

Rapport de stage de recherche – Utilisation des réseaux de neurones artificiels pour l'estimation des flux en surface

Laboratoire : Laboratoire d'Aérodynamique (LAERO, CRA)

Responsables de stage : Fabienne LOHOU, Marie LOTHON

REMERCIEMENTS

Je tenais tout d'abord à remercier Fabienne Lohou et Marie Lothon pour l'organisation de ce stage mais aussi et surtout pour leur encadrement et leurs conseils tout du long. Je remercie également toute l'équipe du Centre de Recherche Atmosphérique pour leur accueil ainsi que pour le partage de leurs compétences. Finalement, je tiens à remercier Jason Kelley et Eric Pardyjak qui se sont rendus disponibles tout du long de ce stage afin de partager avec nous leur savoir et leur expérience.

RÉSUMÉ

La surface terrestre et ses propriétés impactent à différentes échelles l'atmosphère. La compréhension des échanges surface-atmosphère joue donc un rôle important dans les modèles de prévision météorologiques et climatiques. Cependant, la mauvaise représentation des flux de chaleur sensible et de chaleur latente en surface est devenue une problématique majeure dans la modélisation atmosphérique et climatique. La réduction des biais engendrés par cette mauvaise représentation est l'un des principaux objectifs du projet MOSAI dans lequel ce stage est ancré. Une des raisons qui pourraient expliquer ces biais est la mauvaise représentation des hétérogénéités de surface et donc des flux. Appréhender expérimentalement ces hétérogénéités s'avère être coûteux si l'on utilise des mesures de référence dites « d'eddy-covariance ». Ces dernières s'appuient sur des mesures très précises et acquises à haute fréquence. Mais de récentes études ont montré qu'en appliquant des méthodes d'intelligence artificielle à des stations météorologiques de surface, il est possible d'obtenir une bonne estimation de ces flux à moindre coût.

Dans le cadre de ce stage, l'objectif premier est de tester cette nouvelle méthode d'intelligence artificielle. Nous nous appuyons sur deux jeux de données différents. Dans un premier temps, nous déterminons s'il est possible d'estimer des flux en surface sur une courte période de temps (douze jours discontinus de beau temps entre juin et juillet). Nous réalisons une série de tests afin de déterminer l'influence du jeu d'apprentissage, de la topographie du réseau et des variables mises en entrée sur le réseau de neurones étudié. La conclusion principale à tirer de ces premiers tests est l'importance du choix du jeu d'apprentissage comparé à celui de la topographie ou des variables d'entrée. Dans un second temps, afin d'aller plus loin dans cette étude, nous commençons à appliquer cette même méthode mais sur un jeu de données plus long (une année complète de données). L'objectif final serait de parvenir à appliquer une telle méthode sur le site de Lannemezan dans le cadre du projet MOSAI, afin d'estimer de façon continue et simultanée les flux de surface sur différents couverts végétaux à moindre coût.

TABLE DES MATIÈRES

Remerciements.....	2
Résumé.....	3
Table des matières.....	4
1. Introduction.....	5
1.1. Les flux en surface.....	5
1.1. Mesure des flux en surface : une nouvelle approche.....	6
1.3. Résultats des précédentes études.....	6
1.4. Objectifs du stage.....	7
2. Données et stratégie.....	9
2.1. Les réseaux de neurones artificiels (ANN).....	9
2.2. Données disponibles	10
2.2.1. La campagne BLLAST : jeu de données sur 3 semaines.....	10
2.2.2. Petit mât de 2m instrumenté : jeu de données sur une année	11
2.3. Définition des différents tests.....	12
2.3.1. Jeu d'apprentissage : tests tsl (Tableau 1).....	12
2.3.2. Topographie du réseau de neurones : tests ntt (Tableau 1).....	12
2.3.3. Variables en entrée de notre réseau : test ivt (Tableau 1).....	12
2.4. Critères de comparaison des différents tests.....	13
2.5. Statistiques.....	14
2.5.1. Répétitions indépendantes des tests.....	14
2.5.2. Significativité du coefficient R.....	14
3. Résultats.....	16
3.1. MOOR/CORN : tests pour la lande et le maïs sur une courte durée.....	16
3.1.1. Tests tsl pour la lande : sensibilité au jeu d'apprentissage.....	16
3.1.2. Tests tsl pour le maïs : sensibilité au jeu d'apprentissage	17
3.1.3. Tests ntt pour la lande : influence de la topographie du réseau de neurones.....	17
3.1.4. Tests ivt pour la lande : sensibilité aux variables en entrée de notre réseau.....	19
3.1.5. Tests ivt pour le maïs : sensibilité aux variables en entrée de notre réseau.....	20
3.2. Étude d'un jeu de données sur une année.....	21
3.2.1. Test tsl pour un an de données : sensibilité au jeu d'apprentissage.....	21
3.2.2. Test ntt pour un an de données : influence de la topographie du réseau de neurones	22
4. Discussion.....	24
4.1. CORN/MOOR : discussion des résultats.....	24
4.2. Jeu de données sur une année : discussion des résultats.....	24
5. Conclusion.....	26
6. Conclusion personnelle.....	27
Bibliographie.....	28
Annexes.....	29
Annexe A : Fiche résultat type réalisée pour chaque test.....	29
Annexe B : Tableaux récapitulatifs des différents jeu de tests réalisés.....	32
Annexe C : Récapitulatif des résultats obtenus pour le jeu de tests ntt, pour le maïs (flux de chaleur sensible).	33

