

Chapitre II

Prédiction de l'irradiation solaire :

Etat-de-l'art

II.1 Introduction

Les séries temporelles, appelées aussi séries chronologiques ou même chroniques, occupent une place importante dans tous les domaines de l'observation ou de la collecte de données. Le terme série est employé pour évoquer des objets (des nombres ou des mots) classés dans un certain ordre. L'ordre utilisé est le temps, plus précisément, on utilise une mesure de temps exprimée en : années, mois, jours, minutes ou n'importe quelle autre unité de mesure. En d'autres termes, les séries temporelles associent des objets divers à des marques temporelles successives plus ou moins équidistantes. La série est dite temporelle, parce qu'elle indexe l'objet ou la valeur par le temps.

Dans ce chapitre, nous commencerons par expliquer ce que sont les séries temporelles, Nous dégagerons ensuite les différentes méthodologies de prédiction existantes dans la littérature et plus particulièrement celles liées au rayonnement global d'irradiation solaire (GHI) en détaillant les principaux modèles utilisables. L'état de l'art de la prédiction des séries temporelles va nous servir pour comparer les résultats de notre approche de prédiction avec la littérature.

II.2 Définition d'une série temporelle

Une série temporelle est définie comme étant une suite de mesures ou d'observations au cours du temps représentant un phénomène [29, 30,31]. Il est aujourd'hui, de plus en plus fréquent de parler de « prédiction ». Pour utiliser le formalisme des séries temporelles, il est nécessaire, au préalable de donner certaines définitions. Ainsi, la valeur courante en t de la chronique X est notée X_t où t , le temps, est compris entre 1 et n , avec n le nombre total d'observations. On appelle h le nombre de points ou de valeurs à prédire de la série temporelle. La prédiction de la série temporelle de $(n + 1)$ à $(n + h)$, connaissant l'historique de X_1 à X_n , porte le nom d'horizon de prédiction (horizon 1, ..., horizon h). Aussi pour un

horizon 1 (cas le plus simple), le formalisme générale la prédiction sera représenté par l'équation II-1 où ε représente l'erreur entre la prédiction et la mesure, f_n le modèle à estimer et t le paramètre temporel qui prend les $(n - p)$ valeurs suivantes : $n, n - 1, \dots, p + 1, p$. Où n est le nombre d'observations et p le nombre de paramètres du modèle, on suppose que $n \gg p$.

$$X_{t+1} = f_n(X_t, X_{t-1}, \dots, X_{t-p+1}) + \varepsilon(t + 1) \quad (\text{II.1})$$

où ε est l'erreur entre la valeur prédite et la valeur mesurée.

Il est toujours utile en première analyse de représenter l'évolution temporelle d'un phénomène (profil et allure de la série) à l'aide d'un graphique ayant en ordonnée la valeur du phénomène x_t et en abscisse le temps t . Ainsi sur la figure II-1, on peut se rendre compte que le phénomène de rayonnement global (GHI) X_t est un ensemble de signaux périodiques avec un bruit (lié à la couverture nuageuse) qui semble plus important durant les mois d'hiver que durant ceux d'été.

II.3 Approche stochastique du signal

Les techniques traditionnelles d'analyse des séries temporelles procèdent souvent par décomposition et recombinaison. L'approche de décomposition suppose que la structure de chaque chronique peut être scindée en éléments simples (modélisables), et donc plus facilement prévisibles, pour ensuite être reconstituée pour donner la prédiction de la chronique. Les premières études [30,32] sur les chroniques ont amené à considérer de façon standard trois grandes composantes de séries temporelles:

- **la tendance** : décrit le mouvement sur le long terme (extra-annuel).
- **la composante saisonnière** : est une composante cyclique relativement régulière de période intra-annuelle. Il existe différents types de saisonnalités pour le rayonnement global on peut parler de périodicité intrinsèque « rigide », ou déterministe car elle est bien marquée et répétitive.
- **la composante résiduelle (bruit ou résidu)** : correspond à des fluctuations irrégulières, en général de faible intensité mais de nature aléatoire. C'est une composante qui existe par défaut, elle regroupe ce que les autres composantes n'ont pu intégrer.

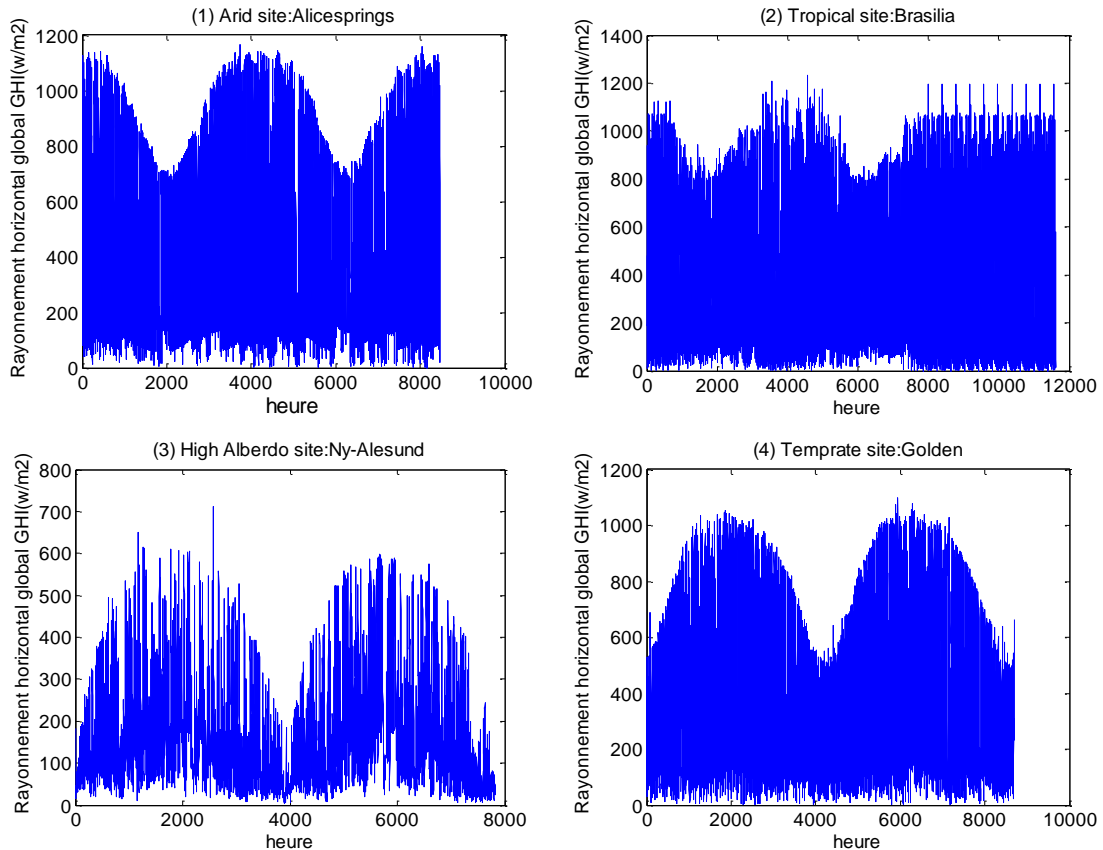


FIGURE II.1 Exemples de rayonnement global horizontal pour les sites Alice Springs, Brasilia, Ny-Alesund et Golden.

