n'a été adaptée que dans le cas des modèles SARIMA. Elle n'a été testée que pour la recherche de meilleurs modèles sur des données artificielles. L'idée était d'inférer une table théorique de performance pour un modèle SARIMA donné. En générant des données artificielles issues d'un modèle SARIMA que l'on se donne, on calcule pour différentes taille d'apprentissage, les erreurs d'estimations des coefficients du modèle théorique. Cette table serait donc un indicateur théorique de performance, théorique au sens où les erreurs indiquées sont celles dans le cas

où les données ont une structure parfaitement modélisable par un processus SARIMA, ce qui n'est en réalité jamais le cas. Le tâche fastidieuse et le temps de calcul était conséquent pour fournir une table fournie, que nous avons décidé de laisser en suspens cette étude.

Algorithm 2: Routine Taille Minimale

Require: $data_{train}$: échantillon d'apprentissage utilisé pour faire correspondre un modèle,
 $TrainSize_{set}$: ensemble d'intervalles d'observations, de tailles croissantes
 $parameters_{set}$: ensemble des paramètres Ω du modèle ARIMA théorique \mathcal{M} ,
 n_{simu} : nombre de séries simulées à partir du modèle théorique \mathcal{M}
 $\{p^*\}_{1..K}$: ensemble des K paramètres exacts

Ensure: df_{set} : contient K dataframes, un pour chaque paramètres de $parameters_{set}$

```

for  $i = 1 .. n_{simu}$  do
    A partir du modèle  $\mathcal{M}$ , générer  $TS_i$  une série temporelle
    for  $Obs\_interval_j$  in  $TrainSize_{set}$  do
        Sélectionner un intervalle d'observation  $Obs\_interval_j$  parmi  $TrainSize_{set}$ 
         $TS_{i,j} := TS_i[1..size_j]$  : on sous-échantillonne la  $i^{eme}$  série simulée avec le  $j^{eme}$  intervalle d'observation  $Obs\_interval_j$ 
        A partir de  $TS_{i,j}$  , on détermine  $\{\hat{p}\}_{1..K}$  l'ensemble des estimations des paramètres du modèle  $\mathcal{M}$ 
        for  $\hat{p}_k$  in  $\{\hat{p}\}$  do
            Calculer l'erreur relative  $e_k^{(i,j)}$  entre l'estimation  $\hat{p}_k$  et le paramètre exact  $p_k^*$  ,
            issu de la  $i^{eme}$  série et du  $j^{eme}$  intervalle d'observation
            Stocker  $e_k^{(i,j)}$  dans le  $k^{eme}$  dataframe, à la ligne  $i$  et colonne  $j$ 
        end for
    end for
end for

```

3

Résultats des modèles de prédition sur des profils d'irradiance globale

On va donc appliquer les routines présentées au chapitre précédent sur des données d'irradiance globale qui sont décrites ci-dessous :

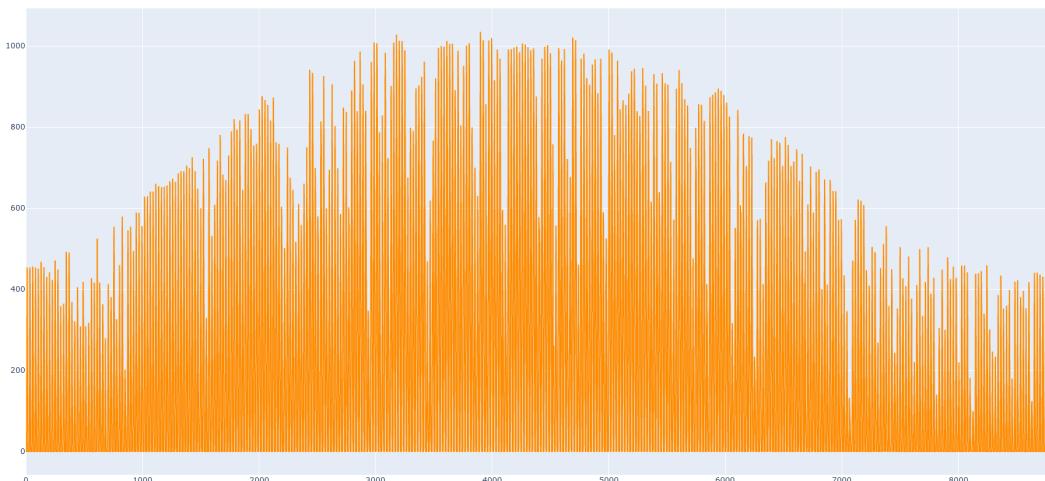


FIGURE 3.1 – Irradiance globale à Odeillo -France sur l'année 2019. En abscisse : index des mesures (pas de temps horaire), en ordonnée : irradiance globale en [W/m²]

La donnée sera pré-traitée par les méthodes énoncées puis on appliquera les modèles sur les données transformées comme décrit dans la section 2.4.1. Ce pré-traitement vise à normaliser la donnée afin d'adapter sa structure aux modèles implémentés. Mais avant de discuter des performances des modèles sur de la prédition d'irradiance, nous allons pour chaque famille de modèles, valider la routine sur des données artificielles issues de modèles de référence que l'on construit : des processus SARIMA et des modèles HMM, tous générés en fixant arbitrairement les hyper-paramètres . Ainsi on présentera en particulier la validation sur données artificielles issus d'un processus MA(1) et d'un modèle HMM à émission gaussienne multivariée de dimension 24, avec 3 états cachés.

3.1 Résultats des modèles SARIMA

3.1.1 Validation sur des données artificielles

On valide la routine sur des données artificielles générées. Cette approche permet de valider rigoureusement la recherche de meilleurs modèle, dans le cas idéal où les données sont totalement modélisable par les processus SARIMA.

On a testé la routine pour différents modèles, et on présente ici les résultats sur des données générées à partir d'un processus aléatoire MA d'ordre 1, avec pour paramètres arbitraire : $\theta_1 = 0.6$ et de variance $\sigma^2 = 1.44$, qui s'exprime :

$$Y_t = \theta_1 \cdot \varepsilon_{t-1} + \varepsilon_t$$

, où $\varepsilon_t, \varepsilon_{t-1}$ des variables aléatoires de loi $\mathcal{N}(0, \sigma^2)$.

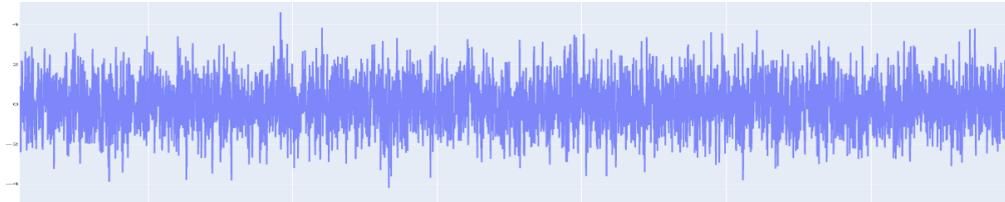


FIGURE 3.2 – Série temporelle de test d'un processus stationnaire MA(1) (taille 10 000 points)

On applique la routine d'optimisation détaillée dans la partie précédente et on essaie de retrouver les coefficients du modèle théorique avec la donnée simulée. On teste toutes les combinaisons d'ordre AR et MA jusqu'à l'ordre 2, avec une différenciation d égale à 0 car les processus MA sont par définitions stationnaires. On teste alors les modèles $((0,0,0), (0,0,1), (1,0,0), \dots, (2,0,2))$ soit 9 modèles possibles en tout. Par souci de lisibilité et d'efficacité, on omet de tester sur des modèles candidats SARIMA avec des coefficients AR et/ou MA saisoenniers.

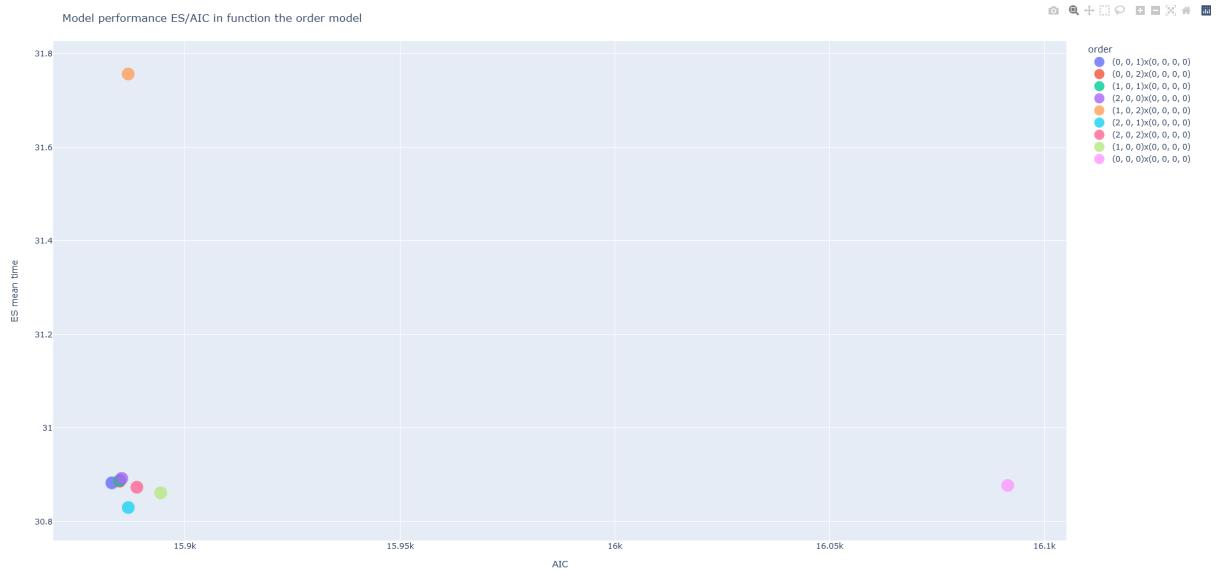


FIGURE 3.3 – Grille de Performance Energy Score - AIC par la routine perfomance pour des données artificielles issues d'un processus MA(1)

Chaque modèle a un couple de score ES-AIC moyen calculé. On voit logiquement se dessiner une distribution de scores, le modèle bruit blanc $(0,0,0)$ ayant l'AIC le plus médiocre. Comme les deux métriques sont orientés négativement, le meilleur modèle estimé est celui positionné dans le coin en bas - gauche du repère.

