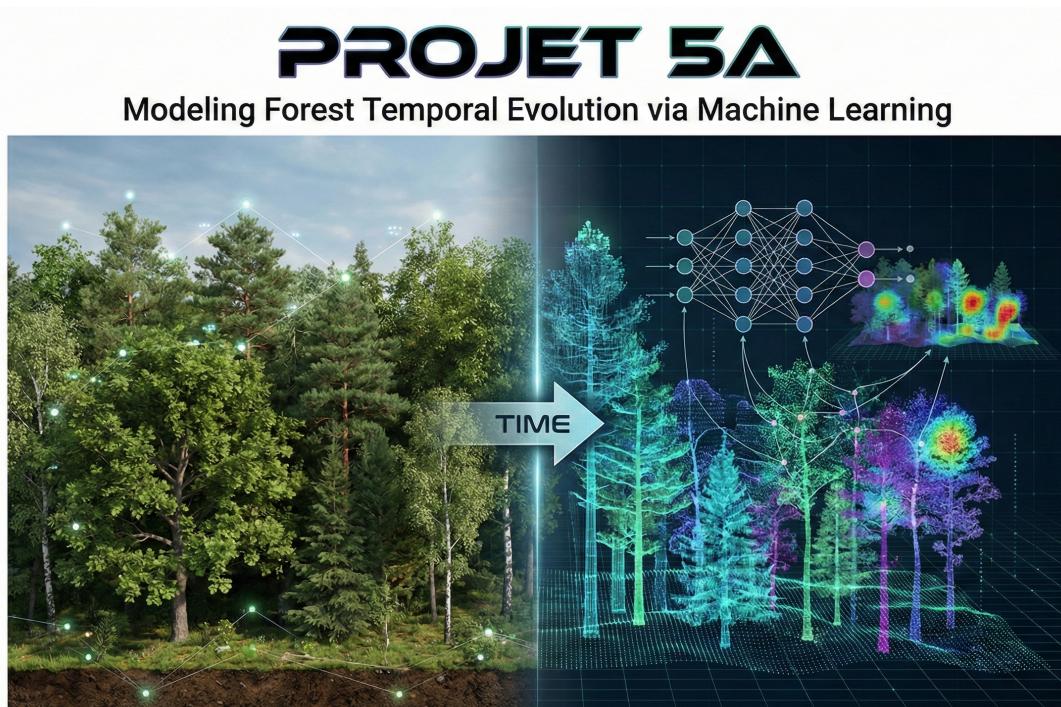




POLYTECH CLERMONT-FERRAND

DÉPARTEMENT INGÉNIERIE MATHÉMATIQUE ET DATA SCIENCE (IMDS)

Modélisation de l'évolution temporelle des espèces forestières par apprentissage automatique



Auteurs :

Ayman ZEJLI
Loïc MAGNAN

Tuteur :

Chafik SAMIR

Année Universitaire 2025 - 2026

Table des matières

Liste des Figures	2
Remerciements	3
Résumé / Abstract	4
Introduction Générale	5
1 Approche de modélisation et fondamentaux	6
1.1 Formalisme théorique des Processus Gaussiens	6
1.2 Étude expérimentale et comparaison de modèles en dimension 1	6
1.2.1 Analyse sur fonction périodique simple : $y = \cos(x)$	6
1.2.2 Analyse sur fonction complexe : $y = \cos(x) + x^2 - 20 \sin(5x)$	7
1.2.3 Analyse sur fonction de haute complexité : $y = \ln(x) + 4x \cos(x^2)$	9
1.2.4 Synthèse et bilan de l'étude 1D	11
1.3 Modélisation spatiale et extension en dimension 2	13
1.3.1 Structure et Approche du Réseau de Neurones Convolutif (CNN 2D)	13
1.3.2 Analyse sur surface périodique simple : $f_1(x, y) = \cos(2\pi(x+y))$	13
1.3.3 Analyse sur surface à motifs périodiques : $f_2(x, y) = \sin(2\pi x) \cdot \cos(2\pi y)$	14
1.3.4 Analyse sur surface complexe et évaluation de la robustesse prédictive : $f_3(x, y) = \sin(3\pi x) \cos(2\pi y) + e^{-5((x-0.5)^2+(y-0.5)^2)}$	15
1.3.5 Synthèse comparative des performances en dimension 2	16
2 Application aux écosystèmes forestiers : Étude de cas réelle	18
2.1 Acquisition et description du jeu de données ERA5-Land	18
2.2 Ingénierie des données et visualisations cartographiques	20
2.2.1 Écosystème logiciel pour l'analyse spatiale	20
2.2.2 Cartes de chaleur et gradients de température	21
2.2.3 Représentation des variables SKT et SKT_C (Année 2000)	22
2.2.4 Analyse de la moyenne min-max par coordonnée	23
2.2.5 Gradient de température : Évolution 2000 vs 2007	24
2.3 Le paradigme LSTM : Mémoire et apprentissage séquentiel	26
2.3.1 Protocole d'évaluation et stratégie de confrontation spatio-temporelle	26
2.3.2 Analyse des résultats : Est-ce que l'IA arrive à prédire l'année 2008 ?	27
2.3.3 Analyse dynamique des tendances temporelles et structures 3D	29
2.3.4 Cartographie des résidus et analyse spatio-temporelle des erreurs	30
2.4 Visualisation cartographique spatio-temporelle grâce aux Heatmaps	33
2.4.1 Architecture et préparation des données spatiales	33
2.4.2 Analyse comparative spatiale : Réalité vs Prédictions (Juillet 2008)	34
2.5 Synthèse globale de l'étude : De l'acquisition climatique à la décision forestière	37
Conclusion Générale	38
Bibliographie	39

Table des figures

1	Prédiction d'un processus gaussien 1D pour l'approximation de la fonction cosinus sur l'intervalle $[0, 10]$	7
2	Approximation de la fonction cosinus à l'aide d'un réseau de neurones (300 époques, batch size 8).	7
3	Comparaison des prédictions des kernels GP sur $y = \cos(x) + x^2 - 20\sin(5x)$. .	8
4	Prédictions d'un réseau de neurones sur $y = \cos(x) + x^2 - 20\sin(5x)$	9
5	Analyse de l'influence des kernels GP sur une fonction à fréquence variable. On observe que seul le noyau RationalQuadratic (en bas à gauche) parvient à suivre l'accélération du signal.	10
6	Prédiction du réseau de neurones.	10
7	Reconstruction par Processus Gaussien de la fonction f_1	14
8	Approximation par CNN 2D de la fonction f_1	14
9	Prédition GP capturant les extrema de la fonction f_2	15
10	Approximation par CNN 2D de la structure complexe f_2	15
11	Reconstruction GP isolant le pic central de la fonction f_3	16
12	Approximation CNN 2D de la fonction hybride f_3	16
13	Portail du Climate Data Store (CDS) permettant l'accès aux réanalyses climatiques ERA5-Land.	18
14	Interface de configuration de la requête sur le Climate Data Store montrant la sélection des variables et de la période temporelle.	19
15	Aperçu de la structure tabulaire des données climatiques extraites (Fichier CSV). .	19
16	Délimitation géographique de la zone d'étude comprenant les départements de l'Allier (03), du Cantal (15), de la Haute-Loire (43) et du Puy-de-Dôme (63). . .	21
17	Représentation de la variable SKT (Kelvin) sur la région Auvergne en 2000. . .	22
18	Représentation de la variable SKT_C (Celsius) sur la région Auvergne en 2000. .	22
19	Carte de la moyenne annuelle min-max pour SKT_C en 2000 avec délimitation départementale.	23
20	Gradient thermique SKT_C (Δ 2007 - 2000) en Auvergne.	24
21	Comparaison des modèles sur les 12 mois de 2008, l'année prédictive.	29
22	Visualisation 3D (Lon, Lat, Temp) de la réalité 2008 vs la prédition GP et l'erreur entre les deux.	30
23	Cartographie mensuelle des erreurs du modèle LSTM pour l'année 2008. . . .	30
24	Cartographie mensuelle des erreurs du modèle Gaussian Process (GP).	31
25	Architecture logicielle : Organisation hiérarchique des répertoires pour le stockage automatisé des heatmaps interactives, incluant l'historique climatique (2000–2007) et la comparaison prédictions vs données réelles pour l'année 2008. .	33
26	Exemple de Heatmap mensuelle générée via Folium	33
27	Réalité terrain ERA5-Land pour juillet 2008. La palette de couleurs illustre le gradient thermique naturel lié au relief auvergnat.	35
28	Prédition IA par réseau LSTM. On note la conservation des structures spatiales fines et la précision des gradients départementaux.	35
29	Prédition par Processus Gaussien.	36

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à mon tuteur de projet, **Monsieur Chafik Samir**, pour son encadrement de grande qualité, sa disponibilité constante et la pertinence de ses conseils techniques tout au long de ce projet tutoré. Ses orientations stratégiques et son expertise ont été essentielles pour mener à bien la modélisation spatio-temporelle complexe au cœur de cette étude.

Mes remerciements s'adressent également à l'école **Polytech Clermont** ainsi qu'à l'ensemble du corps enseignant de la filière IMDS. Je salue particulièrement la qualité des enseignements prodigués, tant théoriques que pratiques, qui m'ont fourni les compétences nécessaires en science des données et en systèmes d'information pour relever les défis de ce sujet.

Enfin, je remercie les responsables de la formation pour avoir mis en place un environnement de travail propice à l'apprentissage et à l'innovation, permettant ainsi la concrétisation de ce projet de fin de cursus dans les meilleures conditions.

Résumé

Ce projet tutoré, réalisé dans le cadre de la cinquième année à **Polytech Clermont**, propose une méthodologie avancée pour la modélisation spatio-temporelle de l'évolution des écosystèmes forestiers face aux pressions du changement climatique. En exploitant la puissance du langage **Python** et de bibliothèques spécialisées en traitement de données massives, nous avons extrait et structuré les données climatiques afin d'analyser la dynamique thermique des sols.

L'étude repose sur une confrontation rigoureuse entre des approches statistiques classiques par **Processus Gaussiens** et la capacité de mémorisation séquentielle des réseaux de neurones **LSTM**. Les évaluations démontrent la supériorité des architectures de **Deep Learning** : tandis que le modèle statistique peine à capturer les cycles saisonniers complexes, l'architecture récurrente parvient à modéliser les variations thermiques avec une grande fidélité. Ce travail aboutit à la production de **cartographies interactives**, offrant un outil de visualisation indispensable pour identifier les zones de stress environnemental critique.

Mots-clés : Intelligence Artificielle, Apprentissage Profond, Apprentissage Automatique, Science des Données, Systèmes d'Information Géographique (SIG), Modélisation Environnementale, Processus Gaussiens (GP), ERA5-Land.

Abstract

This tutored project, conducted as part of the fifth-year curriculum at **Polytech Clermont**, establishes a sophisticated framework for the spatio-temporal modeling of forest ecosystems under climate change. Leveraging **Python** and high-performance data processing libraries, we structured large-scale climate data to monitor environmental dynamics under thermal stress.

The core of the study involves a comparative analysis between traditional statistical methods and the sequential memory of **LSTM** neural networks. Numerical evaluations highlight the significant advantage of **Deep Learning** architectures : while statistical models struggle with seasonal extrapolation, the recurrent neural network demonstrates high predictive accuracy and strong alignment with observed climate patterns. The project concludes with the generation of **interactive heatmaps**, providing essential granular visualization tools to detect and monitor environmental anomalies in forest plots.

Keywords : Artificial Intelligence, Deep Learning, Machine Learning, Data Science, Geographic Information Systems (GIS), Environmental Modeling, Gaussian Processes (GP), ERA5-Land.

Introduction Générale

Le présent rapport documente les travaux de recherche et de développement menés dans le cadre du **Projet Tutoré de cinquième année (5A)** au sein de l'école d'ingénieurs **Polytech Clermont**. Ce projet s'inscrit dans une problématique environnementale et technologique de premier plan : la modélisation de l'évolution temporelle des espèces forestières par apprentissage automatique. Face à l'urgence climatique, la compréhension de la dynamique des forêts est devenue un enjeu stratégique pour la préservation de la biodiversité et la gestion durable des ressources naturelles.

Les écosystèmes forestiers sont aujourd'hui soumis à des pressions sans précédent. La hausse globale des températures, la récurrence des épisodes de sécheresse intense et l'évolution démographique des peuplements (âge des arbres) modifient profondément la composition des forêts. Dans ce contexte, la capacité à prédire les changements de répartition des espèces n'est plus seulement un défi académique, mais une nécessité opérationnelle. Ce projet propose de relever ce défi en exploitant la synergie entre l'**Intelligence Artificielle (IA)** et les **Systèmes d'Information Géographique (SIG)**, afin de transformer des données de terrain complexes en modèles prédictifs robustes.

La démarche scientifique adoptée repose sur l'analyse de données spatio-temporelles massives, notamment issues du jeu de données **ERA5-Land** [1]. La complexité de l'étude réside dans la nécessité de capturer à la fois des motifs temporels cycliques (saisonnalité) et des variations spatiales fines liées à la topographie locale. Pour ce faire, nous avons mis en concurrence deux approches technologiques majeures : les **Processus Gaussiens (GP)** [2], pour leur excellence en interpolation spatiale, et les réseaux de neurones récurrents **LSTM** [3], pour leur capacité unique à mémoriser les dépendances temporelles sur le long terme.

L'objectif final de ce travail est de fournir une preuve de concept capable de caractériser les variations d'espèces sur une période définie et d'en évaluer la précision par rapport à la réalité terrain. L'ensemble du développement a été réalisé sous VS Code [4] via des Jupyter Notebooks [5], et le code source est archivé sur GitHub [6]. Pour une meilleure lisibilité, le présent rapport est structuré selon le plan suivant :

Annonce du plan : Le premier chapitre est consacré à la **préparation et à la visualisation des données**, détaillant les étapes de nettoyage et d'organisation nécessaires à l'entraînement de l'IA. Le deuxième chapitre traite de l'**entraînement et de l'évaluation des modèles**, où nous comparons les performances intrinsèques du GP et du LSTM sur des signaux théoriques et réels. Le troisième chapitre présente la **phase de prédiction opérationnelle** sur l'année 2008 et l'**évaluation de la précision** via la génération de heatmaps interactives. Enfin, une conclusion générale synthétisera les résultats obtenus et proposera des perspectives pour l'aide à la décision forestière en Auvergne.

1 Approche de modélisation et fondamentaux

1.1 Formalisme théorique des Processus Gaussiens

Le processus gaussien (GP) constitue une généralisation du concept de loi normale appliquée aux fonctions continues. Alors qu'une loi normale classique décrit la distribution d'un scalaire ou d'un vecteur aléatoire fini, le processus gaussien définit une distribution de probabilité sur un ensemble de fonctions. Cette approche bayésienne non paramétrique permet de modéliser la forme probable d'une fonction tout en quantifiant précisément l'incertitude associée à chaque prédiction.

Un processus gaussien est intégralement caractérisé par deux fonctions fondamentales :

- **La fonction moyenne** $m(x)$, qui définit l'espérance de la fonction en tout point : $m(x) = \mathbb{E}[f(x)]$. Dans la pratique, on suppose souvent $m(x) = 0$ par défaut.
- **La fonction de covariance** $k(x, x')$, ou noyau (*Kernel*), qui modélise la dépendance statistique entre deux points : $k(x, x') = \text{Cov}(f(x), f(x'))$.

La relation formelle s'écrit alors :

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Pour tout ensemble fini de points d'entrée $X = [x_1, \dots, x_n]$, le vecteur aléatoire correspondant $f = [f(x_1), \dots, f(x_n)]^T$ suit une **loi normale multivariée** :

$$f \sim \mathcal{N}(m(X), K(X, X))$$

où K est la matrice de Gram telle que $K_{ij} = k(x_i, x_j)$. Cette structure permet au modèle de calculer une distribution *a posteriori*, offrant une prédiction ponctuelle accompagnée d'une variance prédictive.

1.2 Étude expérimentale et comparaison de modèles en dimension 1

1.2.1 Analyse sur fonction périodique simple : $y = \cos(x)$

Afin d'illustrer la capacité des processus gaussiens à reconstruire une fonction continue à partir d'un nombre limité d'observations, nous avons utilisé un modèle en une dimension pour approximer la fonction $f(x) = \cos(x)$. Pour cette approche, le modèle a été entraîné sur seulement 20 points choisis aléatoirement parmi 100 points générés uniformément dans l'intervalle $[0, 10]$. Les prédictions ont ensuite été effectuées sur l'ensemble des 100 points pour évaluer la qualité de l'interpolation par rapport au signal originel.