II.4 Fondamentaux pour la prévision du GHI

Dans cette section, quelques concepts de base sur l'irradiation solaire et la production d'énergie solaire sont expliqués, ce qui facilitera la compréhension des parties restantes du chapitre.

II.4.1 Modèles ciel clair

L'irradiation solaire est principalement influencée par la présence de nuages, qui rend difficile la prédiction d'irradiation. Cependant, il est possible d'estimer l'éclairement énergétique par temps clair, c'est-à-dire en l'absence de nuages. Cette valeur peut être utilisée pour calculer des indices solaires, normaliser des mesures et obtenir la production d'une centrale solaire dans des conditions stationnaires.

Généralement, les modèles de ciel clair sont alimentés avec des variables météorologiques et une géométrie solaire, en utilisant des modèles de transfert radiatif pour établir les connexions entre les entrées.

Il existe un grand nombre de modèles de ciel clair, qui diffèrent les uns des autres principalement par les entrées nécessaires à chaque modèle. Certains des modèles de ciel clair

les plus largement utilisés sont le modèle Solis [33], le modèle de l'atlas européen du rayonnement solaire ESRA (European Solar Radiation Atlas) [34], le modèle d'Ineichen [35] et le modèle d'évaluation de référence sur la transmission solaire[36].

Une description détaillée de ces modèles et d'autres modèles de ciel clair sont présentés dans les références [37] et [38]. Certains modèles ne nécessitent qu'une entrée (ESRA) alors que d'autres nécessitent un grand nombre d'entre eux (Solis). Comme détaillé dans [39], le choix d'un modèle à ciel clair pour un lieu déterminé dépend de la disponibilité et de la qualité des données d'entrée, qui constituent le principal facteur limitant.

En pratique, le modèle ciel clair présente les cycles journaliers (jour/nuit) et annuels comme illustré sur la figure II.2.

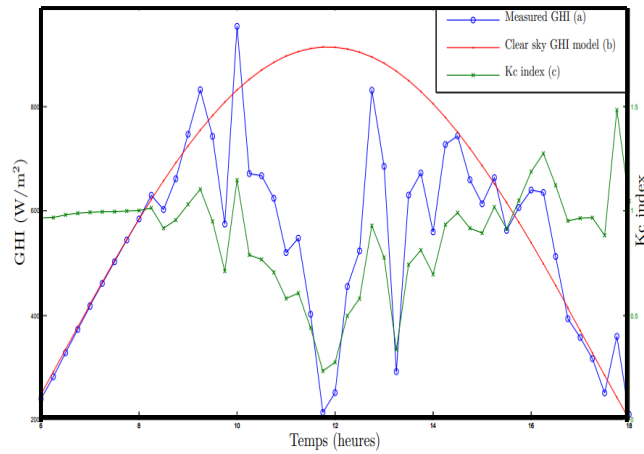


FIGURE II.2 Présentation des courbes du modèle clair ciel, GHI_m et l'indice k_c sur une journée de mesure [38].

II.4.2 Indice de ciel clair K_c

Il existe deux paramètres, l'indice de ciel clair k_c et l'indice de clarté k_t , qui sont largement utilisés pour classifier les conditions météorologiques et pour calculer des modèles de *persistance intelligents*. Ils sont obtenus de manière similaire, mais différents par la variable de normalisation. L'indice de ciel clair est le rapport entre l'irradiation solaire mesurée et l'irradiation solaire modélisée du ciel clair I_c . Cet indice est donc généralement préféré au GHI dans la plupart des études sur la prévision ou l'estimation locale du GHI.

$$K_c = \frac{GHI}{GHI_{\text{ciel clair}}} \quad (II.2)$$

II.4.3 Modèle de référence : la Persistance

Le modèle de persistance est reconnu comme la référence des méthodes de prédiction du GHI pour sa simplicité et son universalité. Il s'appuie sur l'hypothèse de persistance de la

mesure d'un instant à un autre de proche en proche. La prédiction d'une variable temporelle $X(t)$ par persistance peut être formulée par l'équation suivante :

$$\hat{X}(t+1) = \hat{X}(t) \quad (\text{II.3})$$

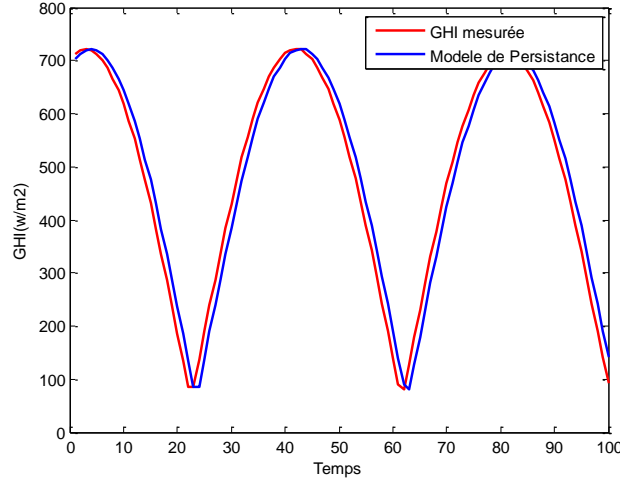


FIGURE II.3 Modèle persistance appliqué directement au GHI pour une journée ciel clair.

L'application de ce modèle à la prédiction du GHI nécessite l'utilisation de l'index K_c . En effet, en appliquant le modèle directement au GHI on obtiendrait une prédiction toujours en décalage avec la mesure à cause de la variation journalière connue (figure II.3). La prédiction de GHI par persistance s'écrit donc avec les équations suivantes :

$$\begin{aligned} \widehat{GHI}(t+1) &= \hat{K}_c(t+1) \times GHI_{\text{cielClair}}(t+1) \\ \hat{K}_c(t+1) &= \hat{K}_c(t) \end{aligned} \quad (\text{II.4})$$

Cette méthode de référence a l'avantage d'être très performante dès que les conditions météorologiques sont assez stables. Les scores de persistance sont d'ailleurs très importants pour évaluer et comparer des méthodes de prédiction développées et testées sur des jeux de données différents. Les scores du modèle de persistance donnent d'une part une référence pour la prévision mais aussi une information sur la difficulté de celle-ci, ou du moins sur la variabilité relative des jeux de données. Un bon score du modèle de persistance indique un site où les conditions météorologiques sont relativement stables et vice versa [40].

II.5 Les méthodes de prédiction des séries temporelles

La prédiction des séries temporelles est un problème qui recouvre de nombreux domaines d'application. Les études menées dans la finance et l'économétrie ont permis de dégager de nombreux modèles plus ou moins sophistiqués. Ces derniers ont été repris dans le

cadre d'autres thématiques, dont la prédiction du rayonnement solaire par modélisation des séries temporelles.

En ce qui concerne l'aspect temporel des prédictions, il est important d'introduire trois concepts : horizon de prédiction h , résolution de prédiction r et intervalle de prédiction f_i .

L'horizon de prédiction est la durée entre le temps présent t et le temps effectif des prédictions. La résolution de prédiction décrit la fréquence à laquelle les prédictions sont émises et l'intervalle de prédiction indique l'intervalle de temps des prédictions.

Il existe de nombreux modèles permettant de faire une prédiction de séries temporelles. Il est possible de les rassembler en quatre grands groupes.