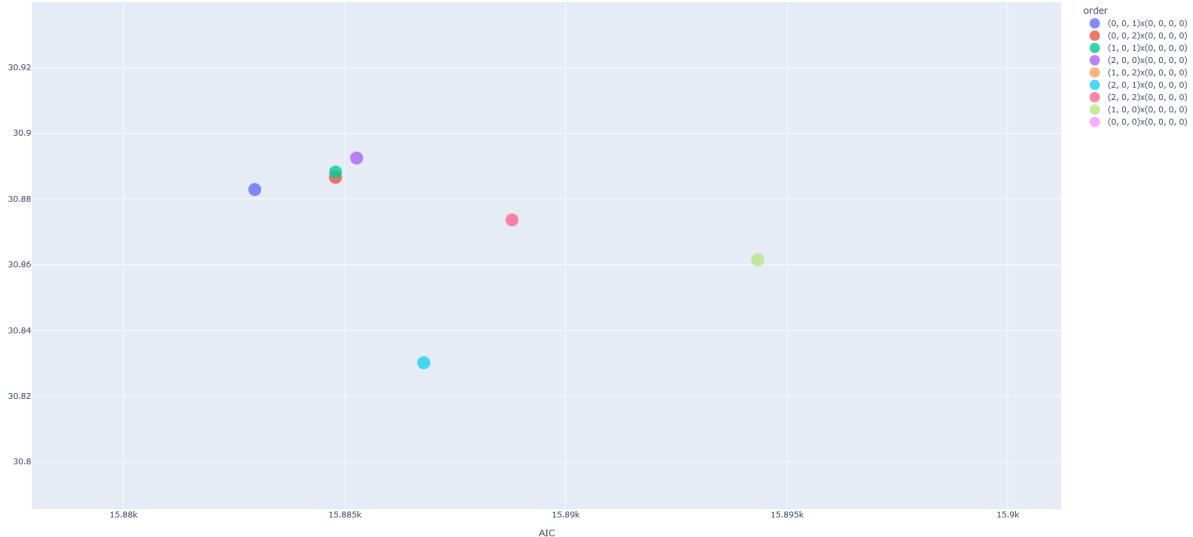


FIGURE 3.4 – Grille de Performance Energy Score - AIC par la routine perfomance pour des données artificielles issues d'un processus MA(1), détail

Finalement en grossissant le graphe, le modèle candidat MA(1) (en bleu dans la légende) se présente comme un des meilleurs modèles, et a bien convergé vers les paramètres théoriques comme décrit dans la figure 3.5

SARIMAX Results						
Dep. Variable:	y	No. Observations:	5000			
Model:	SARIMAX(0, 0, 1)	Log Likelihood	-7966.640			
	AIC	15937.280				
	BIC	15950.315				
	HQIC	15941.849				
	coef	std err	z	P> z	[0.025	0.975]
ma.L1	-0.5860	0.012	-50.698	0.000	-0.609	-0.563
sigma2	1.4172	0.028	49.991	0.000	1.362	1.473
Ljung-Box (L1) (Q):	0.42	Jarque-Bera (JB):		0.17		
Prob(Q):	0.52	Prob(JB):		0.92		
Heteroskedasticity (H):	1.02	Skew:		-0.01		
Prob(H) (two-sided):	0.70	Kurtosis:		3.00		

FIGURE 3.5 – Valeurs des coefficients estimés par le modèle MA(1)

Sur la figure 3.5, les coefficients du modèle estimé MA(1) ont bien convergé vers les paramètres théoriques, avec une erreur sur le paramètre θ_1 de 0.014, et une erreur sur la variance du process de 0.009.

Le modèle MA(1) estimé présente également un des meilleurs energy score moyen sans être dans l'absolu le meilleur. L'energy score n'a pas de sens physique, ce qui importe est d'avantage, ce sont les écart de scores entre les modèles. L'energy score permet d'écartier avec forte présomption des modèles avec de grands scores comme les modèles (0,0,0), (1,0,2) et (1,0,0). On ne peut pas conclure sur des scores très proche hormis de dire que les modèles ont des performances de prédictions similaires. Cela a dû sens car les modèles candidats encore discutables, se présentent comme tous des processus englobant MA(1). A titre d'exemple très simple en vérifiant les paramètres après l'optimisation, le modèle (0,0,2) (en rouge) estimé a un coefficient de moyenne mobile θ_1 très proche du modèle MA(1) mais avec un coefficient θ_2 nul : les deux modèles sont donc équivalents.

Sur la base du score de prédiction, on est capable d'avoir une première sélection de modèles. Les modèles encore discutables sont alors "équivalents" en terme de performance prédictive. C'est là que des critères d'informations deviennent intéressants, car ils permettent de sélectionner le modèle le plus vraisemblable aux données, en pénalisant notamment les modèles trop complexes.

Par conséquent, le modèle MA(1) s'avère être le meilleur modèle, ce qui valide la routine sur des données simulées issues de processus SARIMA.

3.1.2 Evaluation de la performance des modèles SARIMA sur les données d'irradiances

En premier lieu on peut stationnariser la donnée temporelle d'irradiance en la divisant par le flux "top of atmosphere" comme mentionné dans la littérature(12). La série stationnarisée est plus courte que la série test car on supprime les périodes de nuits. Ainsi, dans la suite du rapport on parlera de "jours d'ensoleillement" ou d'heures d'ensoleillement pour faire référence à une journée de cette série stationnarisée dont on a supprimé les nuits.

Les modèles SARIMA suivants seront donc des modèles de prédiction sur la base de l'apprentissage de la série stationnarisée avec uniquement les heures d'ensoleillements.

Comme mentionnée précédemment, on rappelle que l'on apprend sur 3 ans de données, et on mène une cross-validation holdout sur 1 an de l'année 2019.

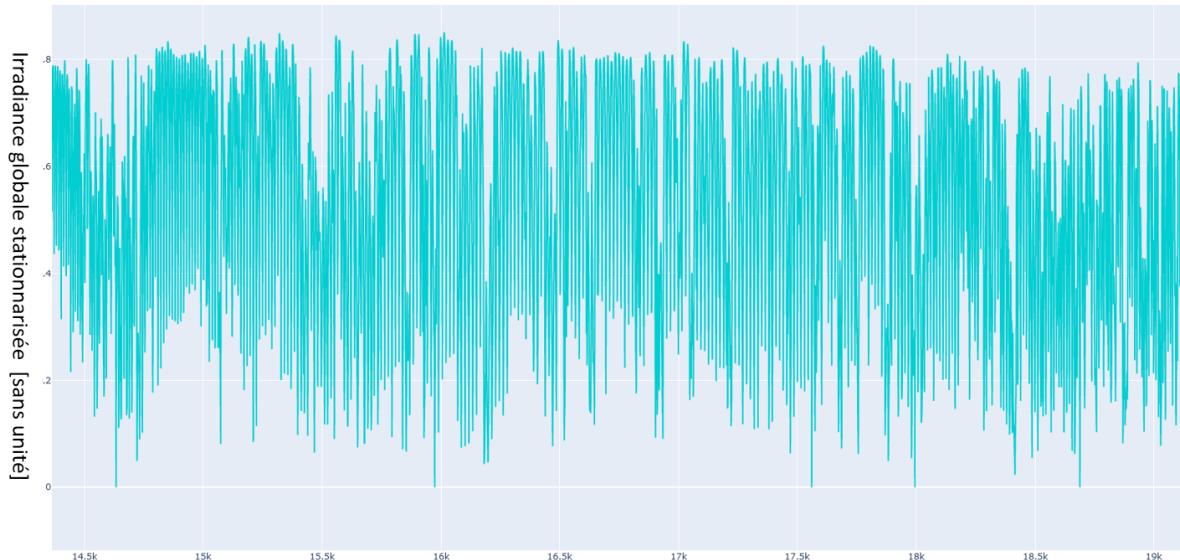


FIGURE 3.6 – Série des heures ensoleillés d'irradiance stationnarisée, à Odeillo sur l'année 2019 En abscisse : l'indice de la valeur d'irradiance stationnarisée. En ordonnée : l'irradiance divisée par le TOA

Comme pour les données d'irradiance artificielle, on cherche parmi plusieurs modèles candidats, le processus SARIMA qui pourrait apprendre au mieux la structure des données présenter. On a mené une recherche de meilleurs modèles, en explorant toutes les combinaisons dans une grille d'ordre maximales ($ar=1, ma=1$, ar saisonnier = 2, ma saisonnier = 2). Comme on travaille sur les heures d'ensoleillement des données d'irradiance, on fixe la périodicité s des modèles SARIMA à 24h d'ensoleillement. Sur toutes les combinaisons possibles, certains modèles n'ont pas convergé, on présente les 14 meilleurs modèles appris sur les 3 ans de données d'irradiance.