En parallèle, un réseau de neurones de type **MLP (Multi-Layer Perceptron)**, composé de deux couches cachées de 64 neurones chacune, a été testé. L'entraînement a été réalisé sur 100 points générés aléatoirement durant 300 époques avec un *batch size* de 8 afin d'évaluer sa capacité à généraliser la fonction cosinus sans connaissance *a priori* de sa régularité".

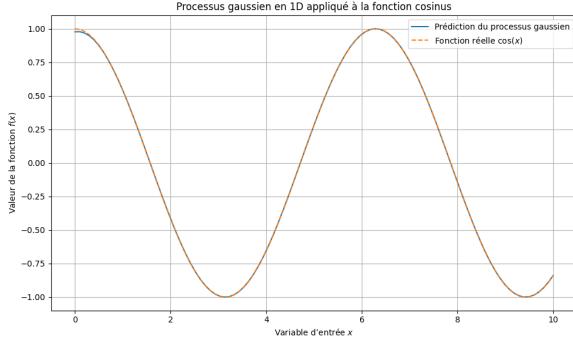


FIGURE 1 – Pr  diction d'un processus gaussien 1D pour l'approximation de la fonction cosinus sur l'intervalle [0, 10].

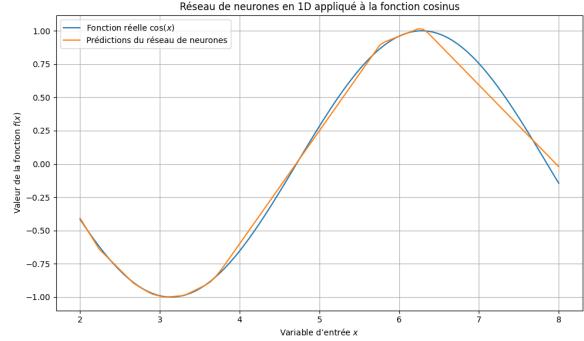


FIGURE 2 – Approximation de la fonction cosinus     l'aide d'un r   seau de neurones (300    poques, batch size 8).

L'interpr  tation visuelle des r  sultats met en   vidence une sup  riorit  notable du mod  le probabiliste dans ce sc  nario de donn  es restreintes. Comme l'illustre la **Figure 1**, le processus gaussien parvient   reconstruire la fonction de mani  re quasi-parfaite malgr  le faible nombre de points d'entra  nement (20 points). L'aspect lisse de la courbe d  montre que le noyau a correctement captur  la structure p  riodeuse du signal sur l'intervalle   tudi  .

À l'inverse, la **Figure 2** r  v  le les limites de l'approche par r   seau de neurones face   cet  chantillonnage. Bien que la tendance globale soit respect  e, la courbe pr  sente des irr  gularit  es locales. Contrairement au GP qui interpose par nature via sa structure de covariance, le r   seau de neurones approxime la fonction par optimisation de poids. Sans une densit  de points tr  s  lev  e, il peine    galer la fluidit  du processus gaussien, confirmant la robustesse de ce dernier pour des signaux p  riodeux lisses avec peu de donn  es.

1.2.2 Analyse sur fonction complexe : $y = \cos(x) + x^2 - 20 \sin(5x)$

Pour cette seconde phase exp  rimentale, nous testons la capacit  des mod  les   g n raliser une structure complexe   partir d'un  chantillonnage partiel. Le protocole consiste    electionner al atoirement n_{train} points parmi un ensemble de N points g n r  s, puis   utiliser ces mod  les pour pr  dier la fonction sur l'int  gralit  du domaine. L'objectif est d'   valuer la pr  cision de la reconstruction via l'**Erreur Quadratique Moyenne (MSE)**.

Dans le cadre des Processus Gaussiens, le choix du noyau (*kernel*) est d terminant car il d finit les corr  lations a priori entre les points. Nous avons test  quatre types de noyaux pour capturer les diff  rentes composantes de notre fonction cible :

- **Radial Basis Function (RBF)** : Le noyau par d faut, supposant une fluidit  infinie.

$$k(x, x') = \exp\left(-\frac{d(x, x')^2}{2l^2}\right)$$

- **Rational Quadratic** : Id al pour des donn  es variant   plusieurs  chelles.

$$k(x, x') = \left(1 + \frac{d(x, x')^2}{2\alpha l^2}\right)^{-\alpha}$$

- **Exp-Sine-Squared** : Con u pour capturer des p  riodicit  es strictes.

$$k(x, x') = \exp\left(-\frac{2 \sin^2(\pi d(x, x')/p)}{l^2}\right)$$

- **Matérn** : Une généralisation du RBF permettant de modéliser des fonctions moins lisses.

$$k(x, x') = \frac{1}{\Gamma(v) 2^{v-1}} \left(\frac{\sqrt{2v}}{l} d(x, x') \right)^v K_v \left(\frac{\sqrt{2v}}{l} d(x, x') \right)$$

Les résultats de ces tests montrent que la flexibilité du noyau **Rational Quadratic** ou l'utilisation du **Matérn** permettent une meilleure adaptation aux oscillations rapides induites par le terme $-20 \sin(5x)$ que le RBF classique. La **Figure 3** présente la prédiction optimale obtenue via le Processus Gaussien.

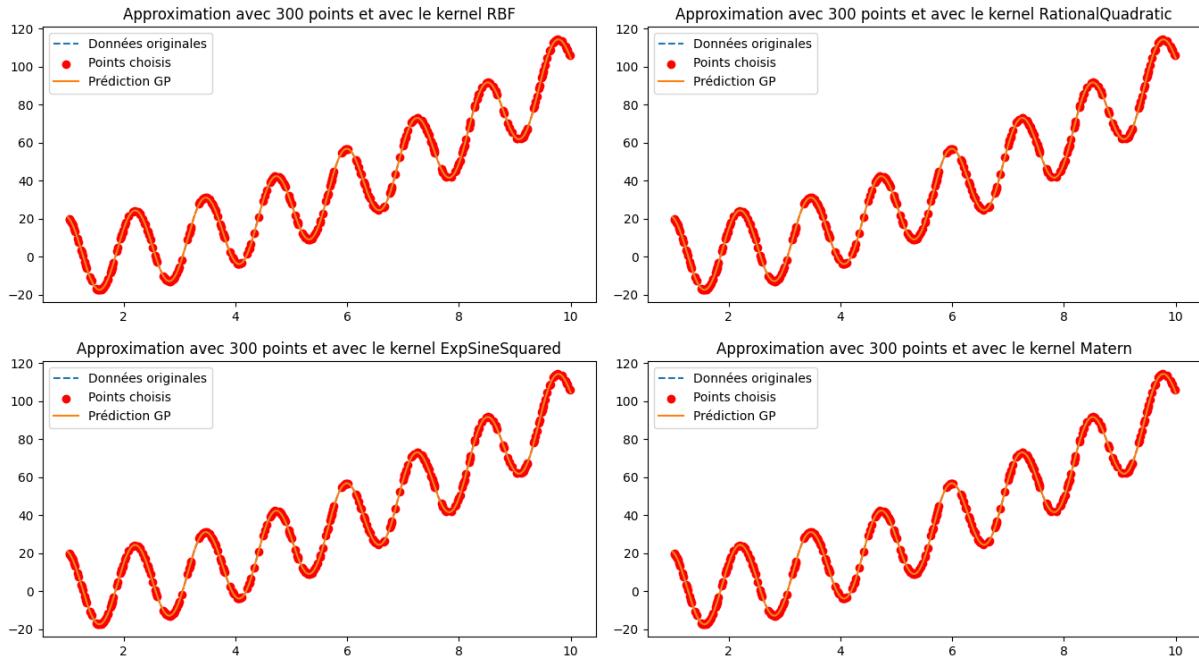


FIGURE 3 – Comparaison des prédictions des kernels GP sur $y = \cos(x) + x^2 - 20 \sin(5x)$.

À l'inverse, l'approche par **Réseau de Neurones** (MLP) aborde ce problème sans hypothèse géométrique préalable. Le modèle doit apprendre la tendance quadratique et les oscillations simultanément. Pour cette fonction complexe, une étape de **normalisation** a été indispensable (Min-Max Scaling sur x , standardisation sur y) pour stabiliser la descente de gradient via l'optimiseur **Adam**.

L'architecture retenue est un réseau profond composé de trois couches denses de respectivement **256 neurones**. L'utilisation de la fonction d'activation **tanh** permet de capturer les non-linéarités. Bien que le modèle minimise l'erreur sur l'ensemble des points, il nécessite un entraînement intensif de **300 époques** avec un **batch size de 128**.

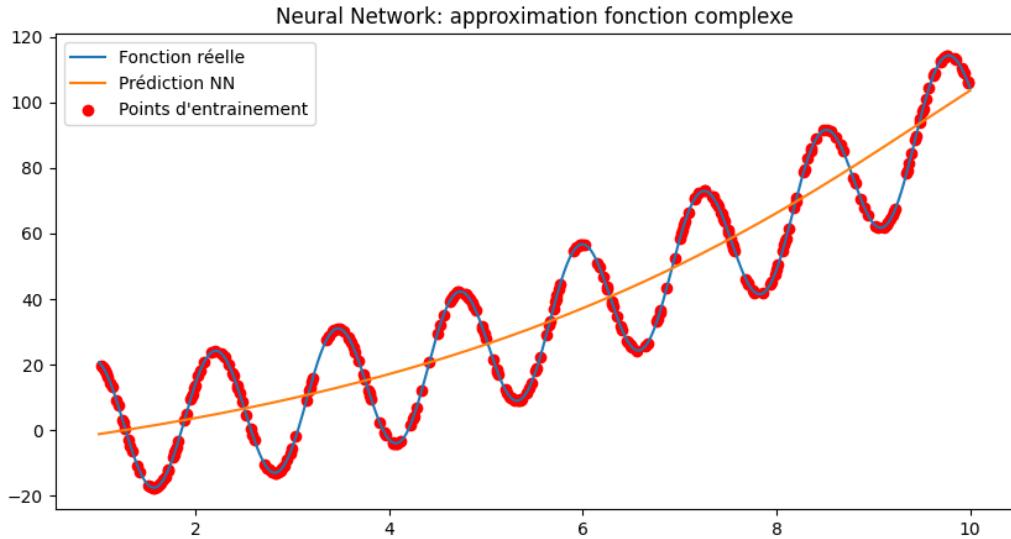


FIGURE 4 – Prédictions d’un réseau de neurones sur $y = \cos(x) + x^2 - 20 \sin(5x)$.

L’interprétation de la **Figure 4** met en lumière une limite structurelle du réseau de neurones face à des signaux multi-échelles. On constate que la prédiction épouse parfaitement la tendance de fond parabolique. Cependant, elle lisse presque intégralement les oscillations à haute fréquence induites par le terme sinusoïdal. En agissant comme un **filtre passe-bas**, le réseau privilégie la minimisation de l’erreur globale sur la composante de forte amplitude, négligeant la structure locale pourtant présente. À l’inverse, le **Processus Gaussien**, grâce à un noyau adapté, parvient à conserver cette fidélité aux variations rapides.

1.2.3 Analyse sur fonction de haute complexité : $y = \ln(x) + 4x \cos(x^2)$

Pour pousser l’évaluation à un niveau critique, nous introduisons une fonction présentant une fréquence non constante et une croissance logarithmique : $y = \ln(x) + 4x \cos(x^2)$. Ce signal est particulièrement difficile à modéliser car la densité et l’amplitude des oscillations augmentent de manière quadratique avec x , rendant l’interpolation extrêmement sensible au choix du modèle.

Étude comparative des noyaux (GP) : La **Figure 5** illustre la sensibilité du Processus Gaussien face à ce signal non stationnaire. Le noyau *ExpSineSquared* échoue totalement à capturer la dynamique car il suppose une périodicité fixe, alors que la fréquence ici s’accélère. Le noyau *RBF*, trop rigide, lisse les oscillations dès que la fréquence augmente. À l’inverse, le noyau **RationalQuadratic** offre la meilleure reconstruction visuelle, s’adaptant aux différentes échelles de variation. Le noyau **Matern** ($\nu = 1.5$) capture bien les pics mais montre des signes de saturation sur les amplitudes extrêmes en fin d’intervalle.

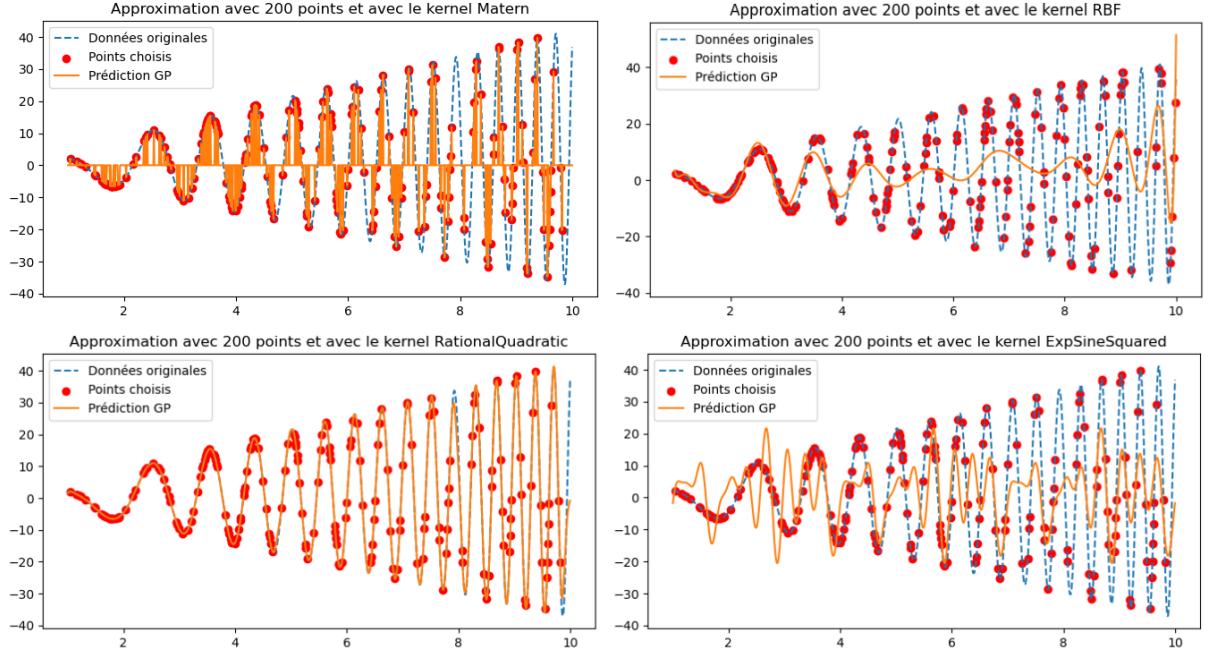


FIGURE 5 – Analyse de l'influence des kernels GP sur une fonction à fréquence variable. On observe que seul le noyau RationalQuadratic (en bas à gauche) parvient à suivre l'accélération du signal.

Évaluation du Réseau de Neurones (NN) : En parallèle, le modèle ReLU implémenté (64 neurones) a été testé sur ce même signal. Comme le montre la **Figure 6**, le réseau est capable de saisir la tendance logarithmique ascendante (la ligne moyenne du signal), mais il est totalement incapable de modéliser les oscillations. L'utilisation de l'activation **ReLU**, associée à une architecture légère, contraint le modèle à une approximation trop simpliste qui ignore les composantes haute fréquence.