- **Les modèles de type « naïf »** qui sont primordiaux pour vérifier la pertinence des modèles complexes. On peut citer la persistance, la moyenne ou les k -plus proches voisins ;
- **Les modèles à probabilités conditionnelles** rarement mentionnés dans la littérature en ce qui concerne le rayonnement global. On peut citer les chaînes de Markov et les prédictions basées sur les inférences Bayésiennes ;
- **Les modèles de type connexionnistes (réseau de neurones)** et plus particulièrement le Perceptron Multi-Couche (*PMC*), qui est un type de réseaux de neurones à fort potentiel prédictif et le plus souvent utilisé.
- **Les modèles de référence**, de par le nombre d'études les ayant utilisés, ils sont issus de la grande famille des modèles autorégressifs ;

Plusieurs modèles de prédiction du rayonnement solaire existent dans la littérature. Ils diffèrent par : les entrées disponibles, leur classification, l'horizon et la méthode de prédiction utilisée.

II.6 Classification de la prédiction de rayonnement solaire

Les chercheurs ont classé la prédiction du rayonnement solaire dans différentes catégories en fonction de différents facteurs qui sont liés à l'horizon de prédiction, données historiques de l'irradiation solaire et autres modèles de données météorologiques.

II.6.1 Classification selon l'horizon de prédiction

Le but et la précision d'un modèle de prédiction dépend de l'horizon de prédiction. Lipperheide et al. [41] ont analysé les performances de l'énergie photovoltaïque sur différents horizons de prédiction, tels que 20s, 40s, 60s,... 180s. L'erreur de prédiction (RMSE) du modèle de prédiction proposé est comprise entre 3,2 et 15,5% pour les horizons de prédiction

compris entre 20 et 180 s. Lonij et al. [42] ont conçu un modèle de prédiction de la puissance PV où les erreurs changent avec les horizons de prédiction allant de 15 min à 90 min. La précision de la prédiction varie avec le changement d'horizon de prédiction dans le même modèle avec les mêmes paramètres. Par conséquent, l'horizon de prédiction doit être pris en compte avant de concevoir le modèle de prédiction approprié. Il n'existe aucun critère bien défini pour classer le modèle de prédiction en fonction de l'horizon de prédiction. Néanmoins, selon la plupart des rapports de recherche, la prédiction de l'énergie solaire peut être divisée en trois catégories en fonction de l'horizon temporel, comme illustré à la figure II.4.

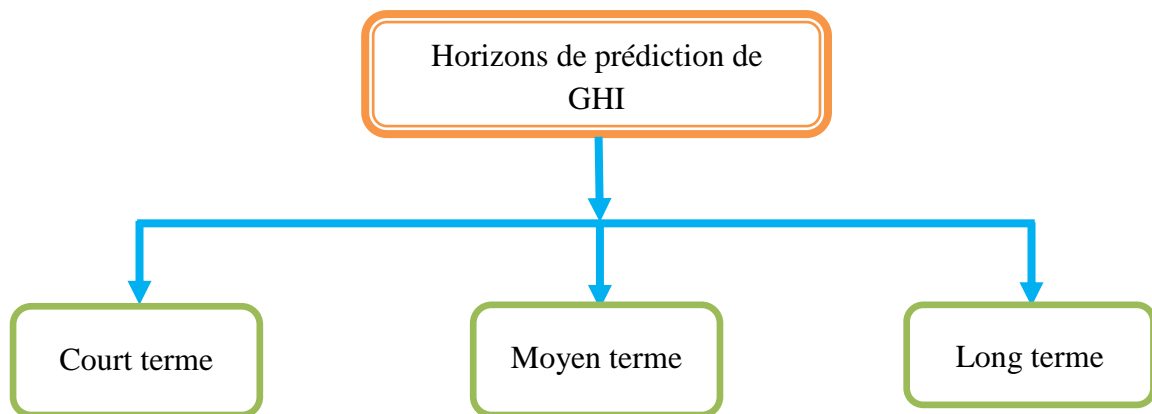


FIGURE II.4 Classification de la prédiction de GHI en fonction de l'horizon temporel.

II.6.1.1 Prédictions à court terme

La prédiction du rayonnement solaire réalisée pendant une heure, plusieurs heures, un jour ou jusqu'à sept jours sont appelées prédictions à court terme. La prédiction à court terme de l'énergie solaire garantit l'engagement, la planification et la répartition de l'énergie électrique des unités. Ce type de modèle de prédiction est utile pour concevoir un système de gestion de l'énergie intégré au système photovoltaïque. La prédiction à court terme améliore également la sécurité du fonctionnement du réseau.

II.6.1.2 Prédiction à moyen terme

La prédiction de rayonnement solaire à moyen terme est effectuée sur plus d'une semaine à un mois. Ce type de prédiction aide à la planification du système d'alimentation et le calendrier de maintenance en prévoyant la disponibilité de l'énergie électrique à l'avenir.

II.6.1.3 Prédiction à long terme

La prédiction à long terme du rayonnement solaire se fait à partir d'un mois à un an. Ce type de prédiction de rayonnement solaire est utile pour la planification de la production,

du transport et de la distribution de l'électricité en dehors des enchères d'énergie et de la sécurisation des opérations.

Cependant, certains chercheurs ont divisé l'horizon prévisionnel de l'énergie solaire en quatre catégories [43]. La quatrième catégorie est appelée "horizon de prédiction à très court terme". La prédiction à très court terme est prise en compte pour quelques secondes, une minute ou plusieurs minutes (<1 h) de prédiction. Ce type de prédiction a été fait pour le lissage de la puissance, répartition de l'électricité en temps réel et réserves optimales.

II.6.2 Classification de la prédiction de rayonnement solaire sur la base des données historiques

Les méthodes de prédiction de rayonnement solaire peuvent être classées en quatre types basés sur l'utilisation de données historiques de l'irradiation solaire et les variables météorologiques associées. Ces modèles sont (a) la persistance, (b) les méthodes statistiques, (c) l'apprentissage automatique et (d) les méthodes hybrides, comme indiqué à la figure II.5 avec les sous-catégories.