1. INTRODUCTION

En février 2019, une étude réalisée par le Groupe de Travail sur l'Expérimentation Numérique (WGNE) a établi que le biais sur les flux de chaleur sensible et latente en surface est la deuxième plus grande source d'erreurs dans les modèles numériques globaux et régionaux [1]. Cette source d'erreur est notamment due à la mauvaise représentation de l'hétérogénéité des surfaces dans ces modèles. Afin de quantifier l'ampleur des hétérogénéités de flux dans la maille des modèles (de 1.3 à 50km selon les modèles) dans le but de mieux les évaluer, il faudrait intensifier les mesures de ces flux en surface.

1.1. Les flux en surface

La surface terrestre et ses propriétés impactent à différentes échelles l'atmosphère, d'où l'importance de la compréhension des échanges surface-atmosphère dans les modèles de prévision météorologique et climatique. Ces échanges entre la surface et l'atmosphère, appelés flux, caractérisent un transport dû à la présence de gradients et dépendent de la surface considérée (paramètres de surface, variables de surface).

Le flux de chaleur sensible (ϕ_H) représente la quantité de chaleur échangée par convection entre la surface et l'atmosphère. Cela correspond donc à un gradient de température. Ce processus réchauffe les basses couches atmosphériques et la convection permet ensuite un transport vertical de chaleur des basses couches vers les hautes couches.

$$\phi_H = \rho C_p \overline{w'T'} = -\rho C_p K_T \frac{\delta T}{\delta z}$$

avec ρ : masse volumique de l'air sec

$$C_p = 1005 \text{ J.kg}^{-1}.\text{K}^{-1}$$

K_T : coefficient d'échange turbulent (température) en $\text{m}^2.\text{s}^{-1}$

$\overline{w'T'}$: covariances vitesse verticale et température

$\frac{\delta T}{\delta z}$: gradient vertical de température dans l'atmosphère

Le flux de chaleur latente (ϕ_{LE}) correspond à l'énergie échangée lors d'un changement d'état de l'eau entre la surface et l'atmosphère (évaporation de l'eau en surface). Cet échange est aussi nommé évapotranspiration (ET) et décrit le transport de l'eau depuis la surface vers l'atmosphère (gradient d'humidité). C'est donc la somme de l'évaporation directe de l'eau du sol et de la transpiration par les plantes.

$$\phi_H = \rho L_V \overline{w'q'} = -\rho L_V K_q \frac{\delta q}{\delta z}$$

avec L_V : chaleur latente de vaporisation = $2,45.10^6 \text{ J.kg}^{-1}$

K_q : coefficient d'échange turbulent (humidité) en $\text{m}^2.\text{s}^{-1}$

$\overline{w'q'}$: covariances vitesse verticale et humidité

$\frac{\delta q}{\delta z}$: gradient vertical d'humidité dans l'atmosphère

Afin de définir le bilan d'énergie à la surface terrestre, il nous faut introduire le flux dans le sol ϕ_{sol} ainsi que le rayonnement net R_n . ϕ_{sol} représente le flux de conduction de chaleur dans le sol, défini comme :

$$\varphi_{\text{sol}} = -k_g \frac{\delta T}{\delta z}$$

avec k_g : conductivité thermique du sol en $\text{W.m}^{-1}.\text{K}^{-1}$

Le rayonnement net, R_n , représente la quantité d'énergie radiative disponible à la surface de la terre. Il peut être défini comme la somme algébrique des rayonnements reçus et perdus par la surface :

$$R_n = (\varphi_{\text{solaire}\downarrow} - \varphi_{\text{solaire}\uparrow}) + (\varphi_{\text{atmosphérique}\downarrow} - \varphi_{\text{tellurique}\uparrow})$$