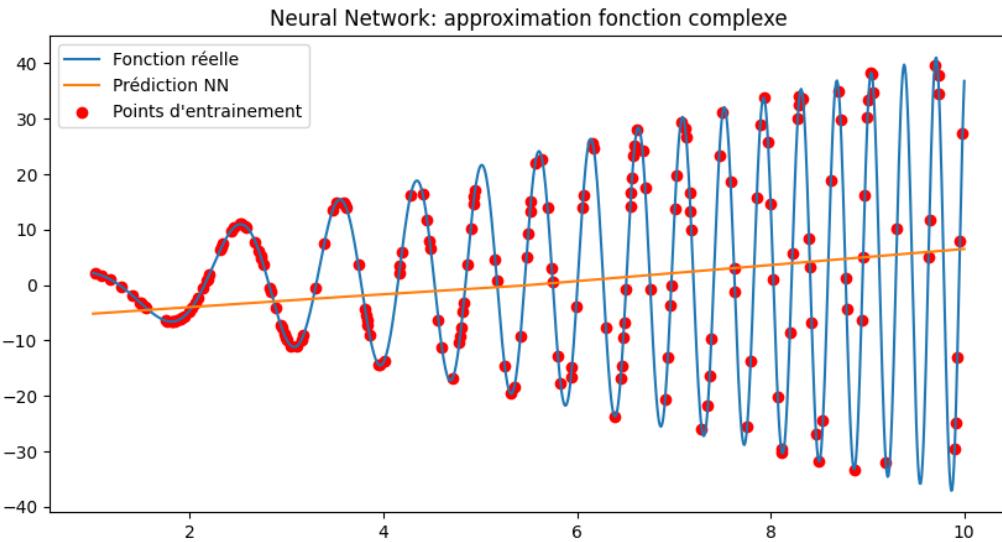


FIGURE 6 – Prédiction du réseau de neurones.

La comparaison visuelle est sans appel : le GP avec un noyau **RationalQuadratic** (Figure 5) surpasse largement le réseau de neurones (Figure 6) en termes de fidélité au signal. Bien que le NN soit beaucoup plus rapide à s'exécuter sur de grands jeux de données, il agit ici comme un simple lisseur de tendance. Pour la modélisation forestière, où les micro-variations temporelles sont essentielles, le GP reste l'outil de référence, malgré un coût de calcul plus élevé lié à l'inversion de la matrice de covariance du kernel.

En conclusion, sur des signaux à fréquence variable et amplitude croissante, la flexibilité du GP permet une reconstruction structurelle là où le NN ReLU standard échoue par manque de profondeur ou de points d'entraînement massifs.

1.2.4 Synthèse et bilan de l'étude 1D

Après avoir analysé individuellement chaque fonction cible, nous proposons ici une vue d'ensemble des performances comparées. Cette synthèse permet de quantifier l'efficacité des modèles en fonction de la complexité du signal et de la densité de l'échantillonnage, simulant ainsi les contraintes réelles de collecte de données sur le terrain.

Premier scénario : Échantillonnage faible (20 points)

Ce premier scénario simule des situations critiques où les relevés de terrain sont rares ou coûteux à obtenir. Nous avons d'abord testé la reconstruction d'un signal périodique simple ($y = \cos(x)$) pour évaluer la capacité d'interpolation de base. Le modèle de réseau de neurones utilisé est un MLP composé de deux couches cachées de 64 neurones, entraîné sur 300 époques. L'implémentation du modèle de régression par Processus Gaussien repose sur la bibliothèque Scikit-learn [7, 2]. Les résultats montrent que le Processus Gaussien surpasse structurellement le réseau de neurones, ce dernier peinant à lisser parfaitement la courbe sans une densité de points plus élevée.

Erreurs quadratiques et des temps de calcul pour $y = \cos(x)$ (20 pts)

MODÈLE	TEMPS (S)	MSE	PARAM.
GP (RatQuad)	0.0284	7.3×10^{-12}	2
GP (ExpSine)	0.0173	7.3×10^{-12}	2
NN (MLP)	3.004	0.149	4353

Pour complexifier l'évaluation, nous avons introduit un signal multi-échelle intégrant une tendance quadratique et des oscillations rapides. Dans ce contexte, le choix du noyau (*kernel*) pour le GP devient déterminant : les noyaux Rational Quadratic et Matérn offrent une flexibilité supérieure au RBF classique pour suivre les variations brusques. Pour le réseau de neurones, une architecture plus profonde (3 couches de 256 neurones) et une normalisation des données (*Min-Max Scaling*) ont été nécessaires pour capturer la dynamique. On observe toutefois que le réseau agit comme un filtre passe-bas, privilégiant la tendance de fond parabolique au détriment des micro-oscillations locales.

Résultats sur $y = \cos(x) + x^2 - 20 \sin(5x)$ (20 pts)

MODÈLE	TEMPS (S)	MSE	PARAM.
GP (RatQuad)	0.0058	146.9	2
GP (RBF)	0.0008	101.1	0
NN (MLP)	2.961	217.3	4353

Second scénario : Échantillonnage dense (200 points)

Le second scénario explore les limites des modèles face à une fonction de haute complexité présentant une fréquence variable et une croissance logarithmique. Avec un échantillonnage plus dense de 200 points, le réseau de neurones ReLU parvient à identifier la tendance ascendante mais échoue systématiquement à modéliser les oscillations dont la fréquence s'accélère. À l'inverse, le Processus Gaussien avec un noyau Rational Quadratic démontre une capacité d'adaptation exceptionnelle, parvenant à suivre l'accélération du signal là où les modèles rigides échouent. Cette étape confirme que, même avec plus de données, la structure probabiliste du GP reste plus fidèle aux réalités physiques complexes que les réseaux de neurones standards.

Résultats sur $y = \ln(x) + \cos(x^2) \times 4x$ (200 pts)

MODÈLE	TEMPS (S)	MSE	PARAM.
GP (RatQuad)	0.188	5.49	2
GP (Matern)	0.0054	245.8	1
NN (MLP)	3.383	295.3	4353

D'après ces observations, on peut conclure que, pour la prédiction de fonctions, le processus gaussien se révèle globalement plus performant qu'un réseau de neurones et s'impose comme l'outil de référence pour la modélisation forestière. En effet, sa rapidité d'exécution, sa précision (MSE) et son très faible nombre de paramètres surclassent les réseaux de neurones, particulièrement pour les signaux non stationnaires. S'il serait certes possible d'augmenter la complexité du réseau de neurones en ajoutant des couches ou davantage de neurones pour tenter d'améliorer ses scores, cela entraînerait inévitablement un temps de calcul prohibitif et un risque accru de sur-apprentissage, pour une précision souvent moins satisfaisante et un nombre de paramètres nettement plus élevé que celui d'un processus gaussien.

Transition vers la 2D : Bien que les Processus Gaussiens dominent en 1D, l'analyse forestière nécessite de prendre en compte plusieurs variables simultanées (coordonnées spatiales, température). Dans la section suivante, nous allons évaluer si cette supériorité se maintient lors du passage à des dimensions supérieures.

1.3 Modélisation spatiale et extension en dimension 2

Le passage à la dimension 2 est crucial pour notre étude forestière, car il permet de simuler la répartition spatiale des espèces sur une parcelle définie par des coordonnées (x, y) . Dans cette section, nous évaluons la capacité des modèles à reconstruire des surfaces continues à partir d'un échantillonnage aléatoire réparti sur un domaine unitaire $[0, 1] \times [0, 1]$.

1.3.1 Structure et Approche du Réseau de Neurones Convolutif (CNN 2D)

Un **Réseau de Neurones Convolutif (CNN)** se distingue des architectures classiques par sa capacité à préserver la **topologie des données** d'entrée en extrayant des caractéristiques locales de manière hiérarchique. Pour l'approximation de fonctions continues en deux dimensions, cette approche repose sur la **discrétisation du domaine** : le modèle reçoit une **grille structurée** où chaque pixel contient ses propres coordonnées spatiales normalisées $[x, y]$, permettant ainsi de capturer les **corrélations spatiales** intrinsèques essentielles à la modélisation d'une surface physique ou climatique.

L'architecture a été conçue pour extraire progressivement les caractéristiques du signal via une première couche de **128 filtres** (4×4) utilisant une activation **tangente hyperbolique (tanh)** pour capturer les non-linéarités complexes tout en favorisant la régularité du relief. Cette structure est complétée par une seconde couche de **32 filtres** dédiée à la modélisation des **interactions globales** entre les axes spatiaux, avant qu'une couche de sortie (1×1) ne produise la valeur prédite par régression pour chaque point de la grille. L'apprentissage est piloté par l'optimiseur **Adam**, visant à minimiser l'**erreur quadratique moyenne (MSE)**. Bien que performante, cette approche reste intrinsèquement liée à la **résolution de la grille** d'entrée. Cette dépendance structurelle peut générer des approximations moins fluides que les méthodes stochastiques et limiter la capacité du réseau à généraliser parfaitement en dehors des points d'entraînement directs.

1.3.2 Analyse sur surface périodique simple : $f_1(x, y) = \cos(2\pi(x+y))$

Afin d'évaluer la capacité des modèles à capturer la régularité d'un signal bidimensionnel, nous avons testé la reconstruction d'une surface définie par une fonction périodique simple présentant des crêtes diagonales. L'objectif est ici d'observer comment chaque architecture gère la continuité de la pente et la périodicité du signal sur l'ensemble du domaine spatial $[0, 1] \times [0, 1]$.

L'interprétation visuelle des résultats met en évidence une divergence nette dans la qualité de la modélisation. Comme l'illustre la **Figure 7**, présentant la reconstruction par Processus Gausien, le modèle probabiliste offre une surface d'une **régularité mathématique parfaite**. Le GP parvient à épouser les crêtes diagonales sans aucune distorsion, préservant une courbure fluide et constante qui témoigne de l'efficacité de sa structure de covariance pour les signaux lisses.

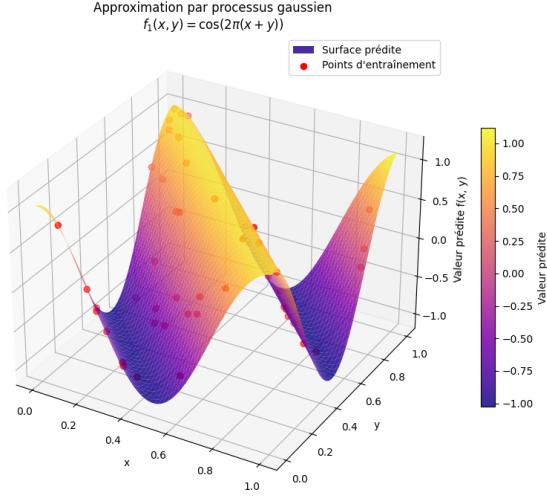


FIGURE 7 – Reconstruction par Processus Gaus-sien de la fonction f_1 .

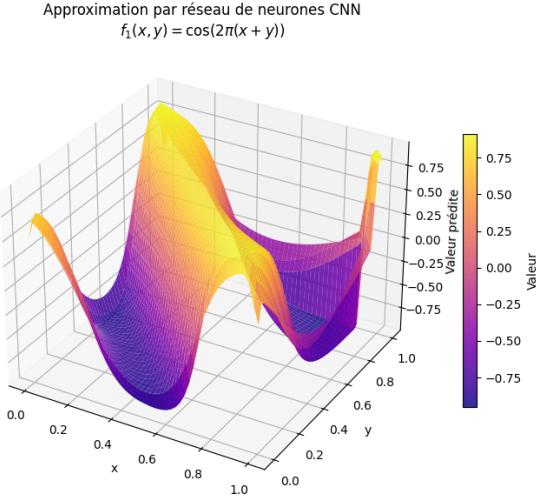


FIGURE 8 – Approximation par CNN 2D de la fonction f_1 .

En revanche, l’analyse de la **Figure 8**, illustrant l’**Approximation par CNN 2D de la fonction f_1** , met en évidence des **contraintes structurelles** intrinsèques à la nature discrète de cette architecture. On observe des **discontinuités locales** et un manque de régularité dans le lissage de la surface, particulièrement prononcés sur les bordures du domaine ainsi que dans les zones de transition thermique. Le réseau de neurones présente des difficultés à restituer une courbure parfaitement fluide, introduisant des **oscillations résiduelles** et des artefacts de discréttisation là où le GP maintient une continuité totale. Cette comparaison confirme que pour des fonctions périodiques régulières, la structure de covariance du GP s’avère plus robuste que l’approche par filtres convolutifs, laquelle reste limitée par sa dépendance à la **Résolution de la grille** pour généraliser une dynamique trigonométrique continue.

1.3.3 Analyse sur surface à motifs périodiques : $f_2(x,y) = \sin(2\pi x) \cdot \cos(2\pi y)$

Pour cette seconde évaluation spatiale, nous avons sollicité les modèles sur une surface présentant des motifs répétitifs de pics et de vallées alternés, définie par la fonction $f_2(x,y) = \sin(2\pi x) \cdot \cos(2\pi y)$. Ce test critique permet de mesurer la capacité des architectures à restituer des extrema locaux multiples avec une amplitude constante sur l’intégralité du domaine $[0,1] \times [0,1]$.

L’analyse comparative des résultats souligne une disparité de performance substantielle, l’écart visuel devenant ici particulièrement marqué. Comme illustré par la **Figure 9**, détaillant la **Pré-diction GP capturant les extrema**, le Processus Gaussien reconstitue avec une fidélité rigoureuse l’amplitude intégrale du signal original. Le modèle parvient à stabiliser les sommets et les vallées sans perte de dynamique, confirmant l’aptitude de sa structure de covariance à modéliser des surfaces hautement non-linéaires avec une précision quasi-parfaite ($R^2 = 0,9996$).

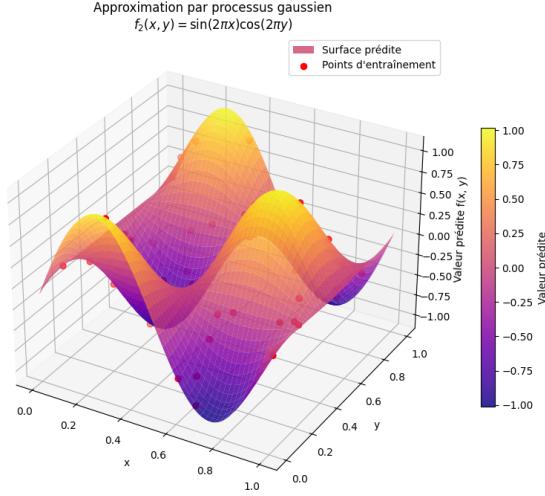


FIGURE 9 – Prédiction GP capturant les extrema de la fonction f_2 .

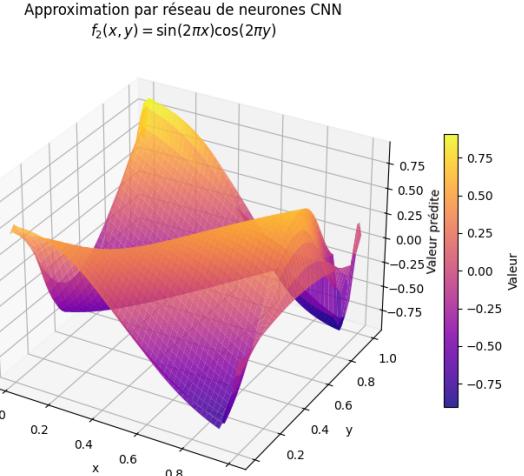


FIGURE 10 – Approximation par CNN 2D de la structure complexe f_2 .

À l'inverse, l'observation de la **Figure 10**, présentant l'**Approximation par CNN 2D de la structure complexe**, révèle un **effondrement notable de la dynamique** du signal prédit. On constate une atténuation systématique des extrema qui apparaissent « écrasés », tandis que la surface manifeste une tendance à l'aplatissement prématûre vers les bordures du domaine. De plus, l'émergence d'**artefacts géométriques** sous forme de lignes de cassure témoigne des contraintes de généralisation du réseau de neurones. Ces distorsions soulignent la difficulté pour les couches convolutives de maintenir la cohérence d'une fonction trigonométrique globale en dehors des points d'entraînement directs, conduisant à une dégradation sensible des métriques de performance ($R^2 = 0,6233$).

1.3.4 Analyse sur surface complexe et évaluation de la robustesse prédictive : $f_3(x,y) = \sin(3\pi x)\cos(2\pi y) + e^{-5((x-0.5)^2+(y-0.5)^2)}$

Pour pousser l'évaluation à un niveau critique, nous introduisons une fonction hybride combinant des oscillations haute fréquence et une singularité centrale exponentielle : $f_3(x,y) = \sin(3\pi x)\cos(2\pi y) + e^{-5((x-0.5)^2+(y-0.5)^2)}$. Ce signal constitue le test le plus exigeant de notre étude, car il impose aux modèles de gérer simultanément des dynamiques à différentes échelles spatiales sans sacrifier la précision locale.