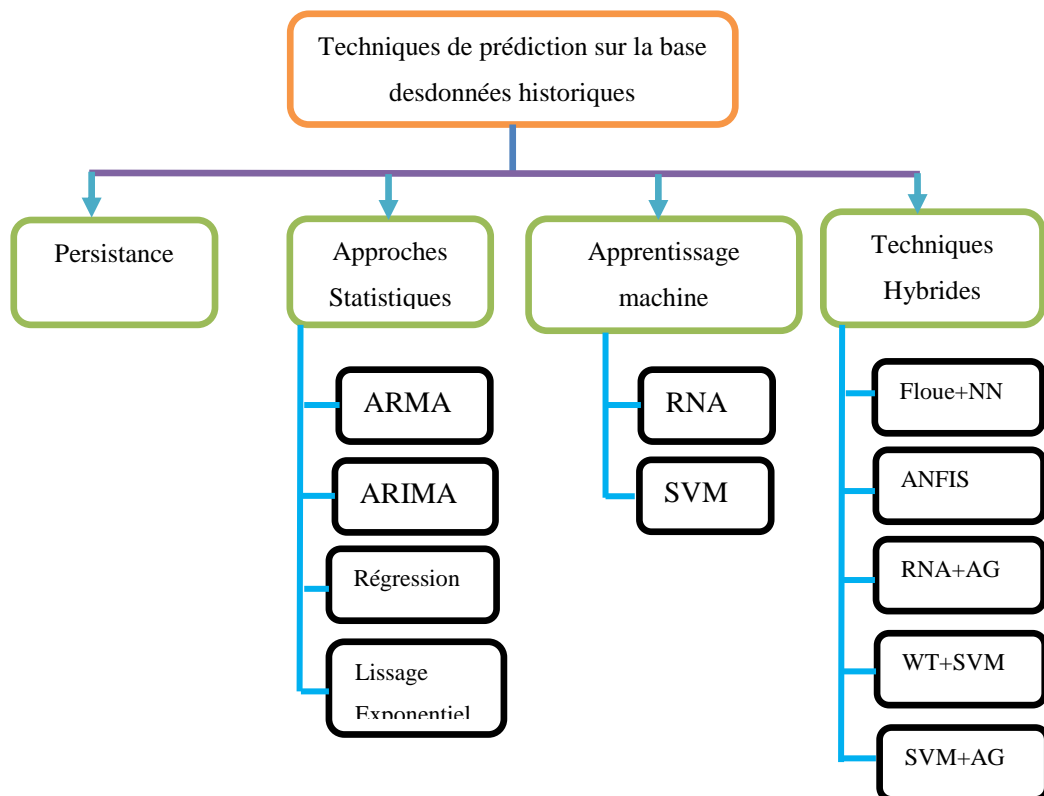


FIGURE II.5. Classification des techniques de prédiction utilisant des données historiques [43].

Dans le modèle de persistance, la valeur de l'irradiation solaire (IS) estimée à l'instant IS_{t+1} est égale à la valeur réelle précédente IS_t . Dans cette méthode, seules les données

historiques sont nécessaires pour prévoir l'IS. Généralement, ce modèle est utilisé comme modèle de référence (benchmark). Dans les méthodes statistiques, l'IS est prédite via l'analyse statistique des différentes variables d'entrée, par conséquent, les données chronologiques historiques sont utilisées dans ces méthodes.

Normalement, les méthodes précédentes sont adoptées pour la prédiction à court terme. Les données historiques récentes devraient être utilisées dans ces méthodes pour augmenter la précision du modèle. L'exigence de la taille de la série de données d'entrée dans ce modèle est moindre par rapport à la méthode d'apprentissage automatique. Inversement, dans l'apprentissage automatique, une grande quantité de données est nécessaire pour prévoir correctement l'IS avec précision. Les modèles d'apprentissage automatique sont des modèles intelligents, ils peuvent modéliser des données linéaires, non-linéaires et non-stationnaires. La combinaison de deux techniques ou plus est utilisée pour concevoir un modèle de prévision appelé modèle hybride. Le modèle hybride affiche de meilleurs résultats qu'un modèle unique pour différents problèmes de prévision en combinant les avantages de chaque technique.

II.6.3 Classification des méthodes de prédiction du rayonnement solaire

De nombreux chercheurs ont étudié les différentes méthodes de prédiction de rayonnement solaire. Ils ont proposé et développé plusieurs techniques et modèles pour prévoir la prédiction de rayonnement solaire. Parmi les méthodes de prédiction utilisées dans la littérature, certaines sont basées sur des modèles linéaires tels que la régression linéaire (Linear Regression LR), la moyenne mobile autorégressive (Auto-Regressive Moving Average ARMA) et le processus autorégressif (Auto-Regressive AR) [44, 45]. Cependant, le comportement non linéaire du rayonnement solaire a suscité les chercheurs à proposer plusieurs autres modèles : ceux basés sur le modèle numérique météorologique, les ondelettes, les modèles flous, les systèmes d'inférence neuro-flou adaptatifs (Adaptive Neural Fuzzy Inference Systems ANFIS), les forêts aléatoires (Random Forests RF), les k-plus proches voisins (k-Nearest Neighbors kNN) ainsi que les Réseaux de Neurones Artificiels (RNA) [46].

II.6.3.1 Processus autorégressifs

Ces techniques permettent d'estimer la relation entre une variable dépendante (irradiation solaire) et des variables indépendantes, appelées des fois *prédicteurs*. Selon la manière dont les séries chronologiques sont traitées (linéaire ou non linéaire et stationnaire ou non-stationnaire), une autre classification apparaît. Les séries temporelles stationnaires sont

des séries chronologiques fluctuant autour d'une moyenne statique alors que les séries non stationnaires ne montrent pas une telle moyenne.

Nous trouvons ici les modèles autorégressifs (AR), qui modélisent la sortie sous forme d'une combinaison linéaire des valeurs retardées des prédicteurs; les modèles simples de moyenne mobile (MA) utilisés lorsque les données présentent une variance constante sur une position d'équilibre autour de la moyenne, pour laquelle la moyenne des données historiques est utilisée comme prédiction; les modèles double MA utilisés lorsqu'il y a une tendance; les modèles à moyenne mobile autorégressive (ARMA) qui tiennent compte à la fois des valeurs antérieures décalées et des erreurs; Les modèles AR eXogenous (ARX), qui ajoutent des données exogènes à un modèle AR, et les modèles de moyenne mobile autorégressive à variables eXogenes (ARMAX), qui introduisent des variables externes. Dans l'analyse des séries chronologiques (c.-à-d. à partir de la prédiction numérique du temps), pour traiter l'aspect probabiliste, il existe certaines adaptations, telles que le vecteur AR (VAR) ou le vecteur ARX (VARX) [47,48].

II.6.3.2 Méthodes de régression

La méthode de régression est une méthode statistique utilisée pour établir une relation entre les variables explicatives et dépendantes. Dans ce modèle, la variable dépendante est prédite en connaissant les variables explicatives. Dans le cas de la prédiction de rayonnement solaire, le RS prédit est considéré comme une variable dépendante et les variables météorologiques sont considérés comme des variables explicatives. Oudjana et al. [49] ont développé un modèle pour prédire la production d'énergie photovoltaïque, comprenant deux modèles de régression différents : les régressions simples et les régressions linéaires multiples. Le modèle de régression utilisant l'irradiation solaire et la température en entrée a donné de meilleurs résultats par rapport au cas où l'un ou l'autre était considéré en entrée. Un modèle mathématique et plusieurs variables explicatives sont nécessaires pour concevoir un modèle de prévision basé sur la régression, ce qui constitue la faiblesse de cette méthode.

II.6.3.3 Méthodes de lissage exponentiel

La méthode de lissage exponentiel, suggérée pour la première fois par Brown [50], est connue sous le nom de méthode de lissage exponentiel simple. Cette méthode a été développée par Holt [51], connue sous le nom de méthode de Holt; Winter a également modifié cette méthode sous le nom de Holt-Winter méthode [52]. Dans la méthode de lissage exponentiel, un ensemble de pondérations des données historiques au lieu des pondérations

égales est imposé aux données passées. Cependant, les poids des données passées diminuent de manière exponentielle des points de données les plus récents aux plus distants. Le lissage exponentiel simple est une méthode simple, également connue sous le nom de MA pondérée de manière exponentielle (EWMA). La forme du modèle EWMA est écrite comme suit:

$$\hat{X}_{t+1} = \alpha X_t + (1 - \alpha)\hat{X}_t = \hat{X}_t + \alpha(X_t - \hat{X}_t) \quad (\text{II.5})$$

où α est la constante de lissage et peut prendre toute valeur entre 0 et 1. Cette méthode nécessite une prédiction initiale qui doit être estimée ou supposée. Par conséquent, l'ensemble initial de prédictions \hat{X}_t est souvent considéré comme $\hat{X}_t = X_t$. Le modèle de prédiction EWMA indique que la valeur de prédiction à la période $t + 1$ est égale à la somme des dernières valeurs prédites \hat{X}_t plus un terme d'ajustement de l'erreur de prédiction $\alpha (X_t - \hat{X}_t)$.