Finalement, nous pouvons définir le bilan d'énergie à la surface de la terre comme étant :

$$R_n = \varphi_H + \varphi_{LE} + \varphi_{\text{sol}}$$

1.1. Mesure des flux en surface : une nouvelle approche

Les covariances $\overline{w'T'}$ et $\overline{w'q'}$ définies respectivement dans les formules des flux de chaleur sensible et de chaleur latente sont estimées grâce à des capteurs à hautes fréquences permettant de couvrir toutes les échelles fréquentielles de la turbulence contribuant à ces échanges convectifs. Ces dispositifs sont appelés stations d'eddy-covariance. Mais, ces stations d'eddy-covariance sont peu déployées car particulièrement coûteuses du fait qu'elles nécessitent des capteurs avec une haute fréquence d'acquisition.

Cependant de récentes études, que nous évoquerons plus en détails dans la partie suivante, ont permises de mettre en avant une technique moins coûteuse permettant l'estimation de ces flux. Cette nouvelle technique se base sur l'intelligence artificielle et notamment les réseaux de neurones artificiels (ANN) appliqués à des mesures météorologiques acquises à faible fréquence. En effet, il est possible d'entraîner le réseau de neurones sur un jeu de données issu de stations météorologiques classiques afin d'estimer des flux de chaleur sensible et de chaleur latente. L'entraînement nécessite tout de même un jeu de flux de référence qui sera fourni par une station d'eddy-covariance déplacée de station météorologique en station météorologique selon un protocole qu'il sera nécessaire de déterminer (**Fig. 1**).

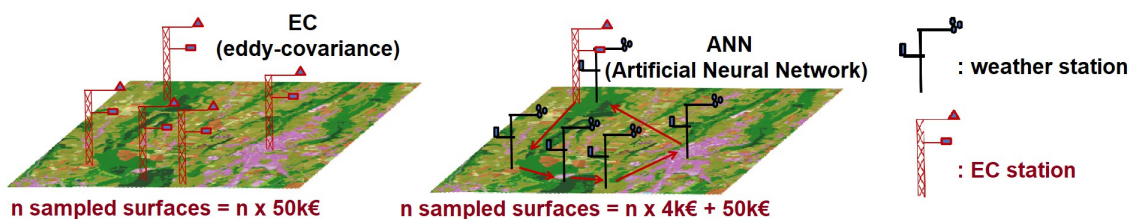


Fig. 1: Utilisation des réseaux de neurones artificiels pour estimer les flux en surface. A gauche, ce que coûterait le déploiement de plusieurs stations d'eddy-covariance afin d'intensifier les mesures de ces flux sur les différents types de couverts. A droite, l'utilisation, moins coûteuse, de réseaux de neurones et de stations météorologiques classiques afin d'estimer ces mêmes flux. Une station d'eddy-covariance est cependant nécessaire et sera déplacée de station en station météorologique en station météorologique pour l'apprentissage du réseau de neurones.

L'objectif final serait d'obtenir des stations météorologiques suffisamment bien entraînées de façon à ne plus avoir besoin de la station d'eddy-covariance déployée pour les entraîner.

1.3. Résultats des précédentes études

En 2011, Kumar et al. proposent une étude sur la mise en place d'un réseau de

neurones afin de modéliser l'évapotranspiration [2]. Plusieurs modèles sont discutés et comparés à l'estimation par réseau de neurones. Parmi les différents cas étudiés, l'estimation par ANN est à chaque fois plus efficace que les modèles existants. Mais les problèmes associés à une telle approche sont aussi relevés, notamment le fait que l'ANN reste assimilable à une boîte noire. De ce fait, nous ne pouvons pas récupérer certains paramètres du réseau (et notamment les poids) ce qui empêche la généralisation du réseau à différentes localisations géographiques. Pour utiliser une telle méthode, il faudrait donc faire en sortes d'entraîner le réseau sur chaque zone étudiée, ce qui représente un désavantage comparé aux modèles d'estimation déjà existants.