Les résultats obtenus confirment de manière univoque la supériorité du Processus Gaussien sur des signaux à forte hétérogénéité. Comme le montre la **Figure 11**, illustrant la **Reconstruction GP isolant le pic central**, le modèle probabiliste parvient à restituer parfaitement la singularité exponentielle tout en stabilisant rigoureusement le relief périodique périphérique. Cette capacité à dissocier les échelles de variation permet au GP de maintenir une fidélité d'approximation exceptionnelle, validée par un score R^2 de 0,9874.

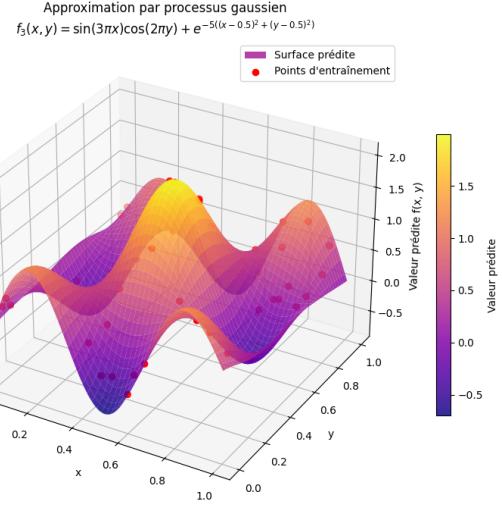


FIGURE 11 – Reconstruction GP isolant le pic central de la fonction f_3 .

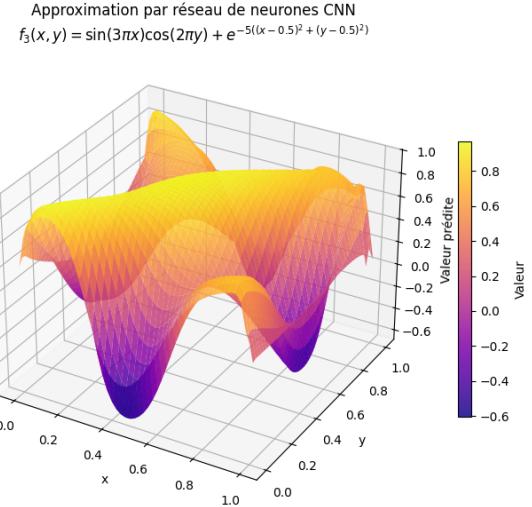


FIGURE 12 – Approximation CNN 2D de la fonction hybride f_3 .

À l'inverse, l'analyse de la **Figure 12**, présentant l'**Approximation CNN 2D de la fonction hybride**, révèle une défaillance quasi totale du réseau de neurones face à cette complexité. On observe que le pic central subit une **déformation structurelle** majeure, tandis que le relief environnant est noyé dans un **bruit visuel important**. Le réseau semble incapable d'arbitrer entre la gestion de la forte amplitude du pic et la finesse des micro-oscillations, produisant un rendu chaotique et scientifiquement inutilisable pour une cartographie de précision. Cette chute drastique des performances ($R^2 = 0,4978$) démontre les limites des couches convolutives standards lorsque le signal s'écarte d'une régularité spatiale uniforme.

1.3.5 Synthèse comparative des performances en dimension 2

Afin d'étayer les observations visuelles, nous présentons une **analyse quantitative** rigoureuse des performances métriques. Cette étape de synthèse est cruciale pour valider la robustesse des modèles face à des surfaces de complexités croissantes, allant de la régularité trigonométrique à l'hétérogénéité d'un signal hybride. Le tableau ci-dessous regroupe l'**Erreur Quadratique Moyenne (MSE)**, le **coefficients de détermination (R^2)** ainsi que l'**incertitude prédictive** propre à l'approche bayésienne.

TABLE 1 – Comparaison des indicateurs de performance GP vs CNN sur les surfaces de test 2D

Fonction	Modèle	MSE (Erreur)	Score R^2	Incertitude (GP)
f_1 : cosinus simple	Processus Gaussien	0.0002	0.9996	0.0090
	Réseau CNN 2D	0.0507	0.8986	/
f_2 : boîte à œufs	Processus Gaussien	0.0001	0.9996	0.0056
	Réseau CNN 2D	0.0941	0.6233	/
f_3 : pic + oscillations	Processus Gaussien	0.0037	0.9874	0.0191
	Réseau CNN 2D	0.1462	0.4978	/

L'analyse de ces indicateurs confirme que le **Processus Gaussien (GP)** s'impose comme le modèle dominant pour la modélisation spatiale en dimension 2. Sur les fonctions f_1 et f_2 , le GP atteint une précision quasi absolue avec des **scores R^2 de 0,9996**. En comparaison, le réseau CNN manifeste des signes de **faiblesse structurelle**, particulièrement sur la fonction f_2 où sa capacité à restituer la dynamique des extrema chute drastiquement, avec un score de 0,6233.

Le cas de la fonction hybride f_3 constitue l'enseignement le plus révélateur de cette étude. Alors que le Processus Gaussien maintient une **excellente capacité d'approximation ($R^2 = 0,9874$)**, les performances du réseau CNN s'effondrent totalement avec un score de 0,4978 et une erreur (MSE) environ **40 fois supérieure** à celle du GP (0,1462 contre 0,0037). Cette divergence majeure s'explique par la supériorité du GP à gérer nativement les **corrélations spatiales** et les ruptures d'échelle via son noyau, là où le CNN peine à généraliser une topologie complexe à partir de coordonnées brutes.

En conclusion de ce volet expérimental, le Processus Gaussien confirme son statut d'**outil de référence** pour la cartographie de précision. Au-delà de sa supériorité métrique, il offre une **estimation de l'incertitude** extrêmement faible (comprise entre 0,005 et 0,019), apportant une garantie de fiabilité statistique indispensable pour la modélisation de données forestières réelles. Cette propriété d'auto-évaluation, absente de l'architecture CNN testée, est un atout majeur pour identifier les zones où la prédiction climatique nécessiterait des relevés de terrain complémentaires.

Transition vers l'application forestière : Cette étude comparative sur fonctions simulées a démontré la robustesse des Processus Gaussiens (GP) pour modéliser des signaux complexes avec une grande fluidité et une gestion fiable de l'incertitude. Forts de ces résultats théoriques, nous allons désormais appliquer ces modèles à des données forestières réelles. L'enjeu sera d'évaluer si la supériorité du GP se confirme face au bruit et à la variance de séries temporelles pluriannuelles (température et humidité), afin de fournir des prédictions robustes pour la préservation de ces écosystèmes fragiles.

2 Application aux écosystèmes forestiers : Étude de cas réelle

Après avoir validé la supériorité des Processus Gaussiens sur des fonctions mathématiques simulées, nous orientons désormais nos travaux vers une application concrète en écologie forestière. L'objectif est de confronter nos modèles à la complexité de données réelles de terrain, caractérisées par un bruit de mesure et une variabilité environnementale élevée.

2.1 Acquisition et description du jeu de données ERA5-Land

Pour cette étude, nous exploitons la base de données **ERA5-Land monthly averaged data** du service Copernicus Climate Change Service (C3S). Ce jeu de données fournit une reconstruction globale de l'évolution du climat terrestre sur plusieurs décennies avec une résolution spatiale fine. Pour notre analyse, nous avons ciblé spécifiquement le département du **Puy-de-Dôme (63)** sur la période allant de **2000 à 2007** (cf. figure 13).

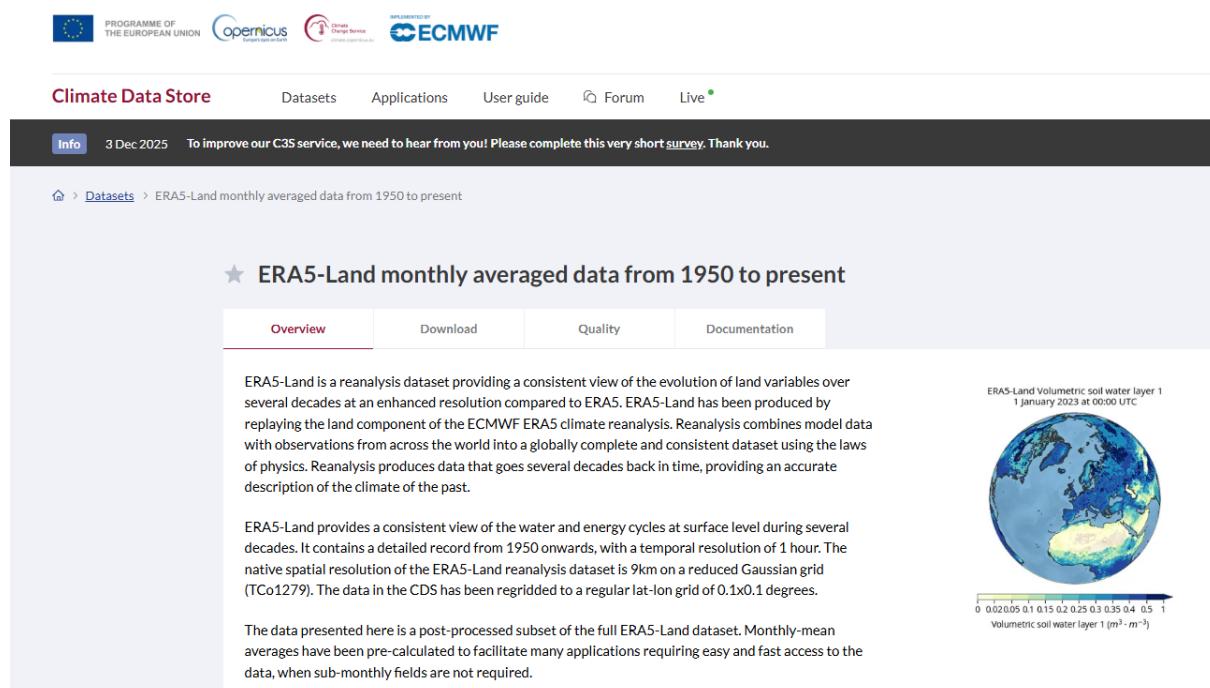


FIGURE 13 – Portail du Climate Data Store (CDS) permettant l'accès aux réanalyses climatiques ERA5-Land.

Afin d'automatiser et de garantir la reproductibilité de l'extraction, nous utilisons l'API **Copernicus Climate Data Store (CDS)** via un script Python dédié. Le code configure une requête client (cf. figure 14) qui définit précisément les paramètres de téléchargement : le type de produit (moyennes mensuelles), le format de sortie (NetCDF compressé en ZIP) et les coordonnées géographiques limitant la zone d'étude aux frontières du Puy-de-Dôme, définies par une boîte englobante (Nord : 45.7, Est : 3.2, Sud : 45.6, Ouest : 3.1).

ERAS-Land monthly averaged data from 1950 to present Details	2025-11-25 10:55:22pm	2025-11-25 11:19:23pm	● Complete	Download	<input type="checkbox"/>
ERAS-Land monthly averaged data from 1950 to present Details	2025-11-25 10:40:05pm	2025-11-25 11:19:16pm	● Complete	Download	<input type="checkbox"/>

FIGURE 14 – Interface de configuration de la requête sur le Climate Data Store montrant la sélection des variables et de la période temporelle.

La requête permet de récupérer un ensemble complet de variables thermiques essentielles pour comprendre les interactions entre l’atmosphère et le sol forestier. Nous collectons notamment la température du point de rosée à 2 mètres, la température de l’air à 2 mètres, ainsi que la **température de surface (Skin temperature - SKT)**. En complément, les températures du sol sont extraites sur quatre niveaux de profondeur distincts (Soil temperature level 1 à 4), permettant une analyse verticale des échanges thermiques. Pour amorcer notre modélisation spatio-temporelle, nous nous concentrerons prioritairement sur la variable **SKT** à l’échelle départementale.

Une fois la requête API traitée et le fichier converti, nous obtenons une base de données structurée au format CSV. La structuration et le nettoyage des données massives ont été effectués à l’aide de Pandas [9]. Ce fichier constitue la matière première de notre analyse. Chaque ligne représente une mesure spatio-temporelle précise sur le département du Puy-de-Dôme.

time	latitude	longitude	skt	sd	swvl1	skt_C
01/01/2000	47.0	2.0	276.47827	2.7656555e-05	0.3859253	3.3282776
01/01/2000	47.0	2.1	276.46265	2.861023e-05	0.38671875	3.3126526
01/01/2000	47.0	2.2	276.42358	2.9563904e-05	0.38624573	3.27359
01/01/2000	47.0	2.3000000000000003	276.31616	3.0517578e-05	0.38690186	3.1661682
01/01/2000	47.0	2.4000000000000004	276.2224	3.33786e-05	0.3874817	3.0724182
01/01/2000	47.0	2.5000000000000004	276.1267	3.8146973e-05	0.388031	2.976715
01/01/2000	47.0	2.6000000000000005	276.01733	4.4822693e-05	0.3888092	2.86734
01/01/2000	47.0	2.7000000000000006	275.95483	5.340576e-05	0.38885498	2.80484
01/01/2000	47.0	2.8000000000000007	275.9431	6.484985e-05	0.3888092	2.7931213
01/01/2000	47.0	2.9000000000000001	276.03296	8.106232e-05	0.38970947	2.882965
01/01/2000	47.0	3.0000000000000001	275.9978	9.727478e-05	0.39015198	2.8478088
01/01/2000	47.0	3.1000000000000001	275.9314	0.00011730194	0.3872223	2.7814026
01/01/2000	47.0	3.2000000000000001	275.8103	0.0001411438	0.37660217	2.6603088
01/01/2000	47.0	3.3000000000000001	275.6267	0.00024986267	0.37815857	2.476715

FIGURE 15 – Aperçu de la structure tabulaire des données climatiques extraites (Fichier CSV).

Le jeu de données se décompose selon les colonnes techniques suivantes :

- **time** : L’horodatage de la mesure, correspondant à un échantillonnage mensuel sur la période 2000-2007.
- **latitude / longitude** : Les coordonnées géographiques précises permettant de situer chaque relevé sur la carte du département.
- **skt (Skin Temperature)** : La température de surface brute exprimée en Kelvin.
- **sd (Snow Depth)** : La profondeur de neige (équivalent eau), paramètre influençant l’albédo et l’isolation du sol.
- **swvl1 (Volumetric soil water layer 1)** : Le volume d’eau contenu dans la première couche de sol (0-7 cm), indicateur de l’humidité de surface.
- **skt_C** : La température de surface convertie par nos soins en **degrés Celsius** pour une interprétation physique plus intuitive.

Cet ensemble de données constitue le socle nécessaire pour engager les travaux d'ingénierie statistique détaillés dans la section suivante. Cette étape cruciale permettra d'identifier les tendances saisonnières et les anomalies thermiques locales à travers la génération de **cartes de chaleur (Heatmaps)**. Ces visualisations seront essentielles pour appréhender la distribution thermique sur la topographie du Puy-de-Dôme et analyser précisément le **gradient de température** au fil des années.

Une fois cette structure spatio-temporelle bien établie, nous passerons à la phase de **prédition**. Nous confronterons alors la robustesse des **Processus Gaussiens** à des architectures de réseaux de neurones plus complexes, intégrant notamment des couches **LSTM (Long Short-Term Memory)**. Ce choix technologique nous permettra de capturer les dépendances temporelles à long terme inhérentes aux cycles climatiques forestiers, afin d'affiner la précision de nos modèles face aux variations réelles du terrain.

2.2 Ingénierie des données et visualisations cartographiques

Afin d'appréhender la structure spatio-temporelle de notre jeu de données ERA5-Land, une étape cruciale d'ingénierie des données est nécessaire. Avant de procéder à la modélisation prédictive, nous transformons les relevés tabulaires en représentations géographiques explicites pour identifier les gradients thermiques régionaux.

2.2.1 Écosystème logiciel pour l'analyse spatiale

Le traitement des données géographiques et la génération de visuels dynamiques reposent sur trois bibliothèques Python piliers. Leur utilisation conjointe permet de transformer des coordonnées brutes en une structure cartographique exploitable.