II.6.3.4 Méthodes physiques

La méthode de prédiction physique consiste en un ensemble de méthodes mathématiques. L'équation qui décrit l'état physique et le mouvement dynamique de l'atmosphère [53]. Les modèles physiques sont conçus en fonction des caractéristiques du site, telles que la localisation, les différentes variables météorologiques et les données historiques d'orientation. Ces modèles sont considérablement simples (lorsque basés uniquement sur l'irradiance solaire globale) ou compliqués (s'ils incluent des paramètres supplémentaires) [54]. La précision du modèle de prédiction physique est supérieure lorsque les conditions météorologiques sont stables [55]. Cependant, les performances de prédiction sont largement affectées par les fortes variations des variables météorologiques. Différentes études ont proposé [54,56], la méthode de prédiction physique combinée à différentes méthodes d'intelligence artificielle et statistiques afin de concevoir un modèle de prédiction hybride peut avoir généralement une meilleure précision de la prédiction. Cependant, ces modèles sont très sensibles à la prédiction météorologique.

II.6.3.5 Méthodes d'intelligence artificielle

- **Réseau de neurones artificiels (RNA)**

Le RNA est la méthode la plus efficace et populaire parmi les chercheurs depuis 1980. Cette méthode a été utilisée dans différentes applications de prédiction, y compris la prédiction du rayonnement solaire avec un niveau de réussite parmi les plus élevés. Les RNA sont largement utilisés pour prévoir le rayonnement solaire dans la plupart des recherches, en raison de la non-linéarité des données météorologiques. Le RNA est plus approprié que les

méthodes statistiques lorsqu'un lien non linéaire existe entre les données sans hypothèse préalable.

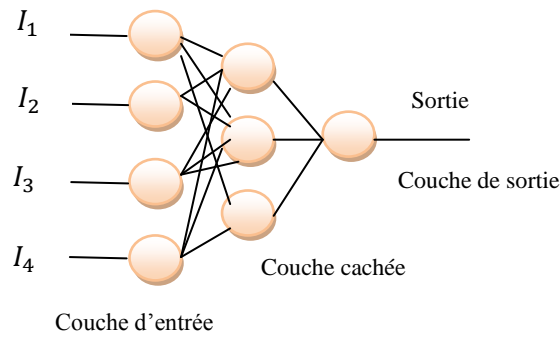
Les principaux composants d'un RNA sont l'entrée, la/les couches cachées, les couches d'entrée et de sortie, les neurones et les connexions. La couche d'entrée reçoit les différentes informations d'entrée. La couche cachée, qui peut consister en un seul ou plusieurs couches, analyse les informations d'entrée. La couche de sortie reçoit les résultats analysés et fournit la sortie (valeur prédite). La connexion fait un lien entre les neurones des différentes couches avec mise à jour des poids. La figure II.6 (a),(b) montre l'architecture de base d'un RNA et un modèle schématique de la procédure de traitement dans une cellule de neurone, respectivement.

La figure indique que la cellule neuronale a deux parties. La première partie est la «fonction de combinaison» qui produit une valeur en résumant toutes les entrées. La deuxième partie est la "fonction d'activation". La sortie du réseau est générée en additionnant les entrées pondérées en utilisant la fonction d'activation. Par conséquent, la fonction d'activation du réseau agit comme une fonction de compression pour transférer l'entrée sous la forme de sortie. La formule mathématique de base du RNA exprimée par [57] :

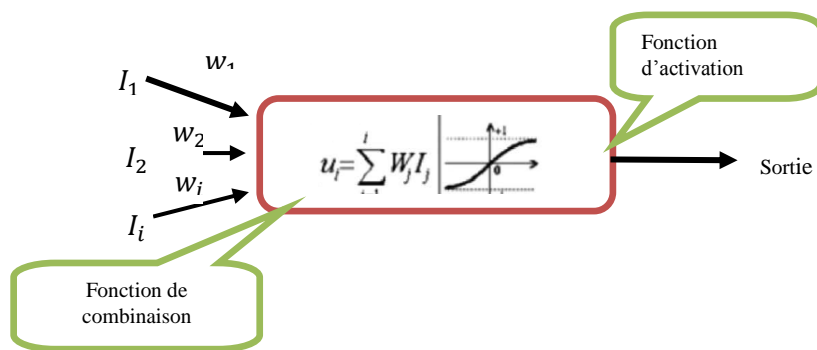
$$U_N = b + \sum_{j=1}^N (W_j \times I_j) \quad (II.6)$$

Où U_N , b , N , W_j et I_j sont la sortie du réseau, poids de biais, nombre d'entrée, poids des connexions et entrée réseau, respectivement. Dans la plupart des cas, le RNA a plusieurs entrées et une ou plusieurs sorties.

L'apprentissage et les tests sont les deux opérations de base d'un RNA. À la phase d'apprentissage, le réseau est formé à l'aide d'une base de données *d'apprentissage* via l'algorithme d'apprentissage. L'algorithme d'apprentissage du RNA tente de mapper les relations d'entrées et de sortie(s) en mettant à jour les valeurs de poids synaptique. La sortie générée par le réseau est comparée à la valeur souhaitée, l'erreur de la sortie est ensuite calculée. Par conséquent, les poids et les biais (agissant en tant que compensateurs), les valeurs de RNA sont mises à jour en fonction de l'erreur. Ce cycle continue jusqu'à ce que la sortie souhaitée soit atteinte. Ainsi, le réseau fournit la sortie finale en fonction de la base de données utilisée et des valeurs de pondération du modèle. Cependant, la sortie du réseau varie en fonction du changement d'architecture, de la fonction d'activation et des entrées



(a)



(b)

FIGURE II.6 (a) Diagramme schématisique d'une structure RNA composée d'une couche d'entrée, une couche cachée et une couche de sortie.(b)modèle mathématique du neurone artificiel [59].

Dans la plupart des cas, les RNA ont une seule couche et sont utilisés pour résoudre les différents problèmes. Cependant, un nombre important de problèmes complexes qui sont liés au type de données peuvent utiliser un RNA monocouche. Dans plusieurs cas, des relations complexes d'entrée et de sortie existent entre différentes variables. Afin de surmonter ce problème complexe, les RNA ont été modifiés en plusieurs types qui suivent différentes architectures et différentes procédures de mappage entrée-sortie. Parmi les plus utilisés sont le perceptron multicouche (MLP), RNA à fonction de base radiale (RBF), RN récurrent (RNN), RN à régression générale (GRNN) et systèmes adaptatif d'interface neuro-floue (ANFIS).