Ces travaux sont repris quelques années plus tard par Jason Kelley et Eric Pardyjak [3]. Ces derniers ont cherché à utiliser les réseaux de neurones pour estimer, à partir de stations météorologiques classiques, l'évapotranspiration de cultures afin d'estimer leur besoin en eau. Cela permettrait à la communauté agricole d'estimer l'évapotranspiration et ainsi de gérer et optimiser l'apport en eau à partir de stations météorologiques classiques bien plus accessibles que la mesure de flux d'énergie et d'expertise associée. Des résultats concluants ont été obtenus en utilisant seulement quatre paramètres physiques (température, rayonnement solaire, humidité et vitesse du vent) ainsi que le temps sur un temps d'entraînement bref d'une semaine. Ajouter le temps pour estimer une évapotranspiration par ANN peut à premier abord sembler incongru. Mais les flux obéissant à un cycle diurne, la présence du paramètre temps permet une bonne représentation de cette périodicité. A noter aussi que choisir l'humidité du sol comme paramètre du réseau n'améliore pas les résultats, au contraire.

La période d'apprentissage varie entre 1 et 21 jours et a elle aussi été discutée. Des expérimentations sont réalisées sur deux sites différents et les mêmes limites que l'étude précédente [2] ont été relevées : le modèle n'est valable que dans des conditions environnementales et de culture similaires à la période de formation. Une période d'apprentissage de sept jours peut être suffisante pour estimer l'évapotranspiration sur un site mais seulement dans des conditions homogènes.

En 2020, Jason Kelley et al. reprennent ces résultats [4] et cette fois-ci, les données utilisées sont collectées sur une période plus longue que ces précédentes recherches [3] et dans des régions distinctes aux conditions environnementales différentes. Les périodes d'apprentissage du réseau varient entre 1 et 21 jours, mais la période d'application du réseau entraîné est plus longue qu'auparavant (plus ou moins 4 mois, contre moins d'1 mois pour les précédentes études). Les estimations qui en résultent sont robustes mais restent sensibles aux données d'entraînement. Les résultats montrent que cette technique peut être utilisée avec des stations météorologiques classiques pour améliorer les estimations locales de la demande en eau des cultures lorsque des mesures de haute qualité peuvent être colocalisées pendant de courtes périodes.

1.4. Objectifs du stage

Dans le cadre du projet MOSAI, il est primordial de parvenir à estimer les flux sur une mosaïque de surfaces. Mais le coût d'un déploiement de stations de mesures d'eddy-covariance dans un tel contexte n'est pas envisageable. L'enjeu du stage s'inscrit donc dans ce cadre et l'objectif principal est de tester cette nouvelle technique d'estimation des flux autour du Centre de Recherches Atmosphériques.

Il est notamment intéressant de reprendre certains résultats des précédentes études sur l'évapotranspiration afin de les comparer avec ce que nous pouvons obtenir à partir de nos données. L'influence de la période d'apprentissage du réseau de neurones, l'influence de sa

topographie mais aussi l'importance des variables météorologiques utilisées (dont l'humidité du sol) sont des aspects qu'il est important de traiter.

Comme l'objectif à long terme est de se séparer de la station d'eddy-covariance permettant l'entraînement des stations météorologiques classiques, il faut parvenir à définir un protocole d'apprentissage le plus juste et le plus précis possible en définissant, par exemple, quels sont les capteurs importants ou encore comment et combien de temps est-il nécessaire d'entraîner une station météorologique classique avant qu'elle puisse devenir indépendante dans l'estimation des flux en surface tout au long de l'année. L'étude d'un jeu de données d'une année permettrait de commencer à traiter ces objectifs à long terme.

2. DONNÉES ET STRATÉGIE

2.1. Les réseaux de neurones artificiels (ANN)

L'intelligence artificielle est une branche de l'informatique qui vise à simuler les comportements du cerveau humain. En biologie, les neurones sont des cellules du système nerveux spécialisées dans le traitement et la communication d'informations. Ces cellules, une fois stimulées, peuvent entraîner la formation de signaux bioélectriques qui sont ensuite transmis à d'autres neurones via les axones et les synapses. Ce signal peut aussi permettre l'activation d'autres tissus, tels que les muscles par exemple.

Les réseaux de neurones artificiels cherchent donc à imiter le fonctionnement de ces neurones biologiques.