- **Cartiflette [11]** : Ce module est une solution de récupération automatisée des fonds de carte officiels (IGN, INSEE). *Utilité dans le code* : Elle est utilisée via la fonction `carti_download` pour requêter directement les vecteurs géographiques des départements français au format GeoJSON. Cela nous permet de délimiter précisément la zone d'étude auvergnate (codes 03, 15, 43, 63) sans manipulation manuelle de fichiers externes.
- **GeoPandas [10]** : Véritable extension de Pandas dédiée à la donnée vectorielle, cette bibliothèque introduit le concept de *GeoDataFrame*. *Utilité dans le code* : Elle assure la conversion des coordonnées tabulaires (latitude/longitude) en objets géométriques projetés. Elle est le moteur de la jointure spatiale (`sjoin`), une opération critique qui filtre nos points de mesure ERA5-Land pour ne conserver que ceux situés physiquement à l'intérieur des frontières administratives de l'Auvergne.
- **Folium [12]** : S'appuyant sur la puissance de la bibliothèque JavaScript *Leaflet*, Folium assure la couche de visualisation interactive. *Utilité dans le code* : Elle permet de générer la carte finale (`folium.Map`) et d'y superposer nos données thermiques sous forme de marqueurs circulaires colorés. Grâce à l'intégration de *Branca* pour la gestion des colormaps, elle transforme nos calculs de moyennes min-max en un gradient visuel intuitif où chaque point de la forêt peut être interrogé via des pop-ups informatifs.

2.2.2 Cartes de chaleur et gradients de température

La méthodologie adoptée pour la création des **Heatmaps** repose sur un processus de raffinement des données brutes ERA5-Land afin de rendre les tendances climatiques intelligibles à l'échelle régionale. Pour optimiser la lisibilité des cartes et éviter une surcharge visuelle, nous appliquons une réduction de la densité de la grille par un échantillonnage d'un point sur deux.

Le traitement statistique consiste à agréger les relevés temporels pour chaque coordonnée géographique : nous calculons la moyenne arithmétique entre les valeurs minimales et maximales de la température de surface (SKT) sur l'ensemble de l'année sélectionnée. Cette métrique "min-max" permet de lisser les variations saisonnières extrêmes tout en conservant une image fidèle du gradient thermique annuel.

Pour garantir la précision géographique de l'étude, nous utilisons les capacités de jointure spatiale de *GeoPandas* afin de circonscrire l'analyse aux limites administratives officielles. Comme l'illustre la **Figure 16**, notre périmètre d'étude se concentre sur les quatre départements constituant la région Auvergne, offrant ainsi un cadre topographique varié (plaines et massifs montagneux) idéal pour tester la robustesse de nos modèles prédictifs.

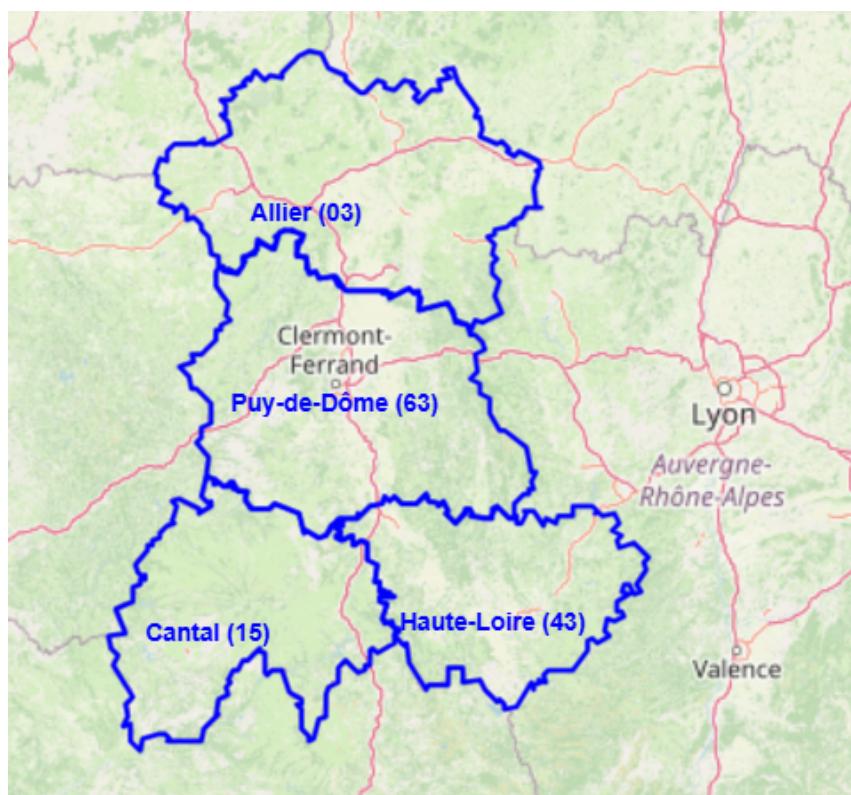


FIGURE 16 – Délimitation géographique de la zone d'étude comprenant les départements de l'Allier (03), du Cantal (15), de la Haute-Loire (43) et du Puy-de-Dôme (63).

2.2.3 Représentation des variables SKT et SKT_C (Année 2000)

Pour amorcer l’analyse spatiale, nous avons procédé à une première représentation de la température de surface brute (SKT) exprimée en Kelvin. Cette étape est essentielle pour valider la cohérence physique des données extraites du service Copernicus avant toute manipulation complexe. La procédure automatisée commence par le chargement du jeu de données nettoyé et l’isolement de l’année cible, ici l’an 2000, tout en opérant une réduction de la grille par un échantillonnage d’un point sur deux afin d’épurer le rendu visuel.

Le cœur du traitement repose sur l’intégration des contours géographiques via la bibliothèque *Cartiflette*, qui nous permet de projeter et de filtrer précisément les données sur les quatre départements de l’ancienne région Auvergne (Allier, Cantal, Haute-Loire et Puy-de-Dôme). Grâce à l’utilisation de *GeoPandas*, nous réalisons une jointure spatiale (*sjoin*) pour ne conserver que les relevés situés à l’intérieur de ces frontières administratives. Nous calculons ensuite la valeur moyenne de la température de surface pour chaque coordonnée géographique, agrégeant ainsi les variations temporelles mensuelles en une donnée annuelle représentative.

La visualisation finale est générée à l’aide de *Folium*, où chaque point de mesure est représenté par un marqueur circulaire dont la couleur est indexée sur une échelle thermique inversée (du bleu pour le froid au rouge pour le chaud). Cette carte interactive permet d’explorer les zones de chaleur par simple survol, affichant des pop-ups informatifs avec les coordonnées exactes et la température moyenne (cf. figure 17). Nous avons appliqué rigoureusement la même méthodologie pour la variable **SKT_C (Celsius)**, confirmant ainsi la validité de notre conversion mathématique ($T_C = T_K - 273.15$) et assurant une base sémantique plus intuitive pour l’interprétation climatique ultérieure (cf. figure 18).

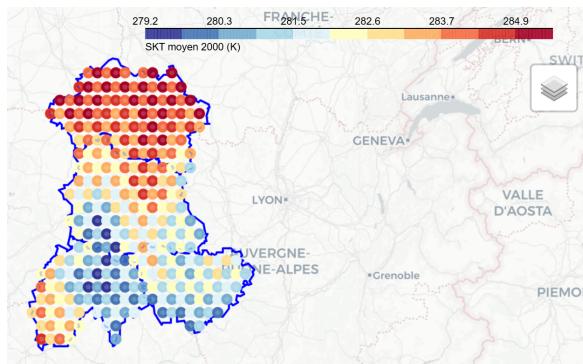


FIGURE 17 – Représentation de la variable SKT (Kelvin) sur la région Auvergne en 2000.

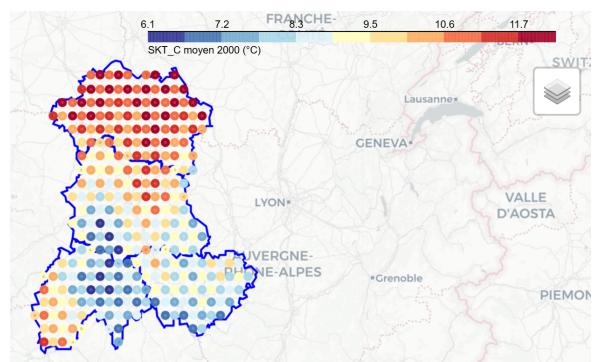


FIGURE 18 – Représentation de la variable SKT_C (Celsius) sur la région Auvergne en 2000.

2.2.4 Analyse de la moyenne min-max par coordonnée

Afin d'obtenir une vision synthétique du climat régional pour l'année 2000, nous avons mis en œuvre une méthodologie d'agrégation statistique simplifiée mais robuste : le calcul de la moyenne min-max. Cette approche consiste à isoler, pour chaque point géographique, les valeurs extrêmes enregistrées sur l'année afin d'en extraire la valeur médiane via la formule $\frac{\min+\max}{2}$. Ce choix méthodologique est particulièrement pertinent dans une étude préliminaire, car il permet de gommer le bruit des fluctuations mensuelles mineures tout en capturant l'amplitude thermique réelle subie par les écosystèmes forestiers.

Le processus de visualisation repose sur un échantillonnage spatial stratégique. Pour assurer une clarté optimale sur la carte interactive sans saturer l'information, nous avons appliqué une réduction de la densité de la grille d'origine en ne conservant qu'un point sur deux. Cette sélection permet de maintenir une résolution suffisante pour identifier les micro-climats locaux tout en offrant une lecture fluide du gradient thermique à l'échelle des quatre départements de l'Auvergne.

La cartographie finale, illustrée par la **Figure 19**, utilise une *colormap* divergente où les tons chauds mettent en exergue les zones de température de surface élevée. Le recours à une jointure spatiale rigoureuse garantit que seuls les relevés contenus dans les frontières administratives officielles sont pris en compte, éliminant ainsi les points marginaux. Chaque marqueur ainsi positionné devient un vecteur d'information complet, permettant par un simple survol d'accéder aux coordonnées précises et à la moyenne thermique calculée, facilitant l'identification visuelle des couloirs de chaleur régionaux.

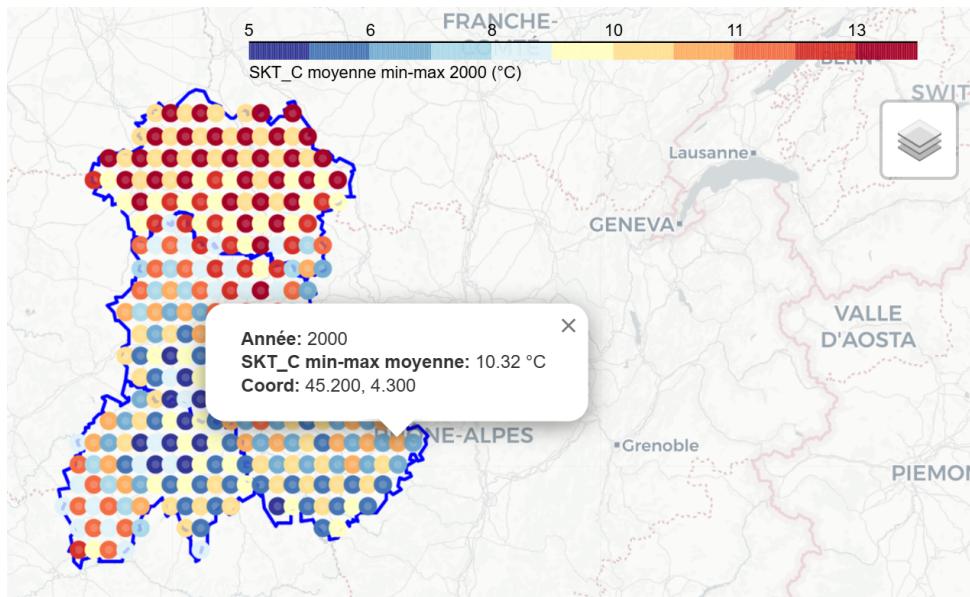


FIGURE 19 – Carte de la moyenne annuelle min-max pour SKT_C en 2000 avec délimitation départementale.

2.2.5 Gradient de température : Évolution 2000 vs 2007

Pour parachever cette phase d'ingénierie des données, nous avons procédé à une analyse comparative visant à isoler l'évolution thermique spatiale sur l'ensemble de la période d'étude. La spécificité de cette approche réside dans le calcul d'un **gradient thermique** (Δ) par point géographique, obtenu en soustrayant les moyennes annuelles de deux années charnières : 2000 et 2007. Cette méthode permet de mettre en lumière les zones géographiques ayant subi les variations de température de surface les plus significatives au cours de ces sept années.

La méthodologie conserve la rigueur des étapes précédentes, notamment l'échantillonnage d'un point sur deux pour garantir la lisibilité et le calcul de la moyenne min-max pour stabiliser les valeurs annuelles. L'innovation ici réside dans la fusion (merge) des données de l'an 2000 et de l'an 2007 par coordonnées via la bibliothèque **Pandas** [9], permettant de quantifier précisément l'écart thermique local. Pour la visualisation, nous avons opté pour une palette de couleurs hautement contrastée (du bleu sombre au rouge vif), facilitant l'identification immédiate du réchauffement ou du refroidissement relatif des parcelles forestières auvergnates.

Comme l'illustre la **Figure 20**, le calcul du delta 2007 – 2000 révèle la progression thermique globale, où les tons chauds marquent une augmentation de la température de surface sur la période. Bien qu'une analyse de la soustraction inverse (2000 – 2007) ait été initialement menée, elle a été écartée de ce rapport car elle présentait une information strictement symétrique. Cette étape de vérification a néanmoins permis de confirmer la stabilité des données et la cohérence des calculs de deltas géolocalisés : les zones de forte augmentation thermique correspondent parfaitement aux zones de retrait dans le calcul inverse, validant ainsi la robustesse de notre jointure spatiale.

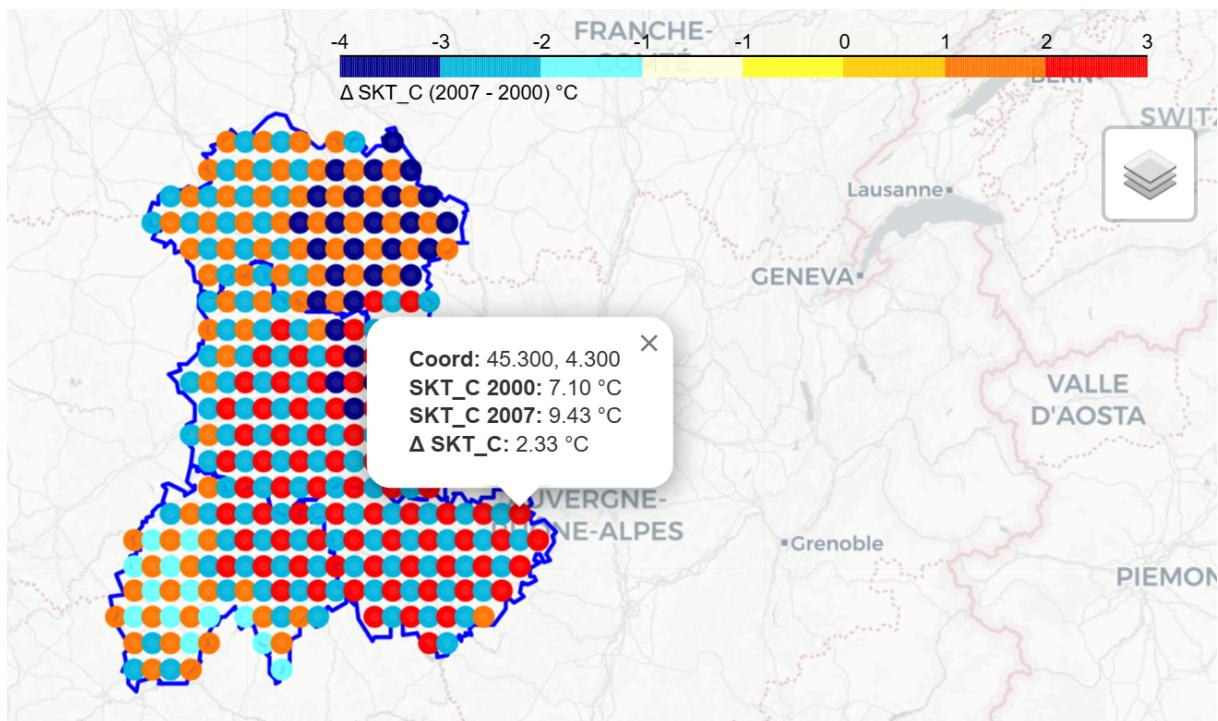


FIGURE 20 – Gradient thermique SKT_C (Δ 2007 - 2000) en Auvergne.

En résumé, l’analyse de cette cartographie nous montre une réalité très importante : le réchauffement en Auvergne n’est pas le même partout. En calculant l’écart entre l’an 2000 et l’an 2007, nous avons réussi à identifier précisément des « couloirs de chaleur ». Ce sont des zones spécifiques où les forêts souffrent beaucoup plus qu’ailleurs à cause d’une hausse rapide des températures de surface.