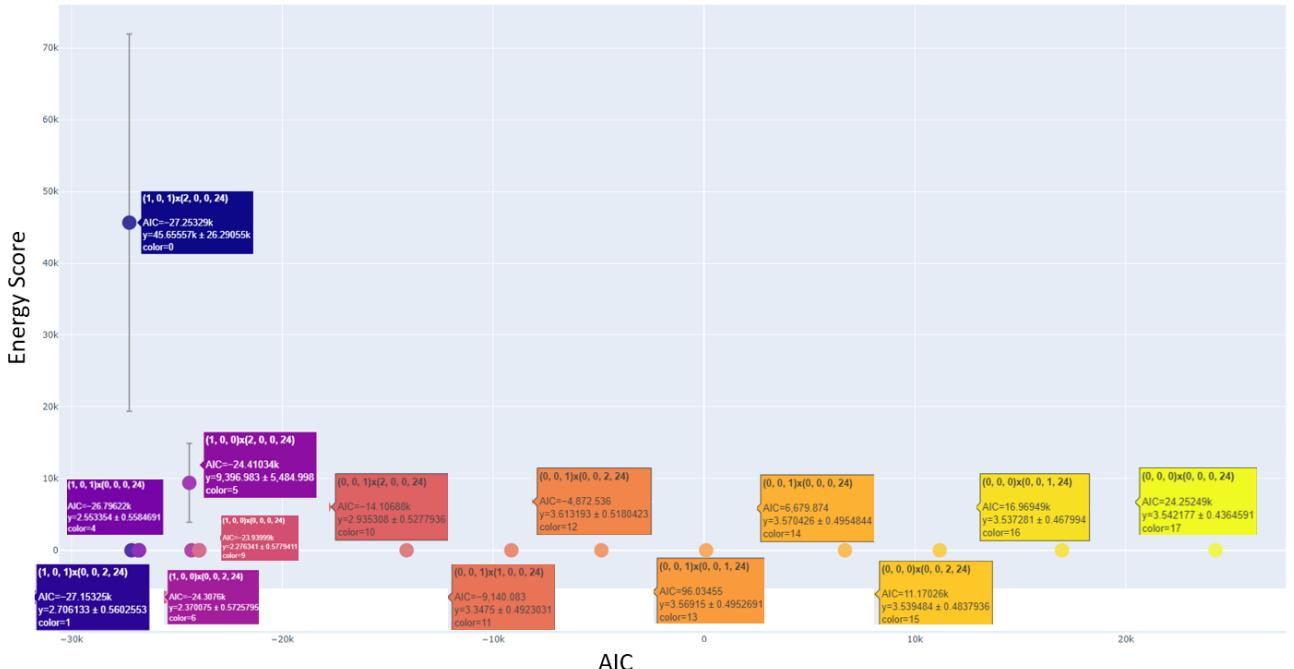


FIGURE 3.7 – Grille de Performance Energy Score - AIC moyens par la routine performance, avec erreur standard, sur les modèles SARIMA d’irradiance stationnarisée. Apprentissage sur les données d’irradiance 2016-2018. Evaluation sur les données 2019 sur les données d’apprentissages

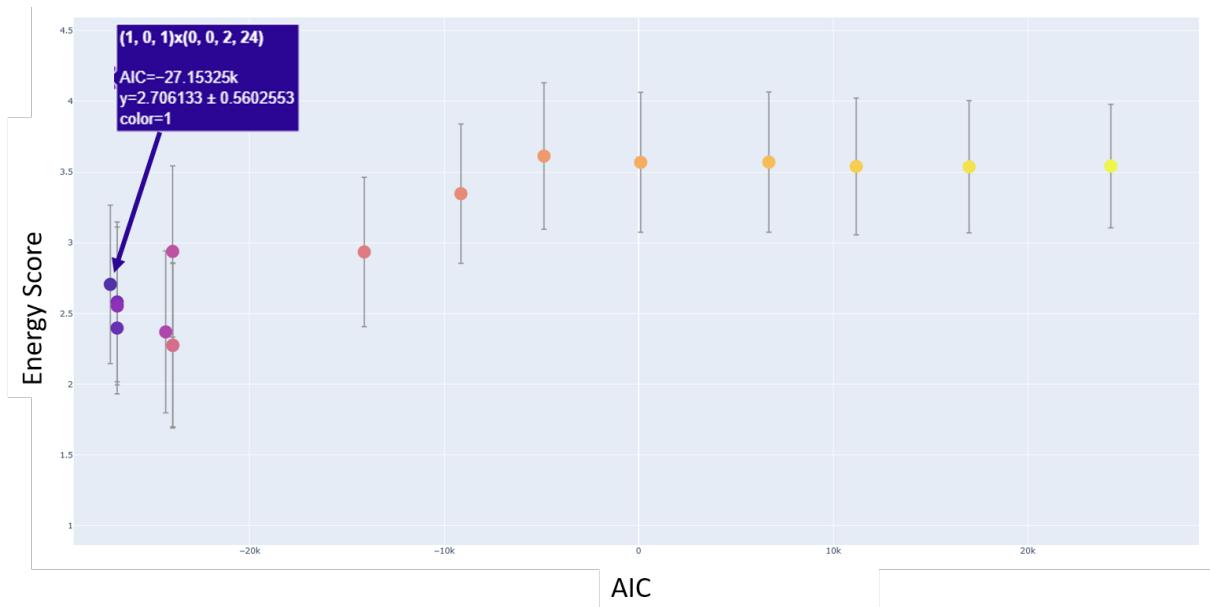


FIGURE 3.8 – Grossissement de la figure 3.7

Pour chaque modèle, la valeur d’Energy Score correspond à la moyenne des scores calculés durant la phase de cross-validation sur les données de l’année 2019 . Les barres d’erreurs standardes des energy scores en gris sont calculées à partir des prédictions dynamiques simulées chaque jour dynamiquement, ici 100 prédictions sont simulées à chaque fenêtre glissante de prédiction.

Sur ce détail illustré sur la Figure 3.8, le modèle SARIMA (1,0,1)x(0,0,2) serait le modèle le plus vraisemblable, et présente un energy score parmi les plus faibles. On peut donc conclure avec conviction que ce modèle intègre le mieux la structure des données d’irradiance parmi les modèles testés.

A partir de ce modèle, on génère une prédition dynamique simulée. En turquoise figure les dix premiers "jours d'ensoleillement" issus de la série de test stationnarisée des heures d'ensoleillement. On a fait apparaître en pointillé le découpage de ces dix "jours" dans le graphe.

On rappelle que les prédictions ont été effectuées comme il a été détaillé dans la section 2.4.3, à savoir des prédictions dynamiques sur chaque "jour" d'ensoleillement.

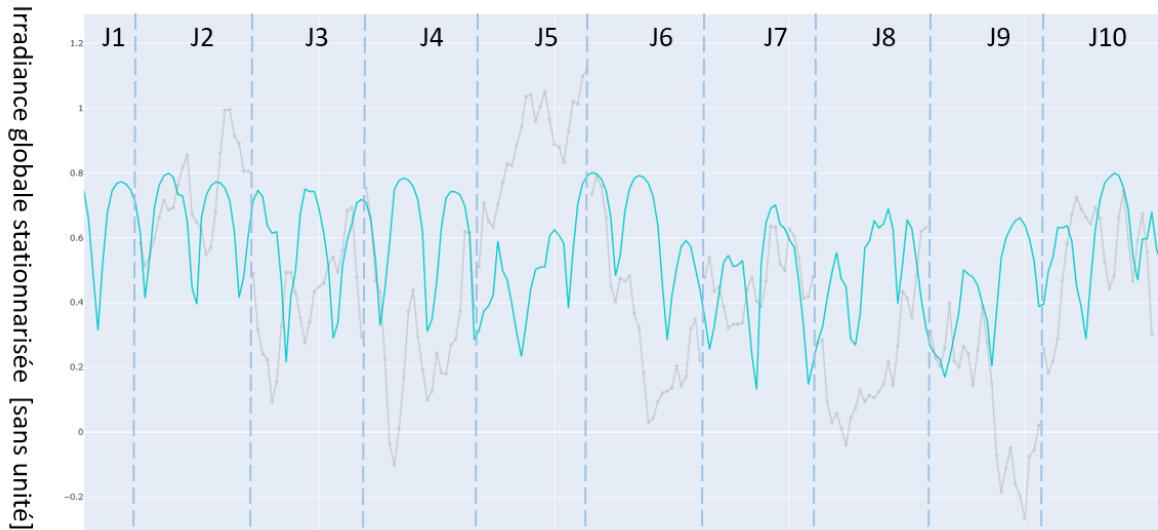


FIGURE 3.9 – Une prédition dynamique (en gris) sur les 10 premiers jours de la série stationnarisée, issue du modèle SARIMA($p=1,d=0,q=1)x(ps=0,ds=0,qs=2,s=24)$ (en turquoise).

Même si les prédictions proviennent du meilleur modèle selon l'étude de recherche de meilleurs modèles, on peut conclure que les modèles SARIMA testées ici n'ont pas capté la nature périodique des données. L'avantage certain de ces modèles employés de cette manière avec des prédictions dynamiques, et que l'on peut bénéficier de diverses prédictions de profils variés, ce qui reste tout de même intéressant dans un contexte de contrôle optimal stochastique. Par ailleurs les modèles SARIMA restent des méthodes de régressions adaptées à des séries évoluant dans le temps. Du fait de leur faible complexité, on s'attendait a priori à avoir des performances somme toutes passables. Suite à cela, un premier élément de réponse que nous pouvons apporter est que les modèles SARIMA ne semblent pas être parfaitement adaptées pour ce type de données, dans une approche où l'on cherche à prédire des trajectoires sur plusieurs jours dans de la co-optimisation de réseau. Il serait intéressant de tester ces modèles très simples pour des pas de temps plus court, de l'ordre de la minute par exemple. En effet, la nature des données est importante, mais on ne doit pas pour autant négliger l'importance de la taille de fenêtre, qui peut donner une dynamique différente si l'on regarde à l'échelle de la minute, de l'heure ou de la journée.