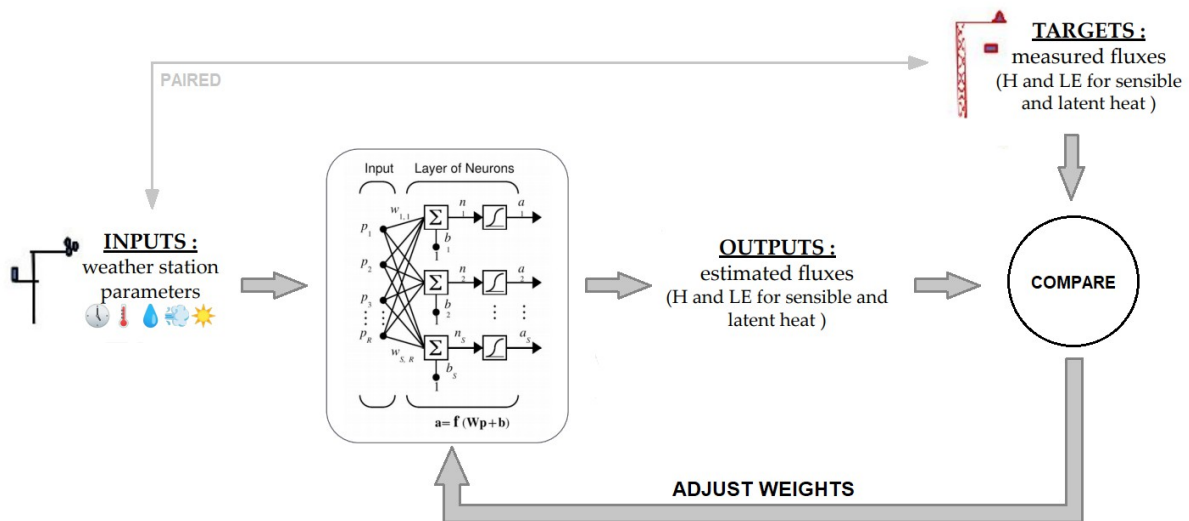


Fig. 2 : Fonctionnement général d'un réseau de neurones artificiels. Ce schéma est adapté à notre situation et reprend les principales étapes d'apprentissage des stations météorologiques classiques.

La **Fig. 2** donne une bonne idée de l'apprentissage par réseaux de neurones que nous allons appliquer ici. Les entrées de notre réseau (inputs, notées p) sont couplées aux targets qui représentent les valeurs cibles que le réseau va chercher à atteindre pendant son entraînement. Notre jeu de données en entrée correspond à nos données météorologiques classiques (température de l'air, humidité, temps,...) et les targets correspondent aux données des stations d'eddy-covariance qui leur sont couplées. Ces entrées et leurs targets associées représentent notre jeu d'apprentissage. En sortie du réseau, nous avons les données de sortie (outputs, notées a). Ces données de sortie sont comparées aux targets ce qui permet de déterminer s'il faut réajuster ou non notre réseau.

Pour l'apprentissage de notre réseau, nous utilisons la régularisation bayésienne (Bayesian Regularization Backpropagation), basée sur l'optimisation de Levenberg-Marquardt. Les réseaux de neurones utilisant ce type d'apprentissage comportent généralement trois couches : une couche d'entrée (input layer), une ou plusieurs couches cachées (hidden layers) et une couche de sortie (output layer). Chaque couche contient son propre nombre de neurones.

L'optimisation du réseau se fait via la modification des poids (W) et des biais (b) : c'est l'étape d'apprentissage. Les poids et les biais sont initialisés aléatoirement avant

l'apprentissage. Les biais sont des valeurs numériques ajoutées une fois que les poids sont appliquées aux données d'entrée. Les poids permettent de pondérer les variables d'entrée. Dans la première couche, les entrées pondérées et les biais sont sommés pour être ensuite multipliés par une fonction d'activation (ou fonction de transfert, notée **f**). Dans notre cas, nous utilisons une fonction tan-sigmoid comprise entre [-1 ; 1]. Cette fonction renvoie à la couche suivante une information, **a**, définie comme étant :

$$\mathbf{a} = \mathbf{f}(\mathbf{Wp}+\mathbf{b})$$

Notre jeu d'entrée peut être partagé en trois jeux de données différents. La répartition de nos données dans ces trois jeux se fait avant la phase d'apprentissage à l'aide de pourcentages. En voici les détails :

- le jeu d'apprentissage : ces données sont utilisées pendant l'entraînement. Ce jeu sert à ajuster les poids et les biais afin d'obtenir un compromis satisfaisant entre la précision de l'approximation sur les ensembles que le réseau a déjà rencontrés et celle obtenue avec de nouvelles données.
- le jeu de validation : ces données sont utilisées pendant l'entraînement. Ce jeu sert à vérifier, sans ajuster les poids et les biais, que le réseau apprend correctement.
- le jeu d'évaluation : ces données servent finalement à tester le réseau une fois que ce dernier a été entraîné.