Cette phase de préparation est fondamentale pour la suite de notre projet. Elle nous a permis de transformer des millions de chiffres bruts en une information visuelle et compréhensible. En repérant ces zones de forte variation, nous ne nous contentons plus d’observer le passé ; nous préparons le terrain pour nos futurs modèles informatiques. Grâce à cette base de données bien organisée, nos algorithmes ne travailleront pas « en aveugle » : ils auront une vision claire des tendances locales, ce qui est indispensable pour obtenir des prédictions vraiment fiables et utiles pour la protection de nos forêts.

Transition vers la modélisation : Ces indicateurs visuels marquent la fin de la phase descriptive et ouvrent la voie à la **prédition**. L’enjeu sera désormais de confronter la robustesse des **Processus Gaussiens**, dont nous avons prouvé l’efficacité pour l’interpolation spatiale en 2D, à des architectures de réseaux de neurones complexes. Nous intégrerons notamment des couches **LSTM (Long Short-Term Memory)**, spécifiquement conçues pour capturer les dépendances séquentielles à long terme, afin de déterminer quel modèle parvient le mieux à anticiper les dynamiques thermiques futures de ces écosystèmes fragiles.

2.3 Le paradigme LSTM : Mémoire et apprentissage séquentiel

La phase précédente d'ingénierie des données a permis de transformer des mesures tabulaires brutes en une vision géographique cohérente, mettant en lumière des gradients thermiques et des hétérogénéités spatiales critiques sur le territoire auvergnat. Ce constat visuel du "stress environnemental" constitue le point de départ de notre démarche prédictive : il ne s'agit plus seulement de décrire le passé, mais d'anticiper les dynamiques futures de la température de surface (SKT).

Pour capturer la dimension temporelle de nos données climatiques, nous introduisons les réseaux de neurones **LSTM (Long Short-Term Memory)**. Contrairement aux réseaux de neurones classiques (MLP) qui traitent chaque donnée de manière isolée, le LSTM possède une architecture récurrente capable de maintenir une "**mémoire interne**" sur le long terme. Le réseau de neurones récurrent LSTM, dont l'architecture théorique a été posée par Hochreiter et Schmidhuber [3], est ici développé avec TensorFlow [8].

Le principe fondamental repose sur une structure de **cellule** régulée par trois portes logiques permettant de contrôler le flux d'information :

- **Forget Gate (Porte d'oubli)** : Elle détermine quelles informations passées, comme des anomalies climatiques ponctuelles ou des cycles saisonniers obsolètes, doivent être supprimées de la mémoire.
- **Input Gate (Porte d'entrée)** : Elle sélectionne les nouvelles informations pertinentes issues du mois actuel pour mettre à jour l'état de la cellule.
- **Output Gate (Porte de sortie)** : Elle décide quelle partie de la mémoire consolidée sera extraite pour produire la prédiction de température à l'instant t .

Dans notre cadre **spatio-temporel**, les LSTM nous servent à modéliser la continuité des cycles biogéophysiques. Là où un modèle classique échouerait à percevoir la saisonnalité, le LSTM apprend la "logique" séquentielle des données **SKT_C**. Il est capable d'identifier que la température d'un mois donné dépend étroitement de l'inertie thermique des mois précédents, permettant ainsi une extrapolation bien plus fine des tendances climatiques forestières.

2.3.1 Protocole d'évaluation et stratégie de confrontation spatio-temporelle

L'objectif de cette phase est de mesurer la capacité de généralisation des modèles en les confrontant à une année "aveugle" pour eux : l'année 2008. Ce protocole repose sur la transformation des prédictions numériques brutes en **heatmaps mensuelles** structurées, permettant une analyse visuelle et géographique approfondie des gradients thermiques sur l'ensemble du département du Puy-de-Dôme.

1. Apprentissage (2000-2007) : Les modèles sont entraînés sur les températures de surface en Celsius (**SKT_C**) couvrant une période historique de 8 ans. Ce volume de données est crucial pour permettre au **LSTM** de stabiliser sa mémoire des cycles saisonniers complexes et au **GP** d'ajuster ses hyperparamètres de noyau sur une variance environnementale réelle.

2. Prédiction et cartographie mensuelle (2008) : Une fois l'apprentissage validé, nous demandons aux modèles de prédire l'intégralité de l'année 2008, mois par mois. Pour chaque échantillon temporel, le modèle reçoit en entrée les coordonnées géographiques (x, y) et produit une valeur de température prédictive T_{pred} . Ce processus automatisé permet de générer une série

de **12 heatmaps distinctes** (de janvier à décembre), offrant une visualisation dynamique de la reconstruction thermique effectuée par chaque algorithme face à la "vérité terrain" d'ERA5-Land.

3. Précision spatiale vs Dynamique temporelle Cette étude met en concurrence deux approches technologiques radicalement différentes pour la surveillance forestière :

- Les **Processus Gaussiens (GP)** misent sur la **précision géométrique**. Ils excellent à lisser les données et à reconstruire fidèlement l'espace thermique à un instant t donné, mais ils traitent chaque mois comme une entité statistiquement indépendante, ignorant ainsi la dépendance séquentielle intrinsèque au climat.
- Les **Réseaux de Neurones LSTM** misent sur la **continuité temporelle**. En apprenant la dynamique cyclique du signal sur plusieurs années, ils cherchent à anticiper le futur climatique en utilisant leur mémoire interne, quitte à être parfois moins précis sur la localisation exacte des micro-variations topographiques locales.

La confrontation finale sur les 12 fenêtres mensuelles de 2008 permettra de déterminer si, pour la gestion durable des forêts, il est préférable de privilégier un modèle capable de reconstruire l'espace avec une haute fidélité mathématique (GP) ou un modèle capable d'anticiper les cycles temporels profonds (LSTM).

2.3.2 Analyse des résultats : Est-ce que l'IA arrive à prédire l'année 2008 ?

L'évaluation finale de nos modèles a consisté en une projection prédictive sur l'intégralité de l'année 2008, période volontairement exclue de la phase d'apprentissage pour garantir l'intégrité de la validation. Cette étape constitue le test de robustesse ultime face à la non-stationnarité des données réelles issues d'ERA5-Land, caractérisées par des cycles saisonniers marqués et des anomalies thermiques locales

Justification des configurations techniques :

- **Modèle Gaussian Process (GP)** : Le choix d'un noyau composite répond à la nécessité de modéliser trois composantes distinctes du signal : la **périodicité annuelle** via l'*ExpSineSquared*, les **dérivés locaux** du relief via le *RationalQuadratic*, et le **bruit d'acquisition** via le *WhiteKernel*. Bien que l'optimisation des hyperparamètres par 5 redémarrages successifs ait permis d'atteindre un optimum mathématique stable, le modèle reste limité par sa nature d'échantillonneur statique face à une dynamique temporelle longue.
- **Architecture LSTM** : Pour surmonter les limites des réseaux denses, nous avons implanté une structure récurrente à deux niveaux. L'usage de **128 et 64 unités LSTM** permet au réseau de hiérarchiser les dépendances temporelles (cycles de 12 mois), tandis que l'encodage cyclique des mois (*sin/cos*) assure la continuité entre décembre et janvier. Le **Dropout (0.2)** a été introduit comme régulateur pour forcer le réseau à apprendre des caractéristiques robustes et éviter la mémorisation du bruit.

Bilan technique (GP) : Malgré une optimisation rigoureuse des hyperparamètres, le modèle se heurte à sa nature d'échantillonneur statique. En l'absence de réelle mémoire du passé, il a tendance à produire des prédictions trop "lisses". Il privilégie une moyenne statistique sûre plutôt que de capturer les pics de chaleur réels, ce qui limite son utilité pour la détection des canicules.

Bilan technique (LSTM) : Le réseau LSTM s'impose comme le grand vainqueur. Sa capacité à mémoriser les dépendances séquentielles lui permet de suivre la courbe réelle des températures avec une fidélité remarquable. Il ne se contente pas de deviner une moyenne ; il anticipe les cycles, atteignant une précision de **1,53 °C** (Score $R^2 = 0,938$).

Les résultats obtenus soulignent une disparité de performance flagrante, illustrée par le tableau ci-dessous :

Tableau 4 : Comparaison des performances de prédiction sur l'année 2008

MODÈLE	MSE	RMSE	MAE	Score R^2
LSTM	2.35	1.53	1.16	0.938
Gaussian Process	37.72	6.14	4.91	0.009

L'interprétation de ces chiffres révèle que le **LSTM capture 93,8% de la variance** du signal thermique sur l'année 2008. Son erreur moyenne (RMSE) de **1,53°C** est particulièrement remarquable pour des données réelles, validant sa capacité à « retenir » la dynamique saisonnière apprise entre 2000 et 2007 pour extrapoler l'année suivante.

À l'inverse, l'effondrement du **Gaussian Process** avec un score R^2 proche de zéro (0,009) constitue un enseignement majeur. Malgré la complexité de ses noyaux, le GP souffre d'un manque de « mémoire » séquentielle : il traite le temps comme une variable d'indexation linéaire et non comme une structure périodique rétroactive. Avec une erreur moyenne de **6,14°C**, le GP s'avère incapable d'anticiper l'amplitude des pics de chaleur ou les chutes de température hivernales, tendant vers une prédiction moyenne lissée totalement inadaptée aux besoins de la gestion forestière de précision.

En conclusion, cette confrontation démontre que si le GP est un outil d'excellence pour l'interpolation spatiale (2D), le **LSTM est le seul paradigme capable de gérer la complexité spatio-temporelle** à grande échelle. Pour la surveillance des écosystèmes forestiers, où la précision temporelle est vitale pour anticiper les stress hydriques ou thermiques, l'architecture récurrente s'impose comme le socle technologique de référence.

Contrairement aux tests sur fonctions théoriques, l'application aux données réelles montre une **domination absolue du LSTM**. Avec un score R^2 de **0,938**, il capture la quasi-totalité de la dynamique thermique, là où le GP échoue totalement à généraliser sur une année complète de données non stationnaires ($R^2 \approx 0,01$).

2.3.3 Analyse dynamique des tendances temporelles et structures 3D

L'évaluation de la pertinence d'un modèle de prédiction forestière ne peut se limiter à une lecture brute des métriques d'erreur ; elle nécessite une confrontation visuelle avec la dynamique saisonnière réelle du milieu. Cette étape permet de vérifier si l'architecture parvient à assimiler l'inertie thermique et les cycles biogéochimiques propres au département du Puy-de-Dôme.

La **Figure 21** illustre de manière flagrante la divergence de comportement entre les deux approches face à la vérité terrain (courbe noire). On observe que la **Prédiction LSTM**, représentée par la ligne rouge pointillée, suit avec une précision chirurgicale l'amplitude thermique de l'année 2008. Le modèle parvient à anticiper le réchauffement printanier dès le mois de mars et capture fidèlement le pic de chaleur estival aux alentours de 292 K (environ 19°C) en juillet.

À l'inverse, la courbe bleue du **Gaussian Process (GP)** révèle une incapacité structurelle à l'extrapolation temporelle. Le modèle semble « lisser » la réalité, produisant une droite quasi-linéaire qui traverse la moyenne annuelle sans jamais épouser les extrema. Cette dérive s'explique par le fait que le GP, malgré son noyau périodique, tend vers une régression à la moyenne dès qu'il s'éloigne de sa fenêtre d'entraînement, là où le LSTM utilise sa mémoire interne pour projeter la dynamique cyclique apprise entre 2000 et 2007.

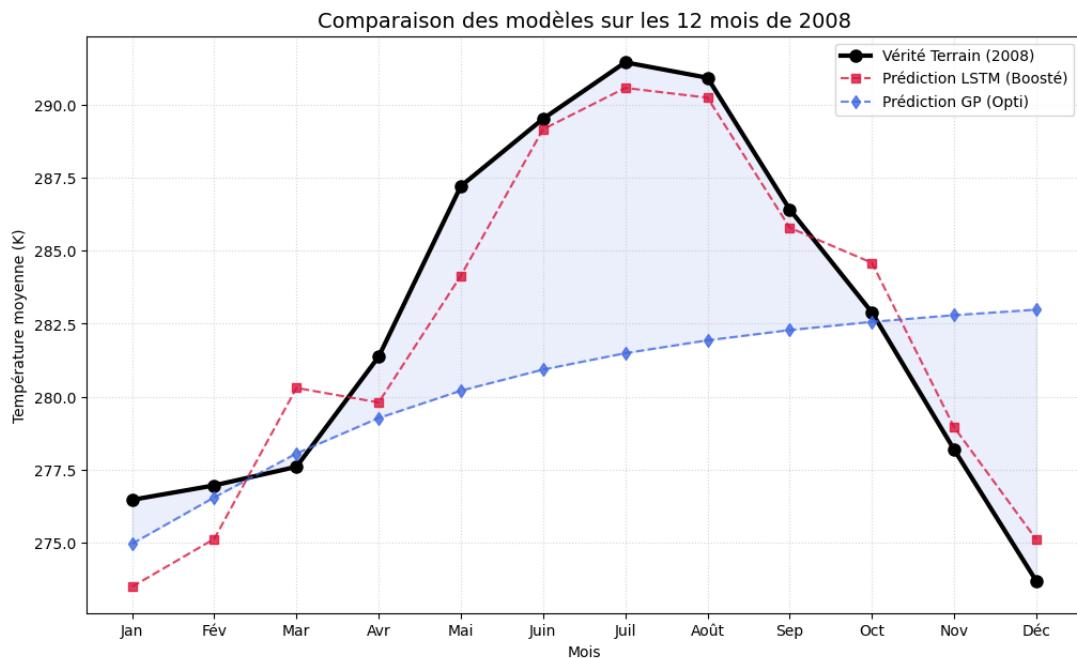


FIGURE 21 – Comparaison des modèles sur les 12 mois de 2008, l'année prédictive.

Pour appréhender la cohérence spatio-temporelle globale, la **Figure 22** projette les données dans un espace tridimensionnel où l'axe Z représente la température SKT. La vue de gauche (*RÉALITÉ 2008*) montre une stratification nette des points colorés par mois, formant des couches distinctes qui témoignent de la progression logique des saisons.

La vue de droite (*PRÉDICTION 2008* par GP) confirme l'analyse précédente : la structure est anormalement compacte. On ne distingue plus les variations d'altitude thermique entre les mois d'hiver (violet/bleu) et les mois d'été (jaune). Cette « compression » du volume thermique montre que le GP échoue à reconstruire la profondeur temporelle du signal, transformant un cycle dynamique en un bloc statique peu représentatif des réalités climatiques.

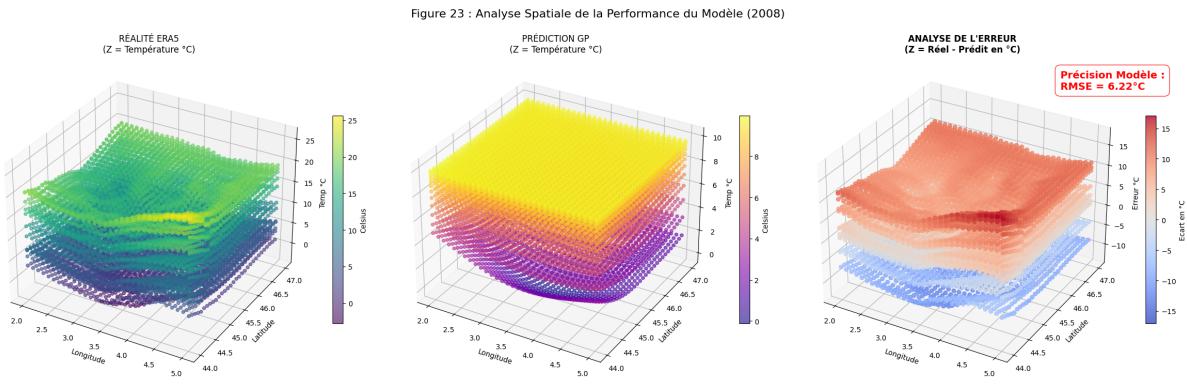


FIGURE 22 – Visualisation 3D (Lon, Lat, Temp) de la réalité 2008 vs la prédiction GP et l’erreur entre les deux.

