L'IMPACT DE LA VARIABILITÉ CLIMATIQUE SUR LE SYSTEME ELECTRIQUE FRANÇAIS

MIG PROSPECTUS
Note de synthèse

AULLEN CHOUBRAC Yannis, ARNAUD Oriane, BALBZIOUI Ziad, BONMARCHAND Goulven, CHRISTMANN Raphaëlle, DAURES–BOUVET Eliott, GIUNTA Romain, HOUMAIRE Marie, LAUVERGNE Alexis, MAZINGUE Léna, MRIMI Mouad, NEVES Tiago, TOOFA Keanu

Valentina SESSA, Chargée d'enseignement recherche (CMA)
Damien CORRAL, Ingénieur de recherche (CMA)

Décembre 2025

Table des matières

1 Introduction

A l'heure où la question du climat et de l'énergie s'imposent à nous, le lien entre les deux est de plus en plus étudié et exploré. Nous pouvons en effet désormais trouver des modèles énergétiques qui intègrent la variabilité climatique, ce qui n'était pas le cas il y a quelques années. En particulier, la variabilité climatique est particulièrement intéressante pour le domaine de l'hydroélectricité, qui représente près de 14% de la production française et qui dépend directement des précipitations par exemple.

COMPLETER

C'est notamment le sujet au centre de l'initiative Clim2Power, un projet Européen d'envergure, qui cherche à évaluer la sensibilité à la variabilité climatique des chemins possibles pour atteindre un système électrique neutre en carbone. A partir de données climatiques, des recherches sont menées pour imaginer quel pourrait être le mix énergétique, d'ici 2030-2050. Au cours de ce MIG, nous avons travaillé dans le cadre de ce projet, en répondant à la question suivante : **comment prendre en compte la variabilité climatique dans le système énergétique français**

Dans ce projet, nous nous sommes concentrés sur deux axes principaux :

— Développer un modèle reliant le climat à la production d'électricité hydraulique

— Élaborer un modèle prospectif pour analyser des scénarios à long terme des mix énergétiques possibles

Par ailleurs, on peut noter qu'il existe trois principaux types de centrales hydrauliques : les centrales de lac ou de haute chute et les centrales d'éclusée ou de moyenne chute (hydroelectric dam), caractérisées par une implantation dans des régions montagneuses, un débit faible à moyen et une hauteur de chute moyenne à haute, et les centrales hydraulique au fil de l'eau (Run of River - RoR), caractérisées par une implantation le long de grands fleuves ou grandes rivières, avec un débit très fort et une faible hauteur de chute (moins de 30m). Nous nous sommes concentrés sur les centrales au fil de l'eau. Certains barrages sont aussi dôtés de la technologie de STEP qui permet de réalimenter le bassin d'eau en hauteur à partir du bassin d'eau en sortie de la centrale par pompage. Cette technologie permet d'entretenir la production énergétique. Les centrales à barrage et par STEP sont des énergies de stock qui sont relativement pilotables et souvent utilisée pour les pics de demande énergétique. L'étude du MIG s'intéresse donc aux centrales au fil de l'eau, dont la production varient instantanément en fonction du débit d'eau turbinable dans les rivières. Elles sont donc les plus sensibles à l'influence des variations climatiques.

De manière plus concrète, nous avons donc modélisé,,en étudiant le cycle de l'eau, la puissance électrique que pourront fournir les barrages en France, en connaissant le volume de précipitations ainsi que la température. On réalise ces prédictions via le facteur de capacité (CF), défini ainsi :

$$CF = \frac{P_{produite}}{P_{capacité\ installée}}$$

2 Données et méthodologie

Afin de faire ces diverses prévisions, nous avons choisi de faire usage de l'apprentissage automatique, également connu sous le nom de Machine Learning, ainsi que du Deep Learning. Ces outils sont particulièrement pertinents lorsqu'un modèle physique est difficile à mettre en oeuvre et lorsque les jeux de données à dispositions ne sont pas massifs. Ces deux conditions sont remplies dans notre MIG : les jeux de données sont réduits et fournis à l'échelle nationale par région ; rendant pertinent l'usage de l'apprentissage automatique.

2.1 Données climatiques d'entrées

Avant toute modélisation, pour réaliser un modèle de Machine Learning il est nécessaire de manipuler un set de données et de les trouver. Pour la réalisation de notre projet, les données étaient déjà pré-nettoyées. Les données sont explorées et analysées pour trouver de nouvelles *features* qui permettront d'améliorer le modèle.