Le RNA donne de bons résultats en apprenant différentes relations complexes et structures de calcul. Par conséquent, il est considéré comme un bon outil pour la prédiction des données chronologiques [60]. Yona et al. ont signalé que des erreurs dans les résultats de prédiction sont considérablement minimisées par RNA par rapport à FFNN [61]. GRNN est un réseau probabiliste qui fait la régression plutôt que des tâches de classification [62]. Avec

une forte capacité non-linéaire de cartographie, GRNN peut résoudre efficacement des problèmes non linéaires [63]. Cependant, la croissance de cette méthode nécessite une grande puissance de calcul. La sélection du paramètre de réseau σ est importante pour l'utilisation de GRNN [64]. Les chercheurs ont introduit un réseau modifié de BP pour remédier à ces limitations [65]. Parmi les méthodes liées à RNA, le réseau de neurones à propagation arrière (BPNN), qui a été largement utilisé en raison de son excellente fonction de cartographie non linéaire, convient à la résolution des problèmes de régression complexes [66].

ANFIS est un type de réseau multicouche adaptatif anticipé ;il est appliqué à la prédiction non linéaire, dans laquelle des échantillons antérieurs sont utilisés pour prévoir l'échantillon à l'avance [67]. Le réglage des fonctions d'appartenance est nécessaire pour améliorer les performances d'ANFIS [68]. Parmi les systèmes flous, l'ANFIS est le plus largement utilisé car il est moins coûteux en calcul et produit des résultats aussi robustes que les modèles statistiques [69].

- ***Machines à vecteurs de support (SVM).***

Il s'agit d'une méthode de modélisation supervisée, introduite pour la première fois par [70], puis développée par [71] pour être utilisée dans les problèmes de classification. Lorsqu'elles sont appliquées à des problèmes de régression, elles sont appelées machines de régression vectorielle (Support Vector Regression ou SVR). Elles se distinguent par leur forte capacité de généralisation et leur capacité à traiter des problèmes non linéaires. Elles fonctionnent comme une régression linéaire multiple utilisant des prédicteurs transformés tout en conservant une faible complexité et un bon ajustement des données. Trois paramètres principaux dominent les performances de la technique et doivent être ajustés : la précision, le paramètre de coût, qui concerne le compromis entre précision et complexité, et γ , qui régule la fonction du noyau, utilisée pour transformer les prédicteurs en un espace de fonctions de plus grande dimension. Les SVM / SVR ont montré un grand potentiel dans plusieurs études [72,73].

- ***k- plus proches voisins (k-NN)***

C'est l'une des méthodes les plus simples d'apprentissage machine. Il repose sur un algorithme de reconnaissance de formes, qui compare l'état actuel à des échantillons d'apprentissage dans un espace de fonctions. Les distances euclidiennes sont ainsi calculées et les k premiers voisins les plus proches sont sélectionnés pour les prédictions.

- **Forêts Aléatoires (RF)**

Développées dans un premier temps par Breiman (2001), elles consistent en un ensemble d'arbres de décision / régression, dont les résultats montrent la prédiction moyenne des arbres individuels. Normalement, en se concentrant sur les arbres de régression, elles se caractérisent par le sur-apprentissage des données de formation, ce qui donne un biais faible mais une variance élevée. Il est courant de cultiver plusieurs arbres pour chaque étude de cas en faisant la moyenne des résultats de plusieurs arbres, il est possible de réduire la variance au prix d'une légère augmentation du biais. Néanmoins, la performance globale du modèle s'améliore. Les arbres de régression simples sont très sensibles au bruit de la base de données d'apprentissage. Le simple calcul de la moyenne des résultats de plusieurs arbres pour résoudre ce problème ne fonctionne pas, car s'ils étaient entraînés avec la même base de données, ils seraient très corrélés. Ce problème est résolu avec la méthode bagging, qui consiste à faire construire des arbres, chacun avec un échantillon initialisé (*bag*) de l'ensemble d'apprentissage. Les échantillons laissés en dehors du modèle sont utilisés pour évaluer les performances de chaque modèle.

Néanmoins, les arbres peuvent toujours être corrélés, dans la mesure où certaines variables sont des prédicteurs très puissants pour la sortie, elles seront sélectionnées dans la plupart des arbres en sac. Les forêts aléatoires traitent ce problème avec la mise en sac des entités, qui consiste à sélectionner un sous-ensemble aléatoire d'entités au niveau de chaque nœud et à réduire ainsi la corrélation.

La Figure.II.7 montre la répartition des études analysées concernant la technique utilisée. Comme on le voit, l'approche la plus courante parmi les articles analysés est l'utilisation de techniques statistiques, en particulier les RNA, qui représentent 24% des études.

II.6.3.6 Modèles hybrides

Certains modèles peuvent omettre certaines informations en raison de la façon dont chaque technique transforme les données. Ainsi, il est également courant de combiner des techniques pour renforcer leurs atouts afin d'améliorer la précision, ce qui est appelé modèle hybride, modèle mixte, modèle combiné ou modèle d'ensemble. Les modèles peuvent être mélangés de plusieurs manières, telles que bagging, boosting, voting ou stacking[76,77].

Deux approches peuvent être suivies, soit en combinant deux techniques statistiques ou plus (hybride statistique), soit en associant une technique statistique à un modèle de performance (hybride physique). Plusieurs travaux peuvent être trouvés en utilisant la

première approche. Bouzerdoum et al. [78] ont associé SARIMA à SVM dans leurs prévisions avec un horizon d'une heure, alors que Ramsami et Oree[79] ont utilisé ARIMA avec ANN, Vaz et al. [80] ont appliqué une technique d'ensemble de RNA et d'ARX non linéaire (NARX). Hossain et al. [81], ont mis au point une méthodologie pour sélectionner la meilleure combinaison de techniques de régression afin de construire un modèle d'ensemble, ils ont montré qu'ils ont des performances meilleures que les techniques individuelles.

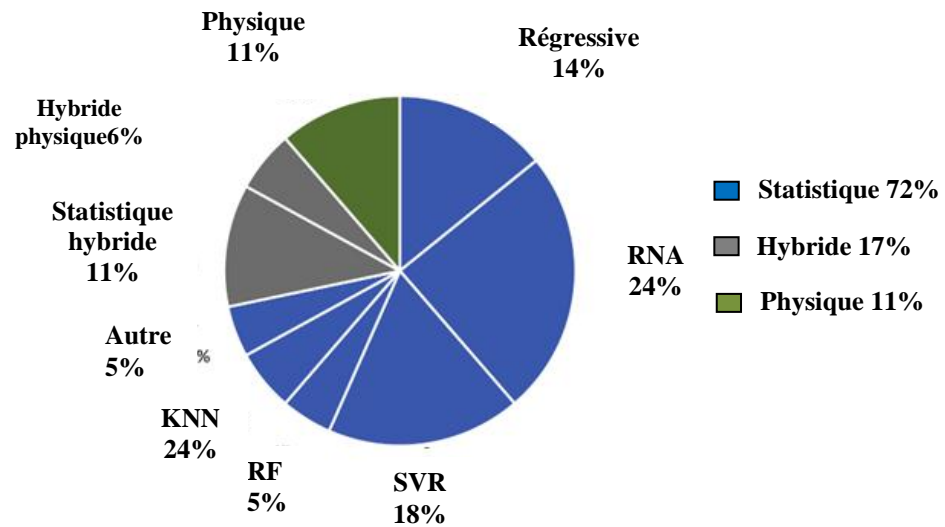


FIGURE II.7 Répartition des études par rapport à la technique utilisée [43].