3.2 Résultats des modèles HMM

3.2.1 Validation sur les données artificielles

Comme dans le cas des modèles SARIMA, on valide la routine sur des données artificielles que l'on contrôle parfaitement. Divers exemples ont été mené pour différents nombres d'états cachés, l'exemple d'illustration portera sur le modèle HMM suivant

- $n = 3$ états cachés
- la dimension $d = 24$ des observations émises par les états cachés
- une matrice de transition $n \times n$.
- n vecteurs d'émission moyen $\{\mu_i\}_{i=1..n}$
- une matrice de covariance $d \times d$ $\{\Sigma_i\}_{i=1..n}$ que l'on considère diagonale, la covariance entre les instants d'émissions et nulles.

Remarques sur la structure de la matrice de transition

La matrice transition a été construite arbitrairement en mettant des poids aléatoire et en respectant la propriété stochastique de la matrice (les coefficients représentent la probabilité de transition et la somme de chaque ligne de la matrice vaut 1). On veut une chaîne de Markov la plus générale possible de sorte que tous les états soient visités indépendamment de l'initialisation. On a alors pris le soin d'avoir aucune transition à 0, cela revient à poser comme condition sur les probabilités de transitions des états cachés $A_{ij} = p(X_{t+1} = j | X_t = i) > 0$. Cela implique naturellement un type de matrice particulier, avec l'existence d'une distribution stationnaire des états.

En effet, on veut tester un modèle HMM artificiel la plus générale, avec une matrice de transition capable de générer toutes les combinaisons possibles de transitions, à savoir également la transition de rester sur le même. Cela implique que notre matrice de transition est apériodique et irréductible. Une matrice stochastique est apériodique lorsqu'il y a existence d'une probabilité non nulle de rester sur un même état à la transition suivante : $\exists i \in \{1, \dots, m\} : A_{i,i} > 0$. Une chaîne de Markov est irréductible lorsque tous les états communiquent entre eux, il existe un chemin où l'on peut visiter tous les états, indépendamment de l'initialisation sur la chaîne de Markov. Tous les états de la chaîne forment une seule classe d'équivalence.

Or, une matrice irréductible et finie (nombre d'états cachés discret) implique que la chaîne soit récurrente positive,, c'dà qu'il existe un temps de retour finie à l'état où l'on se trouve, et ce indépendamment de l'état de la distribution initiale des états cachés.

Finalement, ces trois propriétés impliquent l'existence et l'unicité d'une distribution des probabilités de transitions stationnaires, qui n'évolue plus à chaque transitions. Cette distribution peut substituer l'étape de transition en tirant directement une probabilité issue de ce vecteur.

De manière plus large, il peut être intéressant d'étudier la performance de convergence de l'optimisation des paramètres du modèle HMM, pour des matrices à diverses propriétés (avec plusieurs familles, de la périodicité etc...). On suppose qu'une matrice générale confère un temps de convergence le plus petit, et que de propriétés comme le poids des transitions, ou encore réductibilité de la chaîne (matrice avec plusieurs classes d'équivalences) peut augmenter le temps de calcul avant d'avoir converger vers les paramètres (μ_m, Σ_m) , voire ne jamais pouvoir les atteindre dans des cas où on ne visite jamais certains états cachés.

Remarques sur les émissions issues des états cachés

En supposant une trajectoire émise y_m par un état caché m , cela signifie qu'à tout instant $t = 1, 2, \dots, d$, les valeurs $y_m(t)$ suivent une loi gaussienne $\mathcal{N}(\mu_m(t), \Sigma_m(t, t))$, ce qui avec la structure diagonale de la matrice, impliquent l'indépendance des valeurs entre elles. Ainsi chaque valeur prédictive d'une trajectoire se caractérise totalement par les paramètres $\mu_m(t), \Sigma_m(t, t)$.

On affiche ci-dessous les trois émissions moyennes $\mu_m(t)$ pour les trois états cachés du mo-

dèle HMM artificiel, avec respectivement en gris l'intervalle de l'erreur standard $\sqrt{\Sigma_m(t, t)}$ centré en la moyenne $\mu_m(t)$

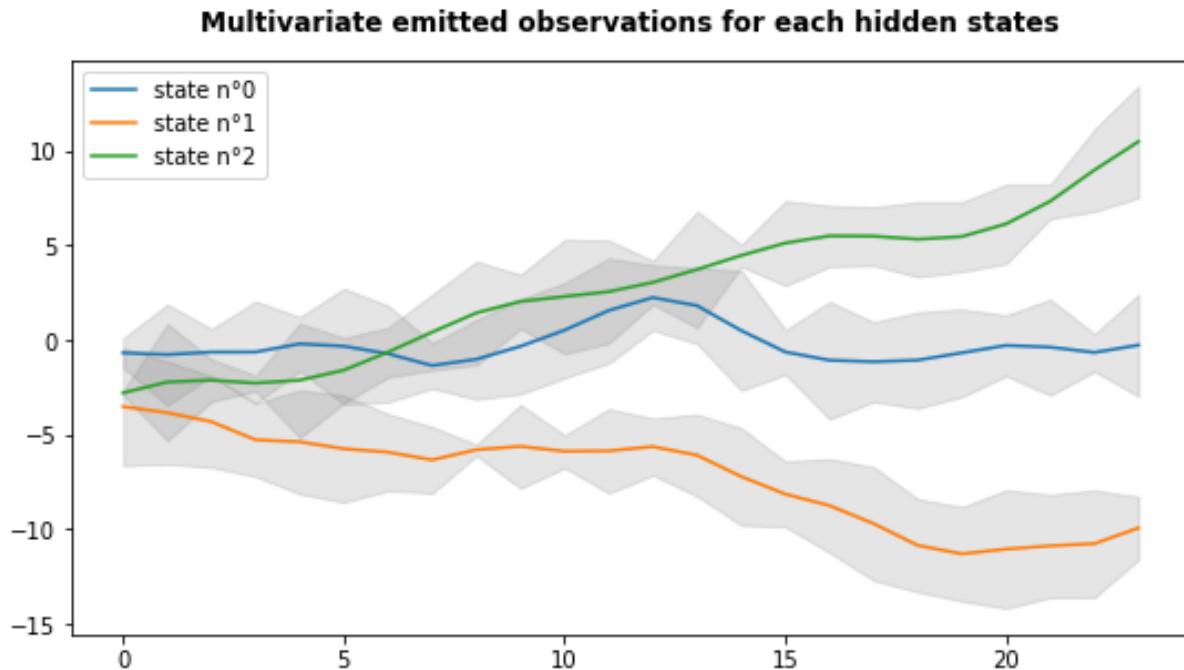


FIGURE 3.10 – Paramètres (μ, Σ) des trois états cachés du modèle HMM, sur un horizon de 24 . En abscisse : évolution temporelle des paramètres (μ, Σ)

A partir de ces paramètres on peut donc générer des trajectoires qui feront office de données ainsi simulées. Une fois les données générées, on essaie de retrouver les paramètres théoriques du modèle HMM par la base de ces données à l'aide de l'algorithme Baum-Welch. On a générée une base d'apprentissage de taille 24000 soit 10000 profils d'apprentissage issus des paramètres théoriques.

Prédiction dynamique

En bleu est représentée une partie des données d'apprentissage sur laquelle on fait apprendre notre modèle pour optimiser les paramètres. En orange figure la trajectoire de test qui n'a pas servi à l'apprentissage et sur laquelle nous allons faire des prédictions dynamiques sur chaque cycle de taille des profils émis $d = 24$. Les prédictions dynamiques en vert sont donc calculées avec comme information a priori l'état caché du jour précédent, en initialisant avec le premier cycle de la trajectoire de test, c'est pour cette raison qu'il n'y pas de prédiction pour le premier cycle de test.

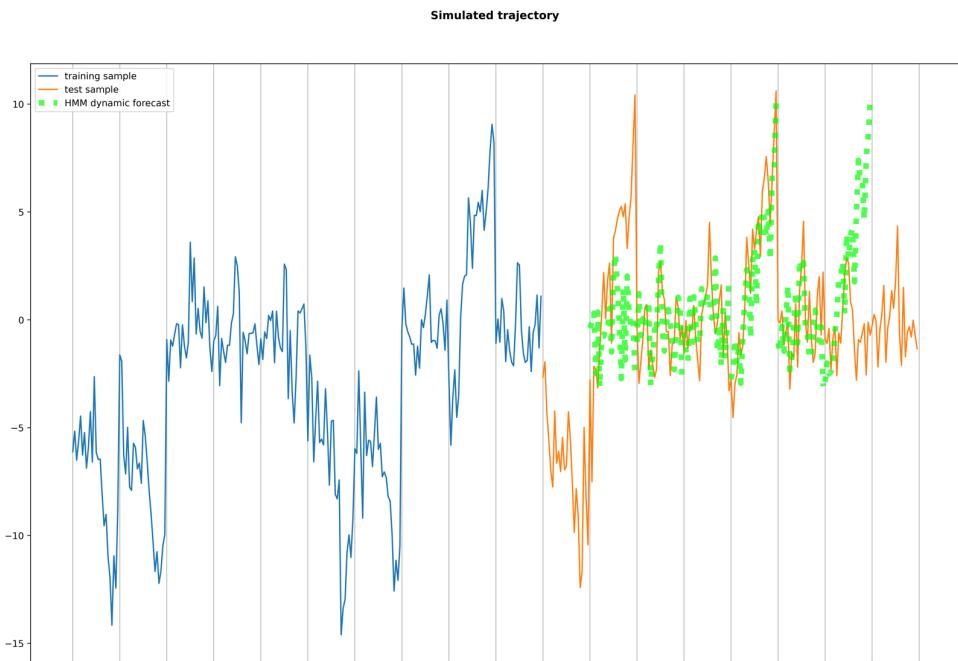


FIGURE 3.11 – Prédiction dynamiques sur plusieurs cycles de 24 points

On génère ensuite plusieurs modèles HMM candidats d'ordre d'états cachés différents, allant de 1 à 10. A l'aide des prédictions dynamiques et la vraisemblance des modèles obtenues dans l'optimisation, on peut déduire les scores ES et AIC/BIC moyens pour chaque modèles, de manière similaire à la validation sur données théoriques pour les modèles SARIMA :