Les données climatiques utilisées comprennent des séries temporelles de précipitations et de température de l'air à 2 mètres au-dessus de la surface. Elles présentent une résolution spatiale d'environ 0,25° × 0,25°.

Les données historiques, couvrant la période 2015–2023, sont fournies par le Copernicus Climate Change Service (C3S) [Ref]. Bien que ces données soient disponibles à une résolution temporelle horaire, nous retenons une résolution journalière pour cette étude. Le recours à des facteurs de charge journaliers agrégés est privilégié afin de conserver une plus grande flexibilité pour la gestion et l'optimisation du fonctionnement infra-journalier.

En ce qui concerne la résolution spatiale, les données climatiques sont agrégées au niveau NUTS2 (régions statistiques utilisées pour l'application des politiques régionales [Ref]) ou au niveau national.

Concernant les projections climatiques, des séries temporelles journalières sont extraites pour chaque variable climatique dans le cadre du projet C2P. Les données comprennent des projections futures selon deux voies de concentration représentatives (RCP 4.5 et RCP 8.5), qui sont des scénarios fournissant des séries temporelles des émissions et des concentrations atmosphériques de l'ensemble des gaz à effet de serre, des aérosols, des gaz chimiquement actifs, ainsi que des changements d'occupation des sols.
Le scénario RCP4.5 correspond à une trajectoire intermédiaire de stabilisation dans laquelle le forçage radiatif se stabilise autour de 4,5 W/m^2, tandis que, dans le cas du RCP8.5, le forçage radiatif dépasse 8,5 W/m^2 à l'horizon 2100 et continue d'augmenter par la suite [Ref].

L'étape de **feature engineering** est cruciale pour améliorer les modèles. Elle consiste à manipuler les données bruts pour extraire des caractéristiques (features) afin de renforcer l'apprentissage du modèle. Pour notre problème, ces caractéristiques sont :

— la saisonnalité analysée sur les courbes de température et de Capacity Factors (CF),

— la fonte des neiges au printemps,

— l'accumulation des précipitations.

[COMPLETER]

2.2 Production hydro-électrique

Les données de production hydroélectrique, agrégées à l'échelle nationale, sont issues de la plateforme de transparence de l'ENTSO-E [Ref], qui collecte de manière systématique des données de demande et de production d'électricité à une résolution horaire depuis le 1er janvier 2015. Afin de garantir la cohérence avec la résolution temporelle des données climatiques, nous calculons des valeurs moyennes journalières.

À partir de cette même plateforme, nous extrayons également les données relatives à la capacité hydroélectrique installée annuelle. La combinaison de ces deux sources d'information permet de reconstituer des séries temporelles du facteur de charge, défini comme le rapport entre la production réelle d'électricité hydraulique et la capacité installée, exprimé en pourcentage.

2.3 Méthodologie

Le Machine Learning est une branche de l'Intelligence Artificielle qui permet à un système "d'apprendre", de faire des prédictions et de prendre des décisions, suite à une période d'apprentissage.

L'objectif principal est que le modèle entraîné puisse fonctionner sur des données nouvelles, inconnues, pour prédire correctement (Robustesse). Ici, l'idée est donc d'entraîner le modèle sur les données des années précédentes, pour essayer de prévoir les facteurs de capacité futurs.

Pour garantir cette généralisation, on divise l'ensemble total des données en trois parties distinctes et chronologiques.

— **Training Set** : c'est le plus grand ensemble de données (souvent 60% à 80% du total), utilisé pour ajuster les paramètres (les poids) du modèle.

— **Validation Set** : il est utilisé pour optimiser les hyperparamètres du modèle et permet de trouver la configuration qui donne la meilleure performance avant de voir les données de test.

— **Test Set** : c'est le plus petit ensemble. Il est utilisé une seule fois, à la toute fin du processus, pour obtenir une estimation non biaisée de la performance du modèle sur des données complètement nouvelles.

Dans notre projet, le *training set* et le *validation set* correspondent aux données de 2015 à 2022, et le *test set* correspond aux données de 2023.

2.4 Erreurs

Afin d'optimiser nos modèles, nous essayons de minimiser plusieurs types d'erreurs. On note y le résultat réel, $\hat{y}$ le résultat prédit, et $\bar{y}$ la moyenne des y_i.