Dans notre cas, la validation est intégrée à la fonction bayésienne utilisée. Il n'est donc pas nécessaire d'attribuer un pourcentage de données au jeu de validation et les données peuvent être réparties en deux jeux seulement : le jeu d'apprentissage et le jeu d'évaluation.

Afin de travailler avec ces réseaux de neurones, nous utilisons la toolbox Neural Network de Matlab dont les fonctionnalités sont décrites dans le manuel Neural Network Toolbox For Use with MATLAB® de Howard Demuth et Mark Beale [5].

2.2. Données disponibles

2.2.1. La campagne BLLAST : jeu de données sur 3 semaines

La campagne BLLAST (Boundary Layer Late Afternoon and Sunset Turbulence) [6], mise en place par le laboratoire d'aérodynamique LAERO de Toulouse sur le plateau de Lannemezan, a permis d'observer les quatre premiers kilomètres de l'atmosphère pour mieux comprendre et mieux représenter dans les modèles la période transitoire. Pour cela, plusieurs sites d'observation ont été mis en œuvre sur différents types de végétations pour une durée de 3 semaines. Le dispositif instrumental et les grandes lignes de la campagne de mesures sont détaillés dans Lothon et al., 2014 [7].

Dans le cadre de notre étude, nous disposons des données météorologiques et des flux mesurés par eddy-covariance sur ces douze journées de beau temps au-dessus de deux types de couverts végétaux différents : la lande et le maïs. Il est intéressant d'étudier ces deux végétations car durant la période d'étude, elles ne vont pas avoir le même comportement : le maïs, au contraire de la lande, va se développer et grandir au cours de la campagne (il est passé de 0.4m à 1.5m en 3 semaines), pouvant donc influencer les flux en surface.

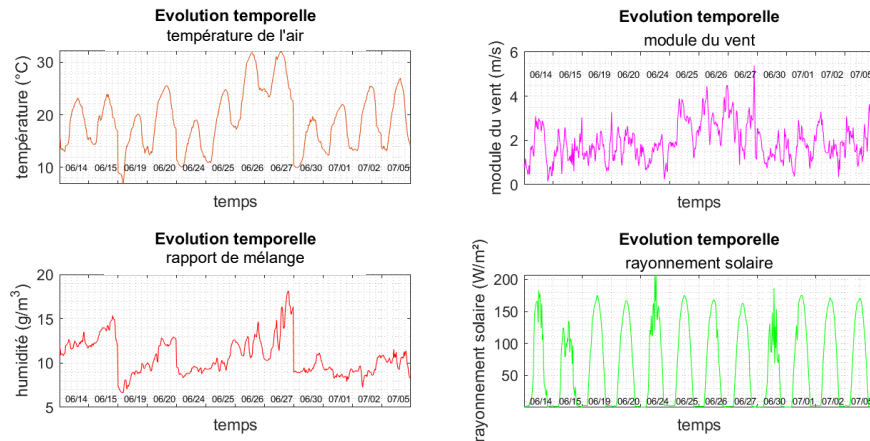


Fig. 3: Conditions météorologiques sur les douze journées de beau temps considérées dans la première partie du stage.

Sur ces douze journées, prises entre le mois de juin et de juillet 2011, nous avons deux journées particulièrement chaudes, apportant un peu de variété à notre jeu de données : le 26 et le 27 juin. Les températures dépassent les 30°C au maximum de la journée pour ces deux jours et pour le 27 juillet, le rapport de mélange dépasse les 15g de vapeur d'eau/kg d'air sec.

Il faut aussi noter deux informations supplémentaires concernant ces douze journées. Premièrement, nous n'avons que très peu de vent. Deuxième, entre les douze journées utilisées pour l'entraînement de notre réseau de neurones, il a plu ce qui implique un sol très humide et des conditions d'humidité du sol très homogènes sur le jeu de données étudié. Cela peut influencer nos résultats et il faudra donc en tenir compte dans la discussion de nos expérimentations.