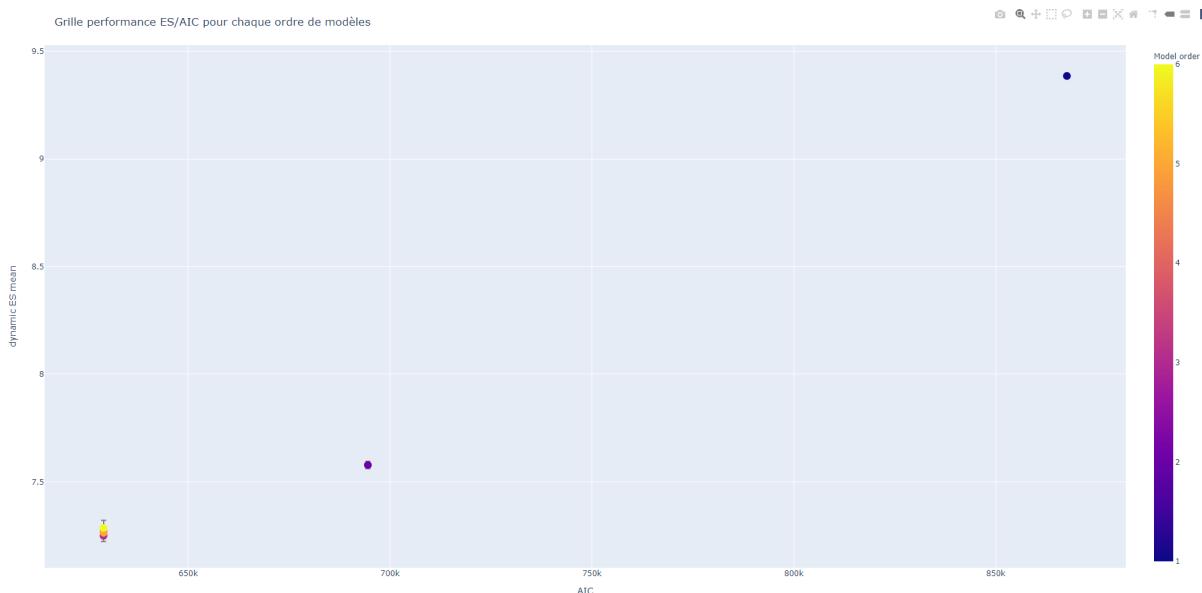


FIGURE 3.12 – Grille de Performance Energy Score - AIC moyens par la routine performance, avec erreur standard. La barre de couleur en légende représente l'ordre des modèles de 1 à 10

Finalement, on retrouve bien le modèle d'ordre 3 comme meilleurs modèles selon les critères d'ES et AIC.

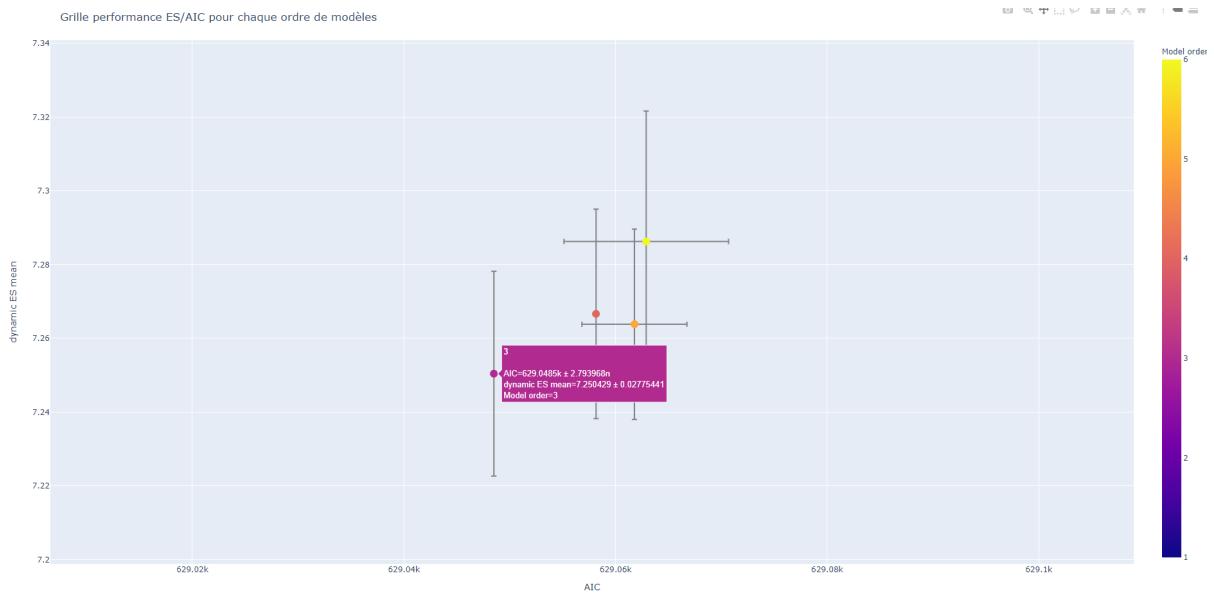


FIGURE 3.13 – Grille de Performance Energy Score - AIC moyens par la routine performance, avec erreur standard, détail des meilleurs modèles

Si en bleu on suppose la série temporelle d'une variable du système à piloter, on peut fournir plusieurs scénarios prédictifs qui serviront au contrôle optimal par MPC stochastique pour trouver la décision optimale minimisant la fonction coût que l'on s'est donnée au préalable.



FIGURE 3.14 – Quelques trajectoires de prédictions générées par le modèle HMM d'ordre 3

3.2.2 Evaluation de la performance des modèles HMM à émissions gaussiennes multivariées sur les données d'irradiance

Une fois après avoir testé la routine et l'optimisation des modèles HMM sur des données artificielles propres et parfaitement contrôlées, nous allons à présent étudier la performance de telles modèles sur les données d'irradiance présentées auparavant. Les modèles sont appris sur les données d'irradiance normalisées comme il a été discuté précédemment.

Ici, on recherche le meilleur ordre de modèles HMM allant de 1 à 10. Pour ce faire, pour chaque ordre de modèles, on lance 20 optimisations par Baum-Welch à des initialisations de paramètre différentes. Pour chaque pas de temps et pour chaque modèle à optimiser, on prédit 100 scénarios tous les 24 pas de temps horaires sur le profil d'irradiance de test 2.2. Ces prédictions dynamiques serviront au calcul de l'energy score de chaque fenêtre de prédition. A chaque décalage de la fenêtre de prédition, on met à jour l'information sur l'état caché de la journée précédente pour prédire les profils du jour actuel et calculer l'energy score associé.

On remarque qu'il y a un optimum d'ordre qui assure un energy score et un AIC. En effet les modèles avec de petits ordres semblent ne pas être adapté, comme le modèle à l'ordre 1 à 4. Les scores évoluent ainsi de manière croissantes au fur et à mesure que l'ordre du modèle augmente. On peut voir qu'à partir d'un certain nombre d'états cachés, approximativement au bout de 5 à 6 états, le "gain" des modèles HMM aux ordres supérieurs n'est plus intéressant. On observe donc graphiquement un optimum entre performance prédictive et complexité du modèle. Sur la figure suivante 3.16, les meilleurs modèles selon l'étude menée sont les ordre 7 à 10 avec des scores assez proches. En terme de vraisemblance, le modèle à 7 états présente un meilleur AIC que les ordres plus élevés. Ainsi les ordres plus élevés sont pénalisés par l'AIC, et n'ont pas de meilleures erreurs prédictions, ce qui nous laisse penser qu'un modèle d'ordre relativement élevé comme des ordres 7 ou 8 seraient les modèles à choisir dans un contexte de prédition d'irradiance.

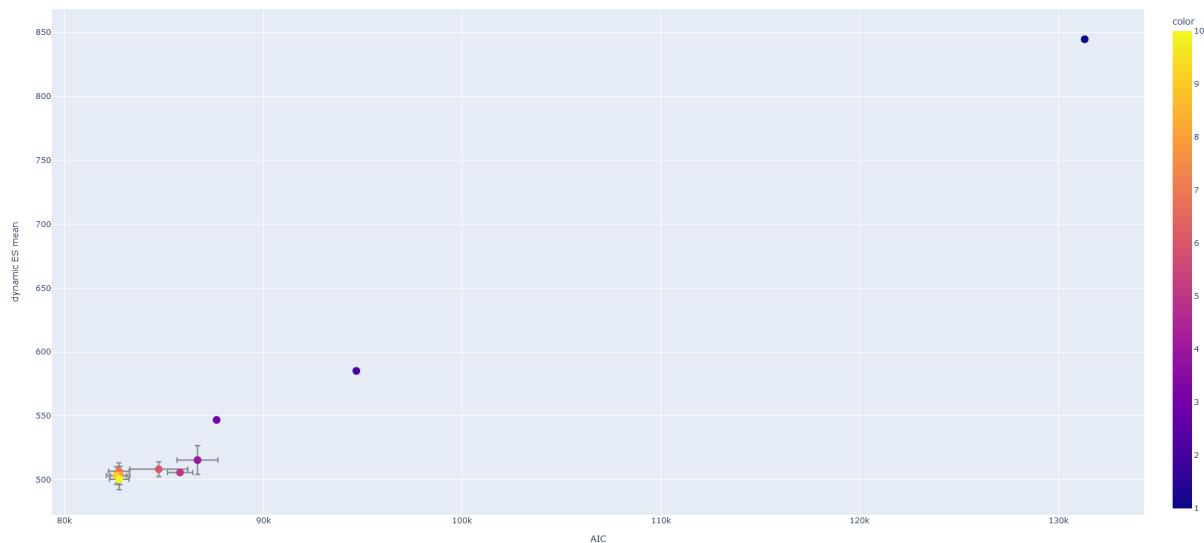


FIGURE 3.15 – Grille de Performance Energy Score - AIC moyens par la routine performance, avec erreur standard