— NMAE : Erreur Absolue Moyenne Normalisée

$$\text{NMAE} = \frac{\text{MAE}}{\text{Range}(y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\max(y) - \min(y)}$$

— NRMSE : Racine Carrée de l'Erreur Quadratique Moyenne Normalisée

$$\text{NRMSE} = \frac{\text{RMSE}}{\text{Range}(y)} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\max(y) - \min(y)}$$

— R : Coefficient de Corrélation

$$\text{R} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

— R^2 : mesure la proportion de la variance de la variable dépendante (la cible, ici le facteur de capacité CF) qui est expliquée par les variables indépendantes (nos features, TA et TP) dans le modèle de régression.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2.5 Minimisation et Optimisation

A la vue de ces différentes erreurs, nous avons mis en place plusieurs stratégies pour les réduire.

Sur le traitement des prévision météorologiques, nous avons effectué des moyennes glissantes sur 10 ans, séparé et codé différents scénarios, et supprimé de la modélisation les régions avec une trop faible corrélation entre les données météorologiques et les CF.

Sur le traitement des données lors de l'entrâinement, nous avons ainsi procédé à la normalisation des données, au calcul du biais, au moyennage. Nous avons égalelement mis en place des fonctions rendant compte des saisons (ici les fonctions cosinus et sinus) et le lag (utilisation de la valeur d'une variable dans une région voisine pour prédire ou expliquer la valeur de cette variable dans la région, avec des régions temporelles ou spatiale).

3 Modèles de Machine Learning

3.1 Approche Linéaire

L'approche linéaire consiste à modéliser le système à l'aide de fonctions linéaires : c'est une approche assez simple à implémenter et à concevoir à l'aide de la descente de gradient, mais ce n'est pas la plus performante. En particulier, nous avons décidé d'implémenter la méthode Ridge, dont l'objectif est de réduire la magnitude des coefficients et donc de rendre le modèle moins sensible aux fluctuations. Cela réduit la variance ie l'erreur quadratique, ce qui se traduit par une diminution des tendances d'overfitting.

En mettant ensuite en place les différentes stratégies,avec un lag de 15 jours sur les températures, nous avons pu obtenir $R^2 = 0.58$ et $nMae = 0.14$.

Et, si ces résultats ne sont pas satisfaisants en soit, ils montrent qu'il est possible avec un bon traitement des données, d'améliorer significativement le modèle, puisque les valeurs de départ pour le R^2 ne dépassaient pas 0.1.

3.2 Random Forest

Après cette technique assez simple, nous nous sommes penchés sur des techniques plus complexes, basées sur les arbres de décisions. La première, *Random Forest*est une technique basée sur les arbres de récursion, qui améliore les arbres baggués pour construire des arbres décorrélés.

En rajoutant les lags, décorrelant les régions et appliquons les méthodes sur les traitements de données habituels, cette technique nous a permis d'obtenir $R^2 = 0.66$, ce qui est mieux que le Ridge mais reste encore à améliorer.

3.3 Gradient Boosting

Le *Gradient Boosting* est également une technique utilisant les arbres et la récursion. Néanmoins, contrairement au *Random Forest* qui construit ses arbres en parallèle et de manière indépendante, le *Gradient Boosting* construit ses arbres de manière séquentielle et additive. Chacun de ces arbres peut être relativement petit et, en général, cela améliorer progressivement la précision du modèle.

De même, avec un traitement des données en amont et un entraînement ciblé, nous avons pu obtenir $R^2 = 0.66$

4 Deep Learning

Nous nous sommes ensuite penchés sur une approche différente et plus complexe : le Deep Learning. C'est un autre sous-domaine de l'apprentissage automatique qui utilise des réseaux de neurones artificiels composés de nombreuses couches pour analyser des données et prendre des décisions.

En particulier, nous avons utilisé le modèle LSTM (Long Short-Term Memory, soit Mémoire à Long Terme et à Court Terme), qui est un type de réseau de neurones récurrents (RNN) particulièrement adapté pour traiter et prédire les séries temporelles et les séquences de données.

Les résultats obtenus sont les suivants :

Pour 2022 :

Pour 2023 :

Ces résultats sont bien meilleurs que les précédents.