2.2.2. Petit mât de 2m instrumenté : jeu de données sur une année

Afin de compléter les résultats obtenus avec les données de la campagne BLLAST, nous utilisons un an de données issues du petit mât instrumenté de 2m disponible sur le site du Centre de Recherches Atmosphériques de Lannemezan. Des flux sont mesurés sur le site d'août 2012 à juillet 2013 dans une prairie. La résolution temporelle de ces flux mesurés par eddy-covariance est de 30 minutes. Les données météorologiques (température de l'air, rapport de mélange, module de la vitesse du vent et rayonnement net) sont donc moyennées pour être ramenées au même pas de temps, soit 48 données couplées par 24h.

Sur ce nouveau jeu de données, plus long que le précédent, il a fallu filtrer les données, notamment les données turbulentes. Nous avons mis de côté certaines journées pour lesquelles des flux décalés dans le temps étaient observés, probablement dus à des coupures d'électricité. Il y a de ce fait des trous dans notre jeu de données qu'il faudra garder en tête lors de notre étude. Les données pourraient être filtrées et corrigées de façon plus rigoureuse (il reste notamment des pics de flux inexplicables pour certains mois) mais il aurait fallu plus de temps pour réaliser de telles corrections.

2.3. Définition des différents tests

Afin de juger de la performance de cette technique sur les différents jeux de données et d'améliorer notre réseau de neurones, nous avons réalisé différents tests. Pour chaque test, une fiche type a été réalisée afin de regrouper les différents résultats obtenus et les figures correspondantes (**Annexe A**). Les résultats présentés plus loin sont des synthèses de ces fiches types.

Tous les tests présentés ci-après seront comparés au test de référence nommé tref (**Tableau 1**) défini en utilisant les résultats de Kelley et al. [3].

2.3.1. Jeu d'apprentissage : tests tsl (**Tableau 1**)

Le premier jeu de tests (**Annexe B**) réalisé concerne la sensibilité de la méthode au jeu d'apprentissage. Lors de ce test, nous allons notamment regarder l'importance du choix des données utilisées pour l'entraînement de notre réseau de neurones. L'objectif principal est de déterminer l'impact de la longueur du jeu d'apprentissage et de la sélection des données qui composent le jeu d'apprentissage (**Fig. 4**).

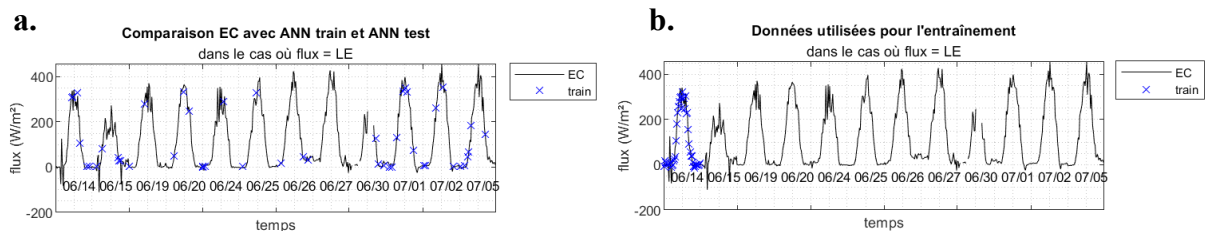


Fig. 4: Flux de chaleur latente de référence mesuré par les stations EC au cours des 12 jours. Les figures (a) et (b) présentent deux exemples de sélection des données, indiquées par des croix bleues, qui composent le jeu d'apprentissage : (a) 48 données réparties aléatoirement sur les douze jours, (b) 48 points concentrés sur une journée.

2.3.2. Topographie du réseau de neurones : tests ntt (**Tableau 1**)

L'objectif de ce jeu de tests (**Annexe B**) est de déterminer la topographie de notre réseau qui nous permet d'obtenir les meilleurs flux estimés. Pour cela, nous pouvons faire varier le nombre de couches cachées ainsi que le nombre de neurones à l'intérieur de ces couches.

2.3.3. Variables en entrée de notre réseau : test ivt (**Tableau 1**)

Nous cherchons avec ce jeu de tests (**Annexe B**) à déterminer l'influence des variables en entrée de notre réseau sur les flux estimés. Pour cela, nous allons ajouter et supprimer des variables de façon à définir quelles sont celles qui semblent avoir le plus d'importance dans l'entraînement de notre réseau.