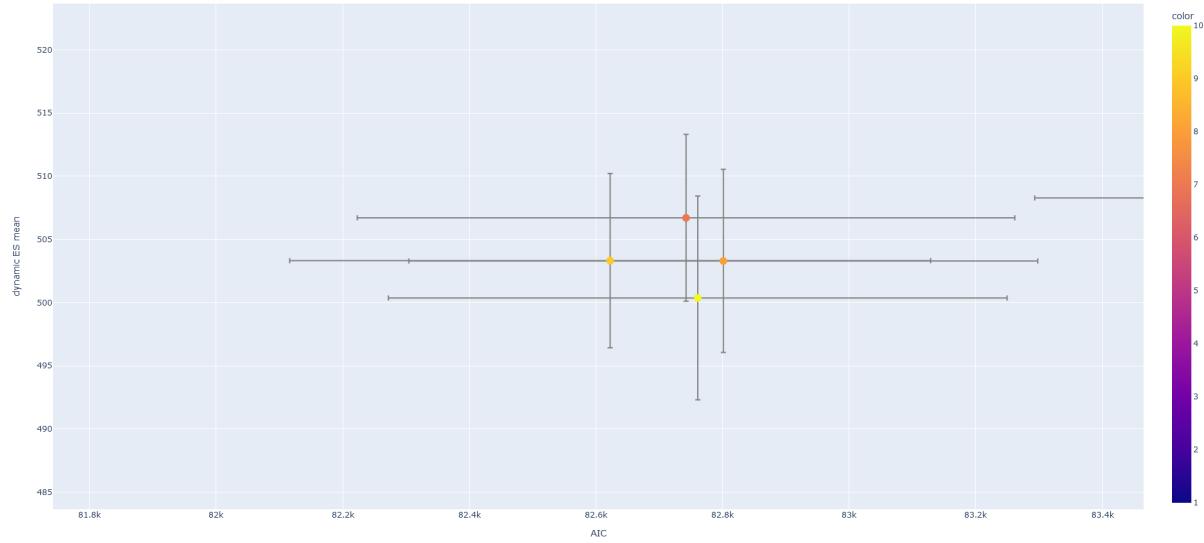
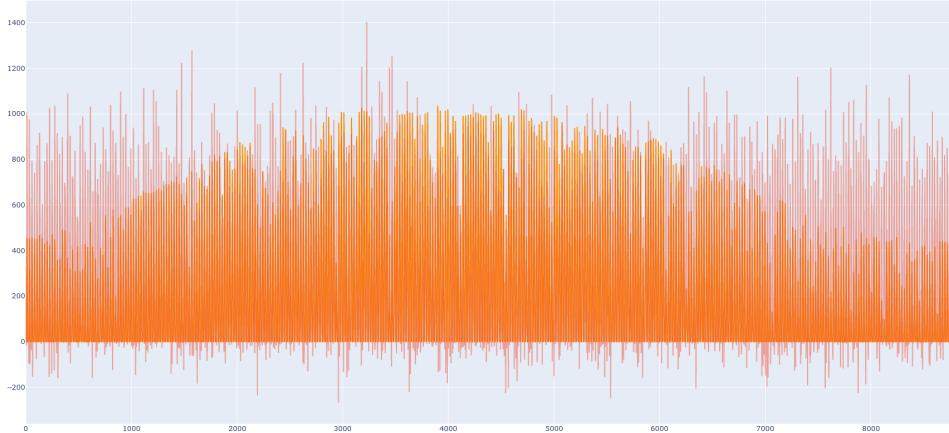
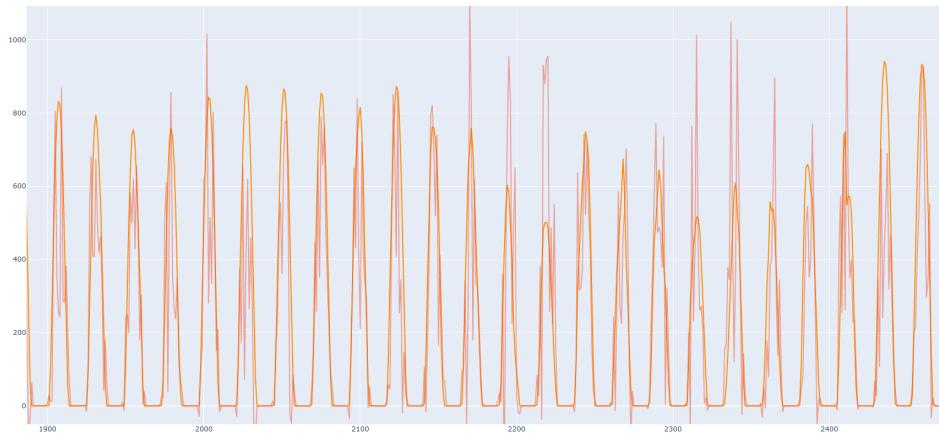


FIGURE 3.16 – Meilleurs modèles à l'issue de la routine

On regarde à présent la structure de prédiction qui découle des différents modèles. Les prédictions des modèle sont au préalable retransformés pour obtenir des prédictions cohérentes afin de les comparer avec les données d'irradiances brutes. Sur la bases des scores prédictif et de vraisemblance, on compare un des "pires" modèles (ordre 1), avec un modèle "moyen" (ordre 3), et le "meilleur" modèle (ordre 7).



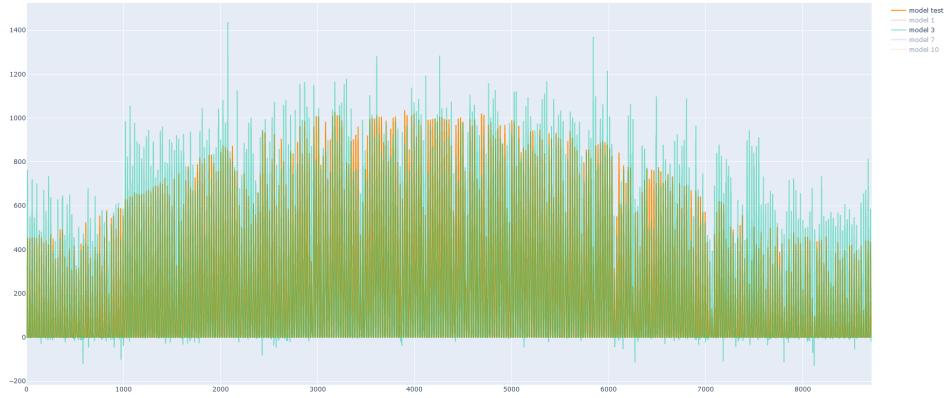
(a) sur l'année 2019



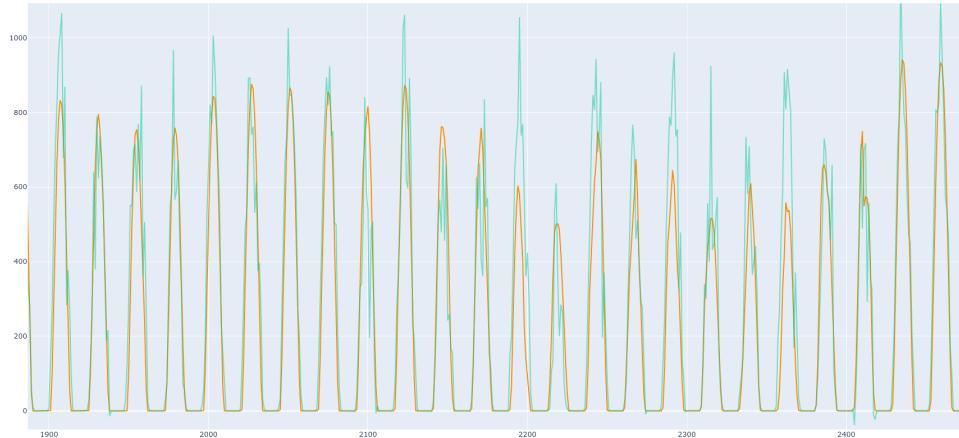
(b) sur quelques jours

FIGURE 3.17 – prédiction avec le "pire" modèle au sens des critères d'évaluation : ordre 1

Plus l'ordre augmente, plus la structure de la prédiction colle avec les données d'irradiance test, ce qui conforte dans l'idée que le modèle apprend bien la dynamique des données. Le modèle d'ordre 1 3.17 n'ayant qu'un état caché, il ne donne que des prédictions émises par un unique état. Les prédictions sont approximatives parfois aberrantes avec des valeurs négatives. Ceci est dû à la nature du modèle HMM à émission gaussiennes qui ne colle pas avec le support strictement positif des données. Un post-traitement des prédictions en seuillant à zéros toutes valeurs négatives est une alternative pour corriger ces aberrations. On peut donc conclure qu'un modèle à un seul état caché n'est pas assez pour les données d'irradiances présentées.