R2: 0.673
NRMSE: 0.176
NMAE: 0.143
R:0.832

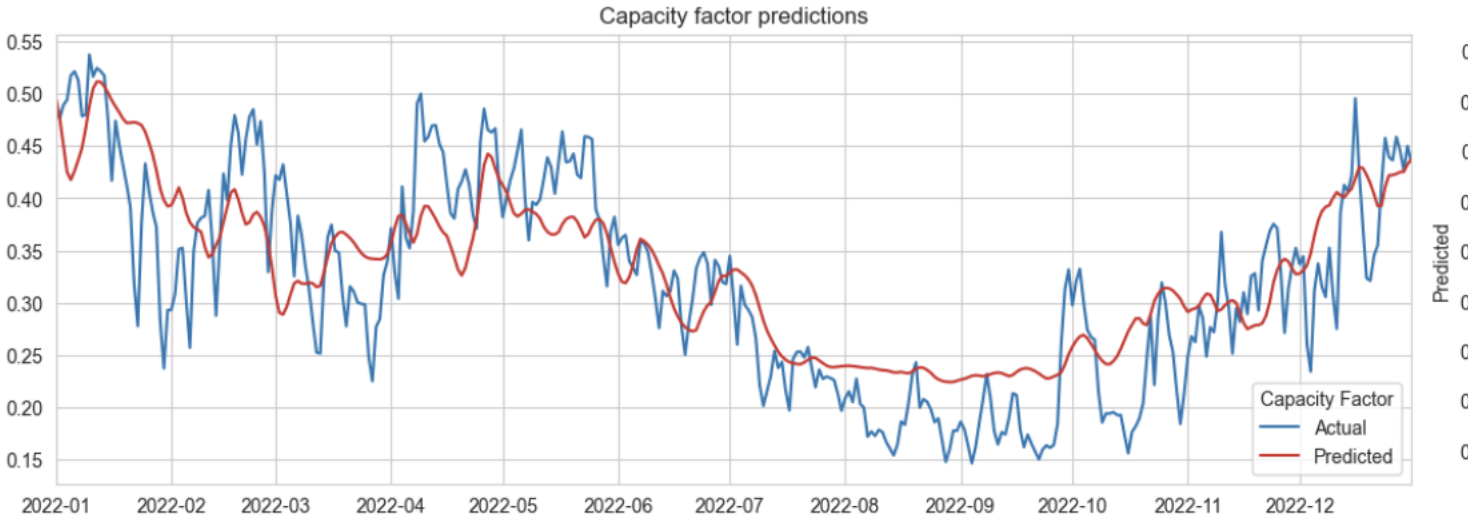

FIGURE 1 – Capacity Factors 2022

R2: 0.713
NRMSE: 0.153
NMAE: 0.117
R:0.853

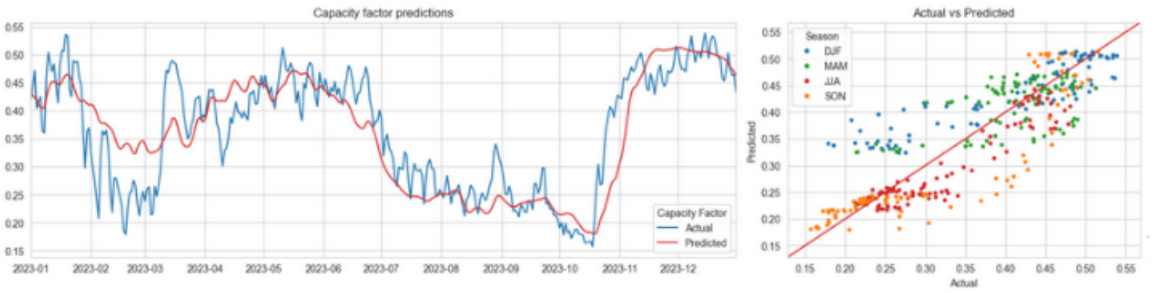

FIGURE 2 – Capacity Factors 2023

5 Comparaison des Résultats

Modèle	R^2	nMAE	nRMSE	R
Ridge Regression	0.58	0.14	0.18	0.76
Random Forest	0.66	0.16	0.19	0.83
Gradient Boosting	0.66	0.13	0.17	0.82
LSTM	0.71	0.11	0.15	0.85

TABLE 1 – Tableau récapitulatif des meilleurs résultats par modèle

6 Prospective

7 Conclusion

Dans le cadre de ce projet, nous avons ainsi pu à partir des données de températures et de precipitations en France entre 2015 et 2023, créer différents modèles d'apprentissage automatique afin de prédire les facteurs de capacités des centrales hydrauliques au fil de l'eau. Afin d'améliorer la précision de nos modèles, un premier travail de préparation des données a ainsi été effectué pour représenter les dépendances d'une centrale hydraulique aux grandeurs climatiques. Nous avons alors pu entraîner nos différents modèles de machine learning et de deep learning sur ces données. Après comparaison de ces modèles à partir de différentes erreurs et des courbes confrontant les facteurs de capacité réels et prédits, le modèle LSTM a été déterminé comme étant le plus efficace. Ce modèle a donc été choisi pour réaliser une étude prospective à horizon 2050. [Parler de la prospective].