Certaines études, cependant, ont construit un modèle boîte noire (black box) pour prédire mais ils ont inclus dans sa structure des expressions physiques. C'est le cas de la méthode du réseau de neurones artificiels hybride physique (PHANN), qui combinait un réseau RNA avec un modèle ciel clair. Ils ont également combiné certaines expressions physiques, telles que le calcul de l'indice de clarté et la température, avec les techniques NAR et RNA [82].

Dans [83], un modèle hybride composé d'un modèle ARMA et d'un réseau de neurones à retard temporel (Time Delay Neural Network TDNN) est examiné. Ce modèle est utilisé pour la prédiction du rayonnement solaire à court terme (une heure). Ce modèle hybride a le potentiel d'exploiter les avantages de ces deux techniques. Le modèle ARMA s'est avéré adapté au problème linéaire et le TDNN est efficace pour le problème non-linéaire. Dans le cas du rayonnement solaire contenant à la fois des caractéristiques linéaires et non linéaires, la précision de ce modèle hybride est assez satisfaisante. Cependant, ces différents modèles

fonctionnent très bien quand le temps est clair, mais lorsque les conditions météorologiques sont mauvaises, la précision des modèles de prévision diminue considérablement.

Cependant, la complexité informatique est accrue dans un modèle hybride en raison de l'utilisation de deux techniques ou plus. La performance d'un modèle hybride dépend de la performance d'un modèle individuel. La précision du modèle hybride est affectée par le choix d'une seule technique susceptible d'être peu performante, ce qui constitue une limitation du modèle hybride.

II.7 Méthode d'ensemble

La méthode d'ensemble est populaire en statistiques et en apprentissage machine, elle utilise plusieurs prédicteurs pour obtenir une décision agrégée qui sera meilleure que n'importe lequel des prédicteurs de base. Selon [84], il existe deux types de méthodes d'ensemble : compétitif et coopératif. Semblable à la classification, la prédiction d'ensemble peut être catégorisée en prédiction d'ensemble concurrentielle et coopérative. Un Ensemble compétitif de prédiction consiste à former différents prédicteurs individuels soit avec différents jeux de données ou bien avec le même ensemble de données, mais avec des paramètres différents, puis la prédiction est obtenue en faisant la moyenne (ou autre approche) de la décision de tous les prédicteurs individuels (prédicteurs de base). D'autre part, la prédiction d'ensemble coopérative consiste à diviser la tâche de prédiction en plusieurs sous-tâches et à sélectionner les prédicteurs appropriés pour chaque sous-tâche en fonction des caractéristiques des sous-tâches. La décision finale est la somme de tous les résultats des prédicteurs de base.

II.7.1 Ensemble de prédiction compétitif

L'approche de prédiction compétitive globale utilise plusieurs prédicteurs construits avec des conditions initiales légèrement différentes ou des paramètres différents, permettant de construire des modèles de prédiction individuels afin de former un modèle de prédiction d'ensemble (ensemble multi-modèle). Les résultats de prédiction de tous les modèles ou des modèles sélectionnés après l'élagage est agrégé par moyenne ou autres méthode de combinaison. Le niveau de confiance peut être mesuré par les variations de la dispersion des résultats individuels utilisés lors du processus de combinaison [85].

La diversité est un élément clé de la prédiction d'ensemble compétitive. Si les prédicteurs de base rendent des décisions similaires, il y aura moins d'amélioration. En prédiction concurrentielle (compétitive), si les sous-tâches sont similaires, les sorties des prédicteurs de base seront similaires et l'amélioration des performances d'ensemble sera

marginale. Pour la classification, la diversité peut être encore catégorisée comme diversité de données, diversité de paramètres et diversité de noyau [86,87]. Un schéma fonctionnel de la prédiction d'ensemble concurrentiel est montré dans la figure II.6. Il existe deux variantes basées sur la diversité des données et la diversité de paramètres.

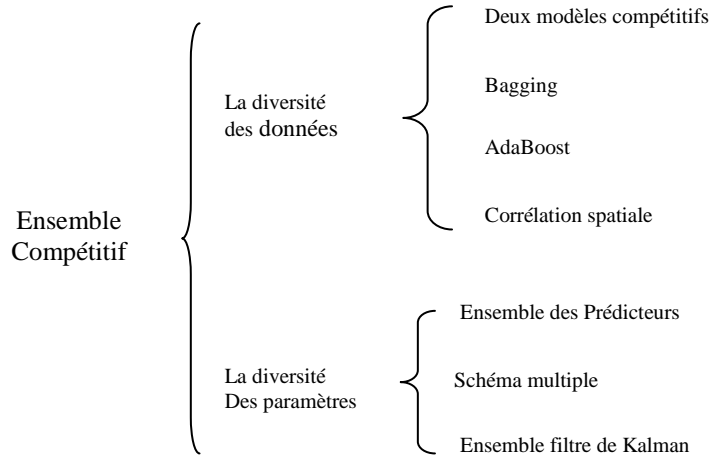


FIGURE II.8 Schéma fonctionnel des méthodes de prédiction d'ensemble concurrentielles.

II.7.1.1 Diversité des données

Pour la prédiction d'ensemble compétitive basée sur la diversité des données, plusieurs ensembles de données d'entrée sont introduits dans le système de prédiction. Il existe deux variantes, comme indiqué dans les équations. (II.7) et (II.8). Éq. (II.7) représente l'approche de prédiction qui applique N prédicteurs. $f_1(.) \dots f_N(.)$ pour N ensembles de données d'entrée $X_1 \dots X_N$ et la prédiction finale est une moyenne pondérée. Eq(II.8) représente une autre approche de prédiction qui utilise un seul prédicteur $f(.)$ pour N ensembles de données d'entrée.

$$\hat{y}(t+h) = \frac{1}{n} \sum_{i=1}^N w_i f_i(X_i(t)) \quad (\text{II.7})$$

$$\hat{y}(t+h) = f(X_1(t), \dots, X_N(t)) \quad (\text{II.8})$$

Où \hat{y} est la valeur prédite et h est l'horizon de prédiction.

Une étude sur la prédiction de l'irradiation solaire à Fontana, en Californie en 2009-2010 a été reportée [88]. Un modèle concurrentiel à deux méthodes de prédiction d'ensemble a été proposé : le premier modèle prédit l'irradiation solaire en utilisant une régression non linéaire des données météorologiques et le deuxième modèle prédit l'irradiation solaire basé sur la reconnaissance des formes. Les sorties de ces deux modèles ont été combinées pour

l'horizon de prédiction de 1h-3h. Les auteurs ont comparé leurs résultats avec les résultats précédemment rapportés et ont déclaré que la méthode de prédiction d'ensemble concurrentiel à deux modèles présente une erreur plus petite que les erreurs des méthodes précédemment rapportées, les auteurs ont également proposé d'utiliser une méthode d'apprentissage adaptatif pour mettre à jour les modèles sur une base journalière.

Le bagging est une méthode populaire de la théorie d'ensemble [89]. Elle comporte deux phases: 1) la phase de bootstrap, consiste à échantillonner la base de données d'origine avec remplacement pour obtenir N bases de données. 2) la phase d'agrégation pour combiner les sorties des N bases, les prédicteurs formés par chaque jeu de donnée sont mémorisés. Dans [90], un modèle RNA avec bagging a été introduit pour la prédiction à court terme de l'irradiation solaire.