(a) sur l'année



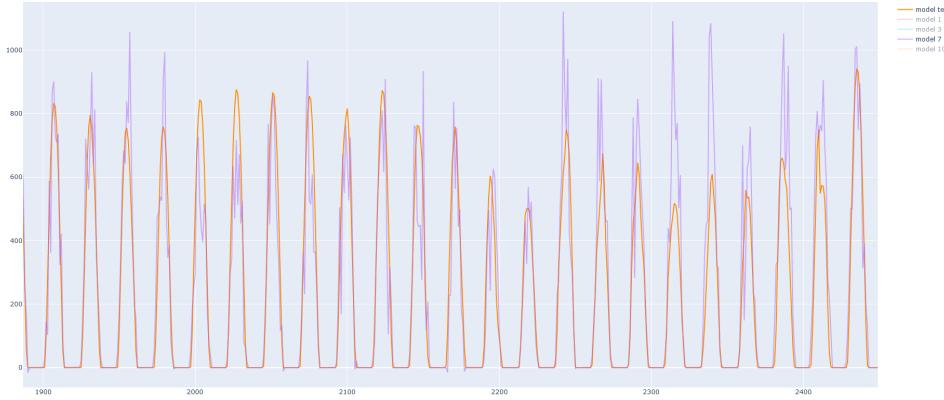
(b) sur quelques jour

FIGURE 3.18 – prédition avec un modèle "moyen" au sens des critères d'évaluation : ordre 3

Le modèle d'ordre 3 3.19 présente une nette amélioration comparé au premier modèle. La structure de la prédition suit la dynamique annuelle de l'irradiance dues à des émissions moyennes bien plus précises et une variété d'états cachées. Les valeurs négatives sont moins présentes et de plus petites amplitudes.



(a) sur l'année 2019



(b) sur quelques jour

FIGURE 3.19 – prédiction avec le "meilleur" modèle au sens des critères d'évaluation : ordre 7

Dans la même tendance, le modèle d'ordre 7 considéré comme le meilleur par l'étude antérieure fournit des prédictions plus précises avec des valeurs aberrantes, trop grandes ou négatives, encore plus rares.

De manière qualitative sur la base des prédictions ci-dessus, les modèles HMM semblent être sur l'expérience de ces prédictions des modèles apprenant facilement des profils redondants dans les données, contrairement aux modèles SARIMA.

Une remarque importante peut être notifiée concernant les valeurs des scores et leur comparaison entre les modèles. Les scores de vraisemblances, en l'occurrence d'AIC et BIC ne sont comparables ici qu'entre les modèles de la même famille, une comparaison "intra-modèles". Une comparaison des scores "inter-modèles", c'est-à-dire entre un modèle HMM et un modèle SARIMA, n'est pas possible dès lors que les méthodes de calcul de vraisemblance ne sont pas les mêmes. Autre chose, la vraisemblance dépend du modèle mais aussi des données. En effet pour un même modèle on peut avoir une vraisemblance - et donc un AIC ou BIC - qui varie pour un set de données différents. Bien que les deux modèles sont évalués sur le même jeu de données, nous devons les pré-traiter différemment ce qui ne rend plus envisageable la comparaison des scores "HMM-SARIMA". Ainsi, les scores étudiés permettent de faire dans un premier temps une comparaison "intra-modèle" entre chaque candidat issu de la même famille, où la vraisemblance a été calculé de la même manière.

Une manière de comparer quantitativement les modèles entre eux serait de trouver le candidat optimal, et ensuite de calculer un critère qui permet de faire un comparaison

"inter-modèle" Ainsi, une fois le modèle optimisé pour une famille donnée, le juge de paix pour comparer les candidats de chaque famille serait de calculer un score final du type RMSE ou cross-entropy pour quantifier les erreurs de chaque modèles entre eux.

4

Conclusions

Au cours de mon stage, j'ai eu l'opportunité de travailler dans un contexte de recherche sur la co-optimisation de réseaux hybrides. Plus précisément, j'ai dû fournir des modèles de prédictions d'irradiance solaire dans un contexte de co-optimisation d'un réseau gérant différents types de production d'énergies, et des postes de consommations. L'irradiance solaire et l'énergie thermique dans son ensemble, sont des ressources jouissant d'applications très diversifiées. L'irradiance solaire est une grandeur variable en temps et en espace, complexe à prédire précisément. Sa connaissance anticipée dans le temps permettrait de gérer le réseau de manière fine : les algorithmes de contrôle dans ce cas d'application tels que les MPC se basent sur cette idée d'anticipation de la dynamique du réseau. A l'aide de la routine de recherche de meilleurs modèles, les modèles de prédictions choisis vérifient des contraintes d'application dans un contexte de co-optimisation et font partie intégrante de la méthodologie de recherche de meilleur stratégie. Ma contribution se résume à avoir fourni une routine de recherche de meilleurs modèles au sein d'une famille de modèles discutée au préalable. La routine a été validée de prime abord sur des données artificielles avant de pouvoir caractériser les modèles sur les données réelles de prédiction.

Les modèles SARIMA et HMM ont été étudiées en particulier, et d'autres modèles pourront être intégrées dans la routine. Des critères d'évaluation ont été également définies, afin de pouvoir quantifier la performance de ces modèles sur les données d'apprentissage. Par conséquent, des premiers éléments de réponses quant à la bonne famille de modèles pour de la prédiction d'irradiance ont été fournies.

En effet les modèles SARIMA sont des modèles de régressions adaptées aux séries temporelles qui ont une nature auto-régressive, les données à un instant influant les données futures à des ordres à définir. Ce sont des modèles très simples et facilement implantables, mais qui peuvent présenter des limites si la donnée se montre non-linéaire avec des tendances et plusieurs dynamiques temporelles à des échelles différentes. A ce titre, la stationnarisation des données est une condition sine qua non pour pouvoir utiliser de tels modèles. Les résultats au cours du stage des modèles SARIMA montrent une performance discutable sur la base de modèles appris sur des données d'irradiance solaire stationnarisée en heure d'ensoleillement.

Des améliorations peuvent être faites pour ces modèles. En effet des modèles SARIMA incluant des régressions linéaires selon des variables exogènes sont utilisées dans la littérature. Par exemple utiliser le taux de couverture nuageuse comme variable exogène sont des modèles classiques dans la littérature et permettrait d'affiner les prédictions. Ainsi de manière plus générale, tous les modèles hybrides mélangeant des modèles de machine learning comme les régressions SARIMA - ou autres - avec des modèles physiques d'imagerie se présentent comme de bons modèles robustes. Le point noir de ces modèles est qu'ils

nécessitent en contrepartie un déploiement d'infrastructures, de capteurs précis pour établir des mesures fiables. Une étude plus fine sur les algorithmes d'optimisations POUR le problème de maximum de vraisemblance servant à estimer les paramètres des modèles SARIMA, serait également une piste d'amélioration.

Concernant les modèles HMM multivariés à émissions gausiennes.

Tels qu'utilisés dans les routines, les prédictions ont été faite sur un horizon de 24h fixé, en d'autres termes le modèle prédit un profil journalier complet d'irradiance, heure par heure. Il est possible, dans l'attente d'acquisition des observations des heures passées, de mettre à jour au cours de la journée l'état caché a posteriori et de prédire sur la base de cette nouvelle information. On pourrait utiliser alors pleinement la structure des HMM, qui permettent de fournir des prédictions probabilités bayésiennes, au sens où les prédictions sont mises à jours par de l'information a posteriori.

Ces modèles se montrent intéressants car contrairement aux modèles régressifs, ils sont adaptés pour des données présentant des profils, des motifs dans la série temporelle. On peut penser que ces modèles pourraient tout à fait décrire des cycles industriels de productions.

Ces modèles peuvent être agrémentés avec des termes auto-régressifs AR sur les émissions, ce qui ouvre sur des modèles nommés "Markov Switching -Autoregressive". On peut également essayer d'implémenter des modèles HMM avec une loi d'émission autre que gaussienne, comme une loi poisson par exemple. Cette loi en l'occurrence serait plus pertinente qu'une loi gausienne du fait du support positif des mesures d'irradiance. Concernant l'optimisation par l'algorithme Baum-Welch, la solution trouvée peut être locale et non globale, et il est préconisé de faire plusieurs simulations avec des initialisations des paramètres différents. L'initialisation opérée était aléatoire : on peut alors affiner l'étape d'initialisation par exemple avec une clusterisation du type k plus proche voisin ou autres.

La routine de recherche de meilleurs modèles peut également être amélioré, en prenant des scores prédictifs plus fins, comme le Brier Score ou le CRPS. Ces deux scores probabilistes requièrent toutefois la connaissance de la distribution des prédictions.

Sur un plan personnel, ce stage a pu confirmer mon intérêt de travailler dans le milieu de la recherche, à l'interface de plusieurs domaines scientifique, avec des applicatifs concrets et des problématiques qui me tiennent à cœur. Le cadre duplice de la recherche et le contact avec les industriel a été pour moi très complémentaire, où l'on peut amener des pistes d'améliorations sur des problèmes actuels ouverts. J'ai également développer la coordination de mon travail avec différents personnes de milieux scientifiques variés, et de bonnes compétences en rédaction d'articles. J'apprécie particulièrement le cadre de la recherche car il permet d'apprendre et tend à comprendre rigoureusement le choses.

J'ai quelques difficultés à adapter ma vision "ingénierie" des choses, qui est différente de ce que l'on peut demander en recherche. En effet je cherchais quelquefois à vouloir adapter le meilleur modèle sur les données, et me suis perdu dans du post-traitement, notamment sur les données SARIMA qui peut s'avérer très laborieux. Ce qui m'a été demandé de faire est différent : il n'était pas question d'optimiser et de calibrer les paramètres des modèles pour s'adapter absolument aux données. Ma contribution ne réside pas dans une méthode ou un modèle de prédition, mais plutôt à amener une discussion de modèles sur un problème et un contexte d'utilisation.

Par ailleurs, combiné avec mon stage de 4A, j'ai pu travailler sur une même thématique pendant une période assez longue. Travailler sur une période de dix mois cumulés sur le même projet, m'a permis d'avoir une vision large du problème, de pouvoir rentrer en détail dans l'élaboration d'une méthodologie, ce qui a été par conséquent une riche opportunité

de développer mes compétences et de me former sur d'autres matières scientifiques. Ce stage a été assez dense pour moi car il a fallu me documenter, me former sur des modèles nouveaux, faire des recherches bibliographiques, expérimenter diverses approches, dont la majorité n'a pas été concluante mais a nourri le cheminement scientifique tout au long du stage.