2.3.4 Cartographie des résidus et analyse spatio-temporelle des erreurs

Pour parachever cette étude, il est impératif d’analyser la distribution spatiale des erreurs afin de comprendre comment chaque modèle se comporte géographiquement face aux spécificités topographiques de l’Auvergne. Cette étape de cartographie des résidus (valeur réelle - valeur prédictive) permet d’identifier les zones de fragilité prédictive et de valider la robustesse locale des architectures.

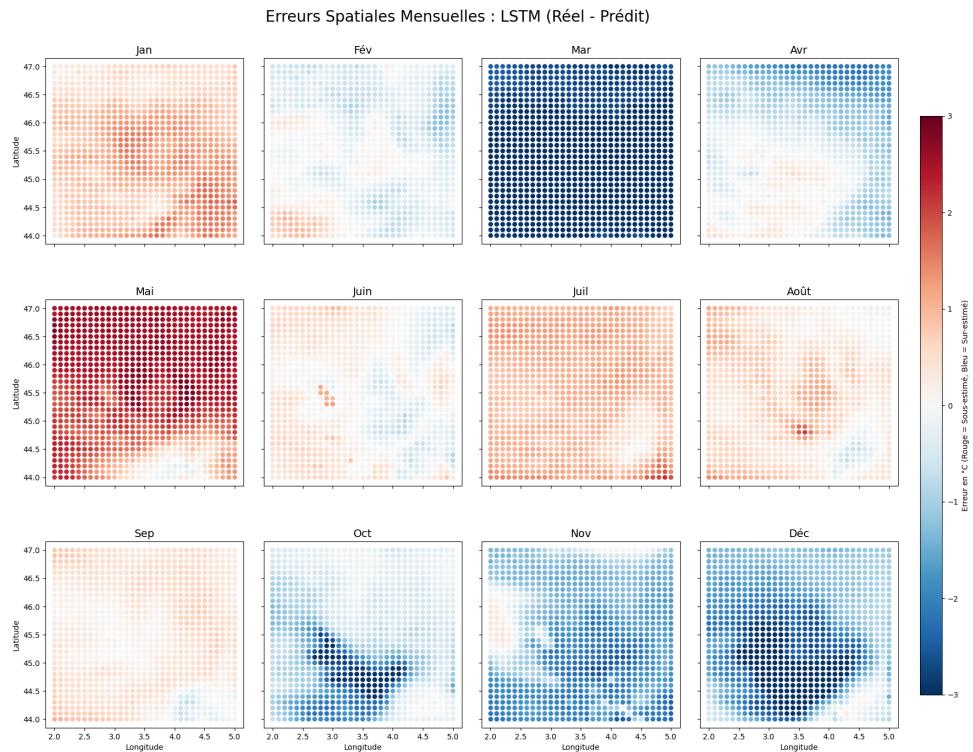


FIGURE 23 – Cartographie mensuelle des erreurs du modèle LSTM pour l’année 2008.

L'examen des résidus du **LSTM**, illustré par la **Figure 23**, révèle une stabilité prédictive remarquable tout au long de l'année. Sur la majeure partie des mois, l'erreur se maintient dans une fourchette étroite comprise entre **-1°C et +1°C**, comme en témoigne la prédominance des teintes claires et neutres sur les cartes. Cette homogénéité spatiale confirme que le réseau récurrent a non seulement appris la dynamique temporelle globale, mais qu'il parvient également à l'adapter aux différentes coordonnées géographiques du département.

On note cependant un résidu plus marqué durant le mois de mai (tons rouges), indiquant une légère sous-estimation de la température de surface. Ce phénomène suggère que des événements météorologiques atypiques ou des transitions thermiques brutales survenues ce mois-là ont partiellement dépassé la capacité d'anticipation de la mémoire à long terme du réseau.

À l'opposé, la **Figure 24** met en exergue l'échec structurel du **Processus Gaussien** à gérer la saisonnalité sur des données réelles. Les erreurs observées sont massives, systématiques et hautement polarisées selon les saisons. En hiver (janvier à mars), le modèle surestime lourdement les températures (tons bleus foncés), tandis qu'en été (mai à août), il les sous-estime de manière critique avec des écarts dépassant souvent les **3°C** (tons rouge sombre).

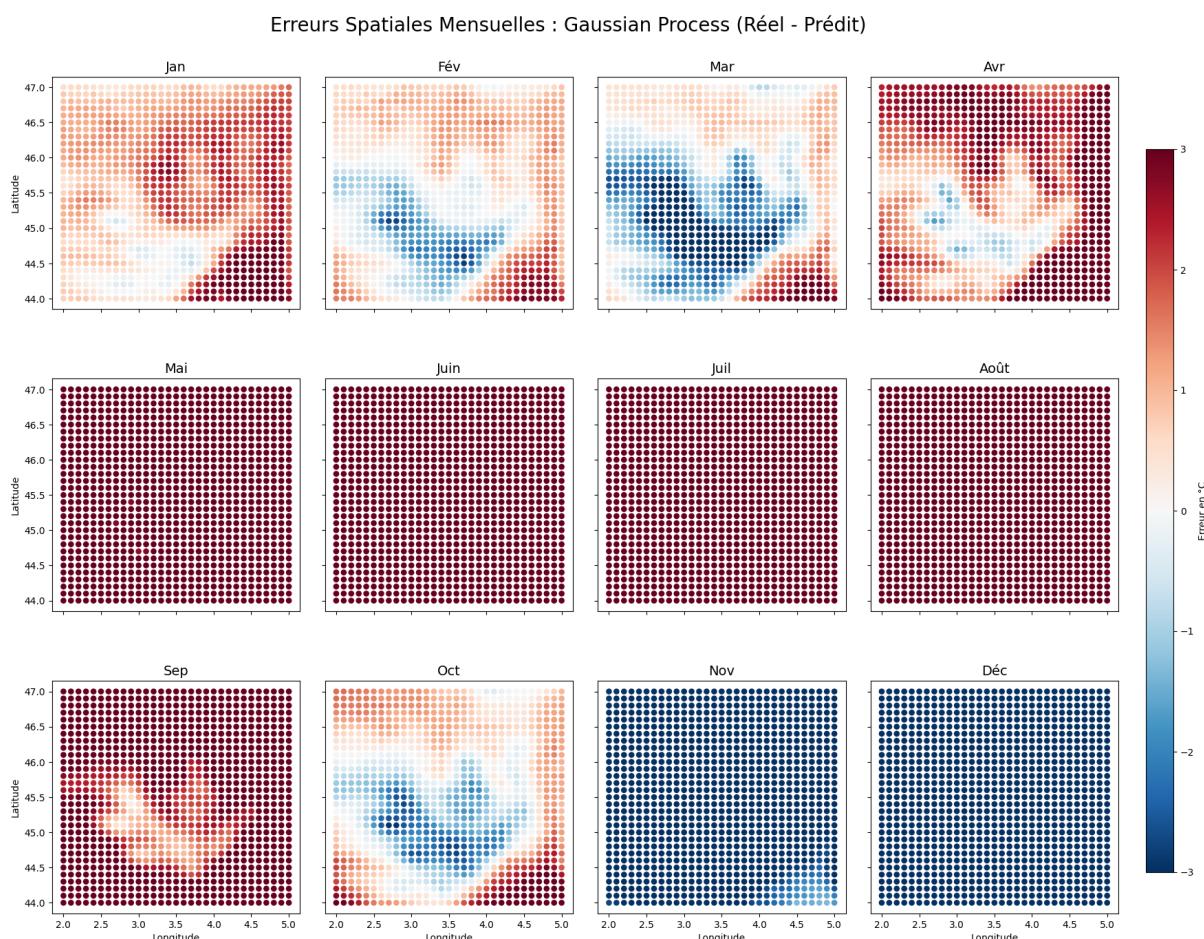


FIGURE 24 – Cartographie mensuelle des erreurs du modèle Gaussian Process (GP).

En synthèse, l'importante dérive saisonnière observée souligne une limite structurelle majeure du modèle par **Processus Gaussien (GP)** : son absence de mémoire temporelle. En se comportant comme un échantillonneur statique, le GP réalise une forme de « régression vers la moyenne » du jeu d'entraînement. En d'autres termes, le modèle privilégié statistiquement des valeurs médianes, ce qui le rend incapable de capturer les extrêmes thermiques, tels que les pics de chaleur ou les gels tardifs, pourtant essentiels à l'analyse climatique.

Techniquement, le GP ignore l'inertie thermique et les corrélations temporelles entre les relevés successifs. Dans le cadre de ce projet, une telle imprécision est préjudiciable : une sous-estimation de 3°C durant la période estivale fausse l'évaluation du stress thermique réel subi par les parcelles forestières d'Auvergne. Là où la réalité du terrain affiche des valeurs critiques, le modèle GP fournit une estimation lissée qui occulte les variations locales de température de surface (SKT).

À l'inverse, l'architecture **LSTM** (Long Short-Term Memory) s'est avérée parfaitement adaptée à la structure séquentielle des données. En traitant chaque mois comme une étape dépendante des précédentes, le réseau récurrent parvient à modéliser l'accumulation de chaleur et les cycles saisonniers avec une grande fidélité. Cette capacité à maintenir un état interne (cell state) lui permet de suivre la courbe thermique réelle avec une précision bien supérieure, là où les méthodes stochastiques classiques ne produisent qu'une vision moyennée et imprécise du signal.

Finalement, cette comparaison démontre que la modélisation climatique nécessite des outils capables d'intégrer la continuité du temps. En privilégiant le LSTM, nous avons sélectionné un modèle capable de fournir des indicateurs thermiques robustes et précis. Ces résultats constituent une base technique fiable pour tout outil d'aide à la décision visant le monitoring des écosystèmes, en offrant une vision dynamique indispensable à la détection des anomalies de température.

2.4 Visualisation cartographique spatio-temporelle grâce aux Heatmaps

Pour transformer nos résultats numériques en outils d'aide à la décision exploitables, nous avons développé un pipeline de génération de **cartes de chaleur (Heatmaps)** interactives. Cette étape permet de confronter visuellement la précision géographique de nos modèles sur les quatre départements de l'ancienne région Auvergne : l'Allier (03), le Cantal (15), la Haute-Loire (43) et le Puy-de-Dôme (63).

2.4.1 Architecture et préparation des données spatiales

Le déploiement opérationnel de notre pipeline de modélisation repose sur une structure de fichiers rigoureuse, conçue pour isoler les flux de données historiques (2000-2007) des projections prédictives de l'année test. Cette organisation permet une automatisation fluide des scripts de rendu cartographique et garantit la traçabilité des versions de modèles (**cf . Figure 25**).

```
Projet_5A/
|-- cartes_climat/
|   |-- historique/
|   |   |-- hist_2000-01.html
|   |   |-- ...
|   |   |-- hist_2007-12.html
|   +-- predictions_2008/
|       |-- MAP_PRED_2008-01.html      (Prediction GP / LSTM)
|       |-- MAP_REEL_2008-01.html      (Ground truth)
|       |-- ...
|       +-- MAP_REEL_2008-12.html
```

FIGURE 25 – Architecture logicielle : Organisation hiérarchique des répertoires pour le stockage automatisé des heatmaps interactives, incluant l'historique climatique (2000–2007) et la comparaison prédictions vs données réelles pour l'année 2008.

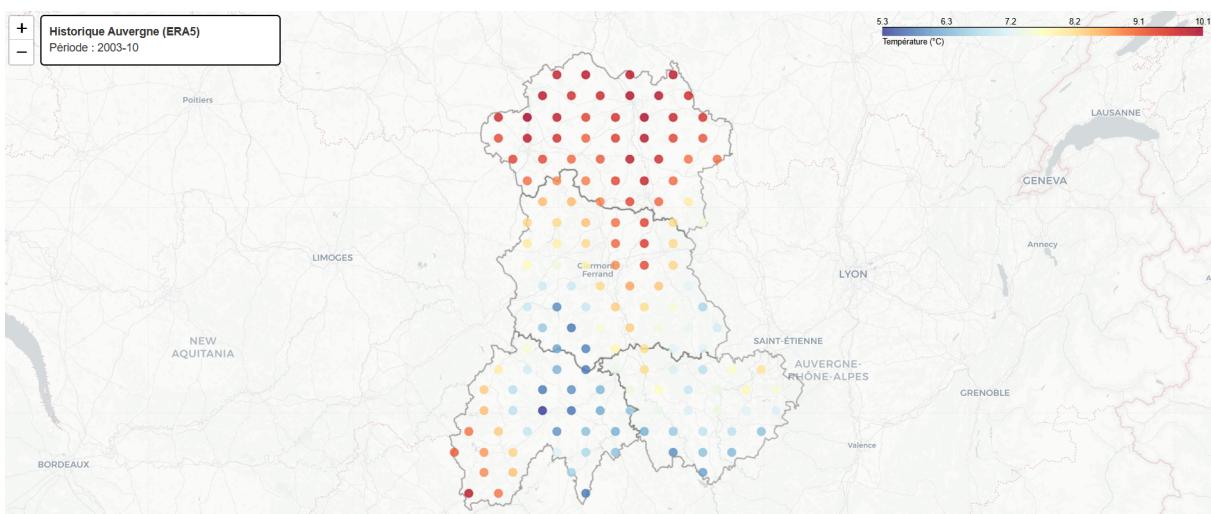


FIGURE 26 – Exemple de Heatmap mensuelle générée via Folium

L'interface *Folium* (**Figure 26**) génère des rendus interactifs haute résolution, permettant une exploration granulaire des températures de surface converties en degrés Celsius pour une interprétation physique optimale par les gestionnaires forestiers.

2. Implémentation du Processus Gaussien (GP) : La modélisation par GP a été stabilisée pour traiter les données réelles via les étapes suivantes :

- **Feature Engineering :** Création d'un encodage cyclique pour le temps (*month_sin/cos*) et d'un index temporel continu pour capturer la linéarité des tendances.
- **Optimisation du Noyau :** Utilisation d'un noyau composite : *ConstantKernel * ExpSineSquared* (pour la saisonnalité) + *Matern* (pour les variations locales) + *WhiteKernel* (pour le bruit).
- **Stabilisation Numérique :** Introduction d'un paramètre $\alpha = 10^{-2}$ pour stabiliser la matrice de covariance lors de l'ajustement sur les 1500 points d'échantillonnage.

3. Architecture du modèle LSTM : En parallèle, le réseau de neurones récurrent a été configuré pour exploiter la mémoire séquentielle :

- **Fenêtrage (Window Size) :** Utilisation d'un historique de 12 mois pour prédire le mois suivant, permettant de capturer les cycles annuels complets.
- **Structure du réseau :** Deux couches LSTM (64 et 32 unités) avec activation *tanh*, suivies d'une couche Dense pour la régression.
- **Normalisation :** Application rigoureuse de *StandardScaler* sur la température (SKT) et les coordonnées géographiques avant l'entraînement sur la période 2000-2007.

2.4.2 Analyse comparative spatiale : Réalité vs Prédictions (Juillet 2008)

Pour valider l'utilité opérationnelle de nos modèles sur le terrain, nous avons sélectionné le mois de **juillet 2008** comme cas d'étude. Ce choix est stratégique : il correspond au pic de stress thermique estival en Auvergne, période durant laquelle la précision de la température de surface (SKT) est vitale pour la surveillance des écosystèmes forestiers.

A. État de référence : La réalité terrain (ERA5-Land)

La carte de référence issue du jeu de données ERA5-Land (cf. **Figure 27**) montre une distribution thermique fortement influencée par la topographie auvergnate. On observe des températures de surface oscillant entre **14,0 °C et 19,3 °C**. Les zones de chaleur intense se concentrent dans les plaines de la Limagne et le nord de l'Allier, tandis que les massifs du Cantal et du Puy-de-Dôme conservent une relative fraîcheur. À un point de contrôle précis (44.90, 2.10), la température réelle enregistrée est de **17,74 °C**.