Le modèle RNA avec bagging se compose de trois types différents des RNA: MLP, fonction de base radiale (RBFNN) et réseau de neurones récurrentes (RNN) avec des données historiques. Les données utilisées provenaient de l'Agence Météorologique Japonaise pendant la période 2007–2008. Les résultats ont montré que les mesures d'erreur de RNA avec bagging étaient plus petites par rapport aux modèles individuels.

Pour la prédiction de l'irradiation solaire, la corrélation spatiale est largement utilisée. Cette méthode est appelée totale satellite image (TSI), elle est habituellement utilisée dans les prédictions à court terme [91]. Les images du ciel sont prises par des satellites avec différentes coordonnées spatiales et temporelles. Dans [92], un prédicteur d'irradiation solaire par corrélation spatiale basé sur le RNA a été mis en œuvre avec 17 bases de données géographiques et météorologiques en Turquie de 2000 à 2002.

II.7.1.2 Diversité de paramètres

La diversité des paramètres par rapport aux variables... $\theta_1 \dots \theta_N$ d'un prédicteur pour produire N prédicteurs $f_1(.) \dots f_N(.)$ et chaque prédicteur apprendra du même base de donnée x. La valeur prédite \hat{y} est obtenue par la moyenne des performances de tous les prédicteurs comme il est montré en équation II.9.

$$\hat{y}(t + h) = \frac{1}{N} \sum_{i=1}^N f_i(X(t), \theta_i) \quad (\text{II.9})$$

Où \hat{y} est la valeur prédite et h est l'horizon de prédiction.

II.7.2 Prédiction d'ensemble coopérative

La prédiction coopérative par ensembles divise une tâche de prédiction en plusieurs sous-tâches et résout chaque sous-tâche individuellement. Il existe deux variantes: l'une est en cours de prétraitement et l'autre en post-traitement.

II.7.2.1 Prétraitement

Le prétraitement consiste à diviser la base de données d'entrée en plusieurs sous-ensembles de données. Chaque sous-ensemble de données est modélisé et prédit par un prédicteur. Généralement, les prédicteurs sont identiques pour tous les sous-ensembles de données. La prédiction finale est une somme de toutes les sorties des prédicteurs.

Une méthode de décomposition en série chronologique appelée décomposition en ondelettes a été rapportée dans plusieurs publications [93,94]. La théorie des ondelettes consiste à étudier les séries temporelles dans le domaine fréquentiel ainsi que dans le domaine temporel. La décomposition en ondelettes consiste à décomposer la série temporelle en un ensemble de sous-séries basées sur une ondelette mère qui peut être prédite avec plus de précision. Il existe deux types de décompositions: continue et discrète. Pour des applications pratiques, la transformée en ondelettes discrète (DWT) est généralement utilisée pour la décomposition. Les équations clés de la prédiction basée sur la décomposition en ondelettes sont présentées dans l'équation II.10.

$$\begin{aligned} \{X_{D_i}(t), X_{A_n}(t)\} &= \text{DWT}(X(t)) \\ \hat{y}_{D_i}(t+h) &= f(X_{D_i}(t)) \\ \hat{y}_{A_j}(t+h) &= f(X_{A_j}(t)) \\ \hat{y}(t+h) &= \sum_{i=1}^n \hat{y}_{D_i}(t+h) + \hat{y}_{A_j}(t+h) \end{aligned} \quad (\text{II.10})$$

Où X est la base de données d'origine, X_{D_i} est la $i^{\text{ème}}$ composante détaillée, X_{A_j} est la $j^{\text{ème}}$ composante approximative, h est l'horizon de prédiction, \hat{y} est la valeur prédite et $f(\cdot)$ est le prédicteur. Dans [95], une ondelette Daubechies de nombres 6 a été utilisée pour la décomposition en ondelettes discrètes. Après décomposition, il y a trois décompositions détaillées ($X_{D_i}(t)$, $i = 1, 2, 3$) et une décomposition approximative $X_{A_3}(t)$. ARIMA a été choisi pour être le prédicteur pour chaque sous-série. Les résultats de prévision ont été agrégés pour obtenir le résultat final. Le modèle d'ondelettes-ARIMA présenté a été évalué dans le cas de la prévision de la vitesse du vent de 3,5 et 10 h, et la performance était meilleure que la méthode ARIMA conventionnelle.

Cao and Cao [96] ont présenté un réseau de propagation rétro-ondulante à répétition (RBPN) pour la prédiction de l'irradiation solaire. Le RBPN est un RNA dynamique qui permet de restituer tout ou partie des sorties. Les auteurs ont utilisé la méthode cut-and-trial pour déterminer le nombre de neurones cachés d'un RBPN à 3 couches.

II.7.2.2 Post-traitement

Une série chronologique peut avoir plus d'une caractéristique et chaque caractéristique est adaptée à une méthode particulière. Par exemple, le modèle ARIMA convient à la modélisation de séries chronologiques linéaires et Les RNA sont plus préférables pour la modélisation de séries chronologiques non linéaires. La prédiction de la série chronologique consécutivement par deux prédicteurs ou plus est considérée comme une prévision d'ensemble coopérative basée sur le post-traitement.

Plusieurs modèles coopératifs de prédiction d'ensemble basés sur le post-traitement ont été décrits dans la littérature, tels que ARIMA – GARCH [97], ARIMA – RNA, ARIMA–SVM [98] et SVR-SVC [99].

Dans [100], un modèle de réseau neuronal à retard temporel (TDNN) ARMA a été décrit. Les auteurs ont tout d'abord, utilisé plusieurs méthodes statistiques pour générer une série temporelle d'irradiation solaire stationnaire, puis ils ont utilisé ARMA pour prédire la partie linéaire de la série temporelle et le réseau TDNN pour prédire la partie non linéaire. Ensuite, ils ont combiné les deux résultats pour avoir la prédiction finale. Les auteurs ont comparé le modèle hybride ARMA–TDNN à un seul modèle ARMA et TDNN et les résultats ont montré une RMSE normalisée (NRMSE) plus petite dans le cas du modèle hybride.

II.8 Conclusion

Comme indiqué dans le présent chapitre, de nombreuses méthodes et types de méthodes sont disponibles. Il y a beaucoup de méthodes d'estimation du rayonnement solaire, certaines sont souvent utilisées (réseaux de neurones, processus autorégressifs,...), d'autres commencent à être utilisées (SVM, GP, LSTM...) et d'autres rarement (boosting, arbres de décision, forêts aléatoires, ...etc). En conclusion, on peut dire que les méthodes basées sur les RNA et ARIMA sont équivalentes en termes de qualité de prédiction dans certaines conditions de variabilité, mais la flexibilité des RNA en tant que modèles non linéaires universels les rend plus préférables que les processus autorégressifs classique.

En règle générale, la précision de ces méthodes dépend de la qualité des données d'apprentissage. En fait, en considérant les articles publiés, ces méthodes produisent des statistiques d'erreur très proches.

A partir de la littérature vue dans le présent chapitre, la méthodologie des ensembles a montré toujours une meilleure performance par rapport aux prédicteurs simples (uniques), ce qui montre l'intérêt à les utiliser dans cette thèse.