Bibliographie

- [1] R. AHMED, V. SREERAM, Y. MISHRA, AND M. D. ARIF, A review and evaluation of the state-of-the-art in pv solar power forecasting : Techniques and optimization, undefined, 124 (2020), <https://doi.org/10.1016/J.RSER.2020.109792>.
- [2] D. ALHAKEM, P. MANDAL, A. U. HAQUE, A. YONA, T. SENJYU, AND T. L. TSENG, A new strategy to quantify uncertainties of wavelet-grnn-psos based solar pv power forecasts using bootstrap confidence intervals, undefined, 2015-September (2015), <https://doi.org/10.1109/PESGM.2015.7286233>.
- [3] I. A. ASMI, K. KNOBLOCH, R. L. G. LATIMIER, T. ESENCE, K. ENGELBRECHT, AND H. B. AHMED, Thermocline thermal storage modeling towards its predictive optimal management, Journal of Energy Storage, 52 (2022), <https://doi.org/10.1016/j.est.2022.104979>.
- [4] I. A. ASMI, R. L. G. LATIMIER, H. B. AHMED, AND T. ESENCE, Impact of coupling thermal and electrical carriers on the optimal management of 0a multi-energy network, (2021), <https://doi.org/10.1109/EVER52347.2021.9456613>, <https://www.researchgate.net/publication/352699245>.
- [5] A. AZADEH, S. F. GHADERI, AND S. SOHRABKHANI, Forecasting electrical consumption by integration of neural network, time series and anova, undefined, 186 (2007), pp. 1753–1761, <https://doi.org/10.1016/J.AMC.2006.08.094>.
- [6] P. BACHER, H. MADSEN, AND H. A. NIELSEN, Online short-term solar power forecasting, undefined, 83 (2009), pp. 1772–1783, <https://doi.org/10.1016/J.SOLENER.2009.05.016>.
- [7] S. BHARDWAJ, V. SHARMA, S. SRIVASTAVA, O. S. SASTRY, B. BANDYOPADHYAY, S. S. CHANDEL, AND J. R. GUPTA, Estimation of solar radiation using a combination of hidden markov model and generalized fuzzy model, Solar Energy, 93 (2013), pp. 43–54, <https://doi.org/10.1016/J.SOLENER.2013.03.020>.
- [8] J. BOLAND, J. HUANG, AND B. RIDLEY, Decomposing global solar radiation into its direct and diffuse components, Renewable and Sustainable Energy Reviews, 28 (2013), pp. 749–756, <https://doi.org/10.1016/J.RSER.2013.08.023>.
- [9] Y. CHU, M. LI, C. F. COIMBRA, D. FENG, AND H. WANG, Intra-hour irradiance forecasting techniques for solar power integration : A review, iScience, 24 (2021), p. 103136, <https://doi.org/10.1016/J.ISCI.2021.103136>.
- [10] R. DE TRANSPORT D'ELECTRICITÉ, Futurs énergétiques 2050 principaux résultats résumé exÉcutif, tech. report, 2022, <https://assets.rte-france.com/prod/public/2021-12/Futurs-Energetiques-2050-principaux-resultats.pdf>.
- [11] U. DEPARTMENT OF ENERGY, Waste heat recovery : Technology and opportunities in u.s. industry, p. 112.

- [12] M. DIAGNE, M. DAVID, P. LAURET, J. BOLAND, AND N. SCHMUTZ, Review of solar irradiance forecasting methods and a proposition for small-scale insular grids, Renewable and Sustainable Energy Reviews, 27 (2013), pp. 65–76, <https://doi.org/10.1016/j.rser.2013.06.042>.
- [13] D. R. DREW, D. J. CANNON, D. J. BRAYSHAW, J. F. BARLOW, AND P. J. COKER, The impact of future offshore wind farms on wind power generation in great britain, Resources, 4 (2015), pp. 155–171, <https://doi.org/10.3390/RESOURCES4010155>, https://www.researchgate.net/publication/276274284_The_Impact_of_Future_Offshore_Wind_Farms_on_Wind_Power_Generation_in_Great_Britain.
- [14] A. T. HOANG, V. V. PHAM, AND X. P. NGUYEN, Integrating renewable sources into energy system for smart city as a sagacious strategy towards clean and sustainable process, Journal of Cleaner Production, 305 (2021), p. 127161, <https://doi.org/10.1016/J.JCLEPRO.2021.127161>.
- [15] R. H. INMAN, H. T. PEDRO, AND C. F. COIMBRA, Solar forecasting methods for renewable energy integration, Progress in Energy and Combustion Science, 39 (2013), pp. 535–576, <https://doi.org/10.1016/J.PECS.2013.06.002>, <https://www.scribd.com/document/501625034/2013-Inman-Pedro-Coimbra>.
- [16] J. B. JØRGENSEN, L. E. SOKOLER, L. STANDARDI, R. HALVGAARD, T. G. HOGAARD, G. FRISON, N. K. POULSEN, AND H. MADSEN, Economic mpc for a linear stochastic system of energy units, 2016 European Control Conference, ECC 2016, (2017), pp. 903–909, <https://doi.org/10.1109/ECC.2016.7810404>.
- [17] P. LAURET, M. DAVID, AND P. PINSON, Verification of solar irradiance probabilistic forecasts, Solar Energy, 194 (2019), pp. 254–271, <https://doi.org/10.1016/J.SOLENER.2019.10.041>.
- [18] M. MOHAMMADI, Y. NOOROLLAHI, B. MOHAMMADI-IVATLOO, AND H. YOUSEFI, Energy hub : From a model to a concept – a review, Renewable and Sustainable Energy Reviews, 80 (2017), pp. 1512–1527, <https://doi.org/10.1016/J.RSER.2017.07.030>.
- [19] D. R. MYERS, Solar radiation modeling and measurements for renewable energy applications : Data and model quality, undefined, 30 (2004), pp. 1517–1531, <https://doi.org/10.1016/J.ENERGY.2004.04.034>.
- [20] B. NANDI, S. BANDYOPADHYAY, AND R. BANERJEE, Analysis of high temperature thermal energy storage for solar power plant, 09 2012, <https://doi.org/10.1109/ICSET.2012.6357438>.
- [21] S. PFENNINGER AND I. STAFFELL, Long-term patterns of european pv output using 30 years of validated hourly reanalysis and satellite data, Energy, 114 (2016), pp. 1251–1265, <https://doi.org/10.1016/J.ENERGY.2016.08.060>.
- [22] M. Q. RAZA AND A. KHOSRAVI, A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings, Renewable and Sustainable Energy Reviews, 50 (2015), pp. 1352–1372, <https://doi.org/10.1016/J.RSER.2015.04.065>.
- [23] O. RUHNAU, L. HIRTH, AND A. PRAKTIKNJO, Time series of heat demand and heat pump efficiency for energy system modeling, Scientific Data 2019 6 :1, 6 (2019), pp. 1–10, <https://doi.org/10.1038/s41597-019-0199-y>, <https://www.nature.com/articles/s41597-019-0199-y>.

- [24] A. SFETSOS AND A. H. COONICK, Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques, Solar Energy, 68 (2000), pp. 169–178, [https://doi.org/10.1016/S0038-092X\(99\)00064-X](https://doi.org/10.1016/S0038-092X(99)00064-X).
- [25] E. SHARP, P. DODDS, M. BARRETT, AND C. SPATARU, Evaluating the accuracy of cfsr reanalysis hourly wind speed forecasts for the uk, using in situ measurements and geographical information, Renewable Energy, 77 (2015), pp. 527–538, <https://doi.org/10.1016/J.RENENE.2014.12.025>.
- [26] J. SHI, W. J. LEE, Y. LIU, Y. YANG, AND P. WANG, Forecasting power output of photovoltaic systems based on weather classification and support vector machines, IEEE Transactions on Industry Applications, 48 (2012), pp. 1064–1069, <https://doi.org/10.1109/TIA.2012.2190816>.
- [27] A. STAID, J.-P. WATSON, R. J.-B. WETS, AND D. L. WOODRUFF, Generating short-term probabilistic wind power scenarios via nonparametric forecast error density estimators, Wind Energy, 20 (2017), pp. 1911–1925, <https://doi.org/https://doi.org/10.1002/we.2129>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/we.2129>, <https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/we.2129>.
- [28] B. YANG, T. ZHU, P. CAO, Z. GUO, C. ZENG, D. LI, Y. CHEN, H. YE, R. SHAO, H. SHU, AND T. YU, Classification and summarization of solar irradiance and power forecasting methods : A thorough review, CSEE Journal of Power and Energy Systems, (2021), <https://doi.org/10.17775/CSEEJPES.2020.04930>.
- [29] H. T. YANG, C. M. HUANG, Y. C. HUANG, AND Y. S. PAI, A weather-based hybrid method for 1-day ahead hourly forecasting of pv power output, IEEE Transactions on Sustainable Energy, 5 (2014), pp. 917–926, <https://doi.org/10.1109/TSTE.2014.2313600>, <https://researchoutput.ncku.edu.tw/en/publications/a-weather-based-hybrid-method-for-1-day-ahead-hourly-forecasting->.
- [30] S. YOUNES, R. CLAYWELL, AND T. MUNEER, Quality control of solar radiation data : Present status and proposed new approaches, undefined, 30 (2005), pp. 1533–1549, <https://doi.org/10.1016/J.ENERGY.2004.04.031>.
- [31] D. ZIYATI, A. DOLLET, G. FLAMANT, Y. VOLUT, E. GUILLOT, AND A. VOSSIER, A multiphysics model of large-scale compact pv-csp hybrid plants, Applied Energy, 288 (2021), p. 116644, <https://doi.org/10.1016/J.APENERGY.2021.116644>.