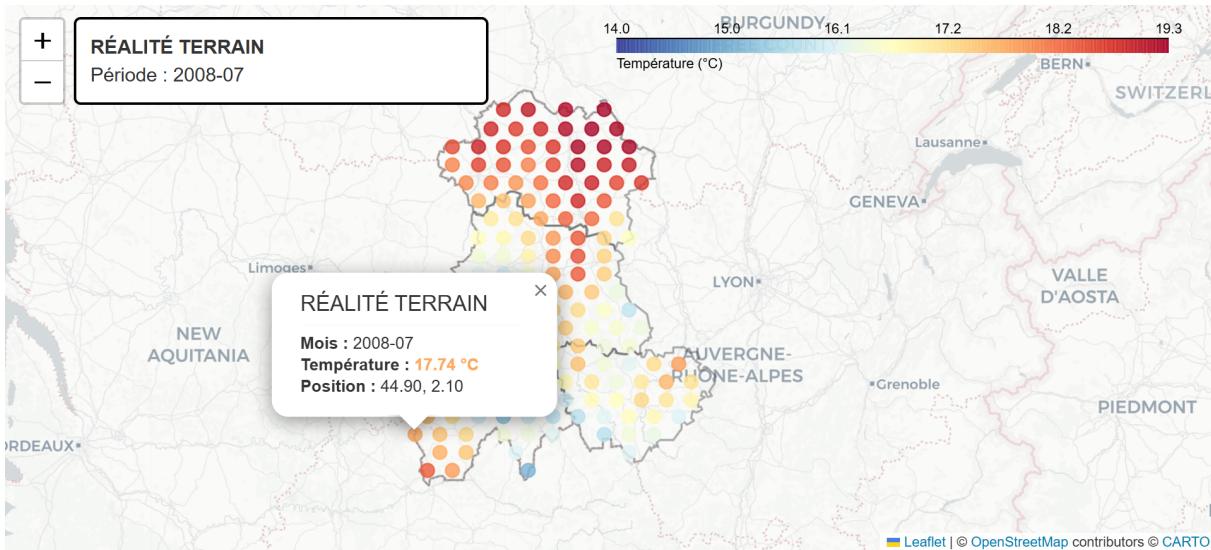


FIGURE 27 – Réalité terrain ERA5-Land pour juillet 2008. La palette de couleurs illustre le gradient thermique naturel lié au relief auvergnat.

B. Prédiction par apprentissage automatique : Le modèle LSTM

Le modèle **LSTM** (cf. **Figure 28**), en exploitant sa mémoire des cycles saisonniers appris sur la période 2000-2007, produit une cartographie d'une fidélité remarquable. La structure spatiale des îlots de chaleur est quasi identique à la réalité terrain. Bien que l'échelle des valeurs prédites soit légèrement plus resserrée (**12,7 °C à 17,4 °C**), le modèle capture parfaitement les nuances géographiques. Cette performance confirme que l'architecture récurrente a assimilé la corrélation entre les coordonnées spatiales et l'inertie thermique des sols.

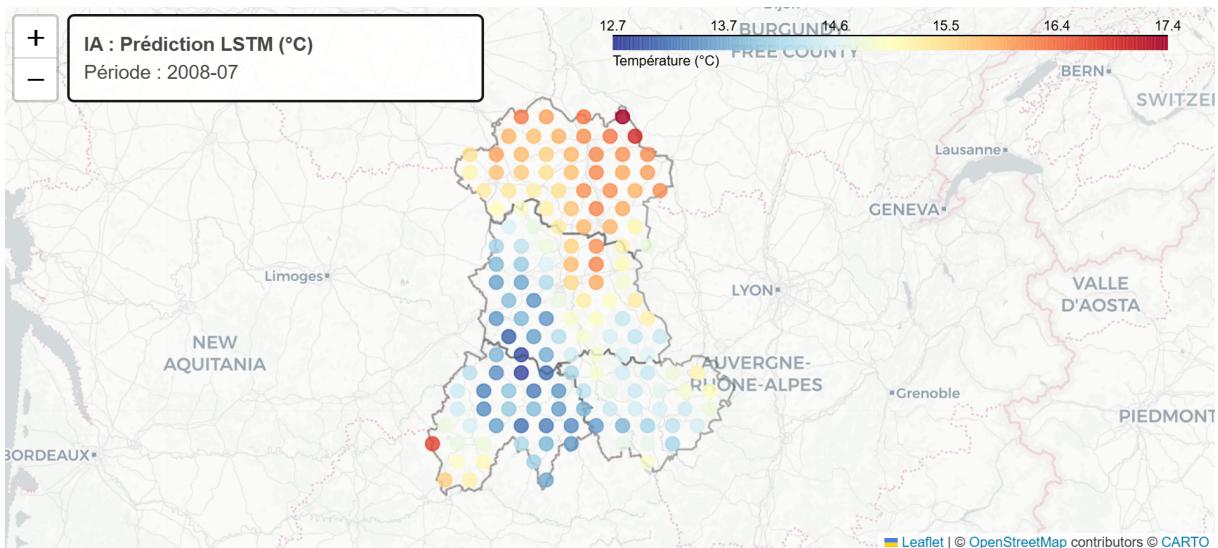


FIGURE 28 – Prédiction IA par réseau LSTM. On note la conservation des structures spatiales fines et la précision des gradients départementaux.

C. Prédiction Stochastique : Le Processus Gaussien (GP)

L'analyse de la prédiction par **Processus Gaussien** révèle les limites critiques de cette approche pour l'extrapolation temporelle réelle. Si le GP identifie le gradient général (nord plus chaud), il échoue massivement sur l'amplitude thermique. Les températures prédictives plafonnent entre **7,14 °C et 7,62 °C**, soit un écart de plus de **10 °C** avec la réalité. Le modèle, dépourvu de mémoire séquentielle, subit un effet de « régression vers la moyenne » : il prédit une valeur statistiquement lissée sur l'historique global plutôt que de capturer le pic saisonnier de juillet (cf. **Figure 29**).

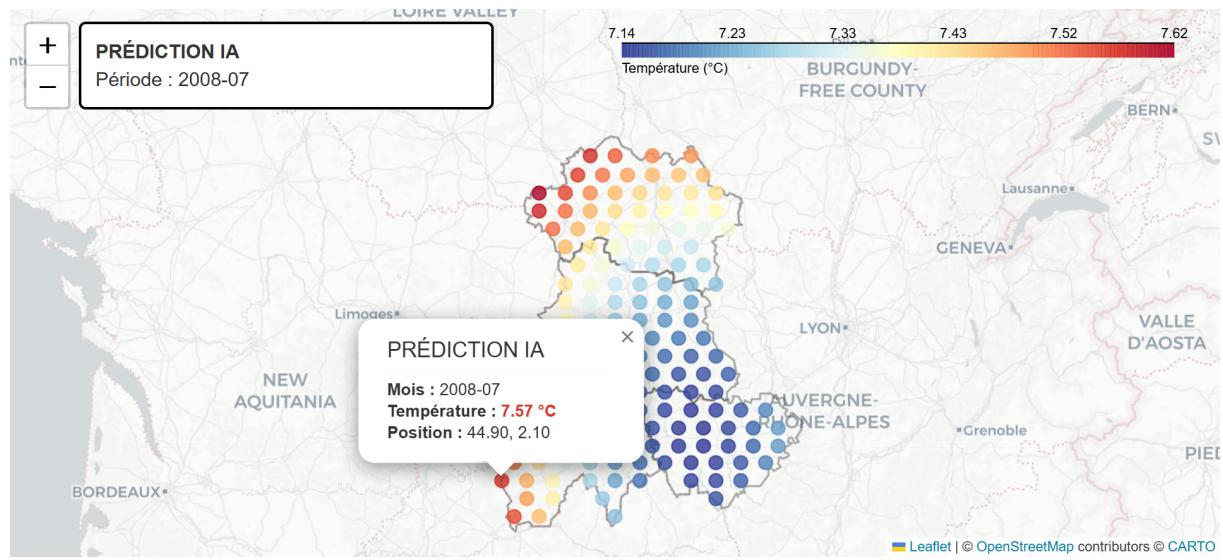


FIGURE 29 – Prédiction par Processus Gaussien.

D. Synthèse et confrontation visuelle synoptique

La confrontation des trois états de la donnée permet de valider notre choix technologique final :

- **Fidélité physique** : Seul le LSTM maintient une cohérence physique avec les températures réelles, indispensable pour détecter les seuils de stress hydrique.
- **Robustesse temporelle** : La défaillance du GP sur l'amplitude montre qu'un modèle statique ne peut anticiper des extrêmes climatiques sans une architecture capable de traiter le temps comme une dimension récurrente.
- **Aide à la décision** : Pour un gestionnaire forestier, la carte LSTM identifie les zones de danger thermique réel, tandis que la carte GP masquerait tout risque en lissant les températures anormalement vers le bas.

En résumé, ces heatmaps constituent la preuve par l'image que le **LSTM est le modèle de référence** pour la surveillance climatique en Auvergne. Il offre le compromis idéal entre résolution spatiale fine et précision temporelle saisonnière.

2.5 Synthèse globale de l'étude : De l'acquisition climatique à la décision forestière

La troisième partie de cette étude a marqué le passage crucial de la validation théorique sur signaux simulés à l'application concrète sur les écosystèmes forestiers de la région Auvergne. Cette démarche, structurée autour du jeu de données **ERA5-Land**, a permis de confronter nos outils de modélisation à la complexité de la réalité de terrain.

1. Maîtrise de la donnée et ingénierie spatiale : L'étude a débuté par une phase rigoureuse d'extraction via l'API **Copernicus (CDS)**, permettant de consolider huit années de relevés thermiques (2000-2007). Le passage de données tabulaires brutes à des **GeoDataFrames** a été l'étape pivot permettant de contextualiser géographiquement le stress thermique subi par les parcelles forestières. Cette phase a abouti à la création d'une architecture logicielle automatisée capable de générer des **heatmaps interactives** via *Folium*, convertissant les données physiques en indicateurs Celsius directement interprétables par les gestionnaires.

2. Confrontation des paradigmes de modélisation : Le cœur de cette section a reposé sur la mise en concurrence de deux approches technologiques majeures pour prédire l'année 2008 :

- **Le Processus Gaussien (GP) :** Bien qu'efficace pour l'interpolation spatiale à un instant T , il a montré des limites critiques lors de l'extrapolation temporelle. Son absence de mémoire séquentielle l'empêche de suivre la dynamique cyclique des saisons, conduisant à une erreur de **6,14°C (RMSE)** et une incapacité flagrante à capturer les pics de chaleur.
- **Le Réseau LSTM (Boosté) :** À l'inverse, l'architecture récurrente s'est révélée être le moteur le plus performant. En apprenant l'inertie thermique et les cycles mensuels, il atteint un score de précision remarquable ($R^2 = 0,938$) avec une erreur résiduelle contenue à seulement **1,53°C**.

3. Validation par les Heatmaps : Le choix de la précision spatiale : La confrontation visuelle des cartes de chaleur a constitué l'arbitrage final. En comparant la réalité terrain avec les prédictions pour juillet 2008, nous avons observé que le **LSTM reproduit fidèlement les îlots de chaleur** des plaines de la Limagne et les gradients d'altitude. À l'opposé, les heatmaps du GP ont révélé un lissage excessif et une sous-estimation massive des températures (écart supérieur à 10°C), rendant ce modèle inapte à la détection des seuils de danger thermique en forêt.

En conclusion, si le Processus Gaussien demeure un allié précieux pour l'analyse statique, la complexité des changements climatiques impose le recours à l'**apprentissage profond séquentiel**. Le **LSTM s'impose comme le modèle de référence** : il est le seul capable de fournir une vision cartographique dynamique et fiable, indispensable pour anticiper les stress hydriques et protéger durablement nos écosystèmes fragiles.

Conclusion Générale

Au terme de ce projet tutoré de cinquième année à **Polytech Clermont**, nous avons pu concevoir et valider une méthodologie complète dédiée à la modélisation de l'évolution thermique des écosystèmes forestiers. Ce travail, à la frontière entre l'**Intelligence Artificielle** et les **Systèmes d'Information Géographique (SIG)**, a permis de démontrer comment l'apprentissage automatique peut transformer des données brutes en indicateurs stratégiques pour la préservation de la biodiversité en Auvergne.

Le premier pilier de ce projet a reposé sur une phase fondamentale de **Data Engineering**. L'extraction de données massives via l'**API ERA5-Land** et leur structuration spatiale sous *GeoPandas* ont constitué un défi technique majeur, soulignant l'importance cruciale de la préparation des données dans tout pipeline d'IA. Cette étape a permis de contextualiser le stress thermique subi par les parcelles, révélant la complexité des interactions entre la topographie locale et les dynamiques climatiques globales.

Le cœur scientifique de l'étude a résidé dans la confrontation de deux paradigmes algorithmiques. Si les **Processus Gaussiens (GP)** ont confirmé leur efficacité pour l'interpolation spatiale statique, ils ont montré leurs limites face à la non-stationnarité des cycles temporels réels. À l'inverse, l'utilisation de réseaux de neurones récurrents **LSTM** s'est révélée être le choix le plus performant. En capturant les dépendances séquentielles sur huit années de relevés, le modèle LSTM a atteint une précision remarquable ($R^2 = 0,938$), validant la supériorité du **Deep Learning** pour la prédiction climatique à long terme.

Au-delà des résultats numériques, ce projet a abouti à une application concrète via la génération de **heatmaps interactives**. Ces outils de visualisation offrent aux gestionnaires forestiers une vision granulaire et dynamique du terrain, indispensable pour anticiper les risques de dépérissement liés aux canicules. La confrontation visuelle entre la réalité terrain et les prédictions IA a définitivement assis le LSTM comme le modèle de référence pour le suivi des stress thermiques.

Sur un plan personnel et professionnel, ce projet de recherche a été extrêmement **formateur**. Il m'a permis d'approfondir mes compétences en ingénierie des données et en apprentissage profond, tout en m'initiant aux rigueurs de l'expérimentation scientifique. La gestion de l'incertitude des modèles et la nécessité d'interpréter physiquement des résultats mathématiques ont constitué une expérience précieuse pour ma future carrière d'ingénieur.

En conclusion, cette étude ouvre des perspectives prometteuses pour le monitoring environnemental de précision. L'intégration future d'autres variables, telles que l'indice de sécheresse ou l'âge des peuplements, pourrait encore affiner la compréhension de la dynamique des espèces forestières et ainsi contribuer plus efficacement à la résilience de nos forêts face au défi climatique.

Références

- [1] **ERA5-Land** : *Copernicus Climate Change Service (C3S) - Données horaires de 1950 à nos jours.* Copernicus Climate Data Store (CDS). Disponible sur : <https://cds.climate.copernicus.eu/>. Consulté en novembre 2025.
- [2] **Processus Gaussiens** : Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, MIT Press. Consulté en septembre 2025.
- [3] **LSTM (Long Short-Term Memory)** : Hochreiter, S., & Schmidhuber, J. (1997). *Neural Computation*, 9(8), 1735-1780. Consulté en décembre 2025.
- [4] **VS Code (Visual Studio Code)** : Environnement de développement intégré (IDE) pour le développement Python. Disponible sur : <https://code.visualstudio.com/>. Consulté le 26 janvier 2026.
- [5] **Jupyter Notebook** : Interface web interactive pour l'informatique scientifique et l'analyse de données. Disponible sur : <https://jupyter.org/>. Consulté le 26 janvier 2026.
- [6] **GitHub** : *Dépôt officiel du projet - Modélisation de l'évolution temporelle des espèces forestières par IA.* Ayman Zejli, Projet 5A Polytech. Disponible sur : <https://github.com/Aymanzej/Projet-5A>. Consulté le 26 janvier 2026.
- [7] **Scikit-learn** : Outils de data science pour les Processus Gaussiens et le prétraitement des données (StandardScaler). Disponible sur : <https://scikit-learn.org/>. Consulté en septembre 2026.
- [8] **TensorFlow** : Bibliothèque d'apprentissage automatique utilisée pour l'implémentation du réseau de neurones récurrent LSTM. Disponible sur : <https://www.tensorflow.org/>. Consulté en septembre 2025.
- [9] **Pandas** : Bibliothèque de manipulation et d'analyse de structures de données tabulaires. Disponible sur : <https://pandas.pydata.org/>. Consulté en novembre 2025.
- [10] **GeoPandas** : Extension de Pandas dédiée à la manipulation de données géospatiales et aux Systèmes d'Information Géographique (SIG). Disponible sur : <https://geopandas.org/>. Consulté en novembre 2025.
- [11] **Cartiflette** : Bibliothèque de récupération des fonds de cartes officiels et découpage administratif (IGN/INSEE). Disponible sur : <https://pypi.org/project/cartiflette/>. Consulté le 26 janvier 2026.
- [12] **Folium** : Bibliothèque de visualisation de données géospatiales sur des cartes interactives Leaflet.js. Disponible sur : <https://python-visualization.github.io/folium/>. Consulté en novembre 2025.