	Variables		Topographie		Jeu d'apprentissage	
	Entrée	Jours	Nb couches	Nb neurones	% train	% test
tref : test de référence	T, t, r _v , w _s , SW _↓	12 journées de beau temps	1	10	40	60
tsl : training set length	-	Détermination de l'influence des jours choisis sur les flux ANN estimés	-	-	Impact de la longueur du jeu d'apprentissage et de la sélection des données qui composent ce jeu	
ntt : network topography test	-	-	Recherche de la topographie du réseau offrant les meilleurs résultats tout en restant le plus simple possible		-	-
ivt : input variables test	Influence des variables d'entrée (ajout et suppression de variables météo)	-	-	-	-	-

Tableau 1: Tableau récapitulatif des différents tests effectués et des paramètres modifiés. Les variables d'entrée présentées dans ce tableau sont la température (T), le temps (t), le rapport de mélange (r_v), le module de la vitesse du vent (w_s) et le rayonnement solaire incident (SW_↓).

2.4. Critères de comparaison des différents tests

Afin de comparer les résultats des différents tests listés ci-dessus, nous avons défini des critères de comparaison. Nous en avons choisi deux : le coefficient de corrélation et l'écart quadratique moyen entre les flux ANN estimés par le réseau de neurones et les flux EC mesurés par eddy-covariance.

Le coefficient de corrélation, noté R, nous permet comme son nom l'indique de juger de la corrélation entre les flux ANN et EC.

L'écart quadratique moyen (noté RMSE pour Root-Mean-Square Error) nous permet de connaître, en W/m², la différence moyenne entre les flux mesurés et les flux estimés. L'erreur RMSE vient donc compléter le coefficient de corrélation puisque nous pouvons tout à fait avoir un R important mais obtenir une grande erreur RMSE (**Fig. 5**).

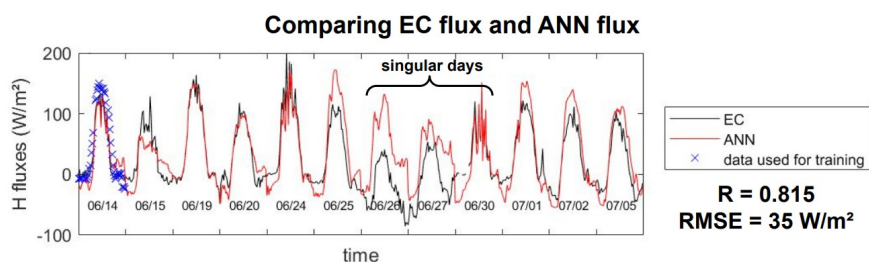


Fig. 5: Évolution temporelle des flux de chaleur sensible EC mesurée et ANN estimés avec un jeu d'apprentissage composé de données mesurées sur la première journée (croix bleues le 06/14).

Sur la **Fig. 5** nous observons l'évolution temporelle des flux de chaleur sensible ANN (en rouge) et EC (en noir). Les croix bleues représentent les données utilisées pour l'apprentissage du réseau de neurones. Nous remarquons donc que dans ce cas-ci, le réseau a été entraîné sur la première journée (06/14) puis appliqué aux onze journées restantes. A première vue, les deux flux ne semblent pas bien corrélés. C'est d'ailleurs ce que vient confirmer le RMSE, puisqu'il est égal à 35 W/m² pour des flux en journée aux alentours de

100 W/m². Cependant, si nous regardons seulement le coefficient de corrélation, nous voyons qu'il est autour de 0.82 ce qui pourrait, dans d'autres circonstances, sembler être une corrélation suffisante. Cette corrélation importante est due au fait que le réseau permet de correctement modéliser le cycle diurne. D'où l'importance dans notre cas d'étudier à la fois R et RMSE.

2.5. Statistiques

2.5.1. Répétitions indépendantes des tests

Les flux estimés dépendent fortement des données choisies pour l'apprentissage. Si on choisit des données à entraîner de façon aléatoire, d'un test à un autre, les résultats peuvent varier significativement.

Afin d'obtenir des résultats représentatifs, nous avons donc cherché combien de répétitions d'un même test nous devions faire afin d'éviter les cas particuliers.

Nous obtenons dans l'ensemble des coefficients de corrélation compris entre 0.96 et 0.98 (**Fig. 6**). Nous constatons que la distribution devient régulière à partir de 100 répétitions. Nous choisissons donc de répéter 100 fois chaque test afin de prendre les valeurs médianes de R et RMSE pour caractériser ce test et le comparer aux autres.

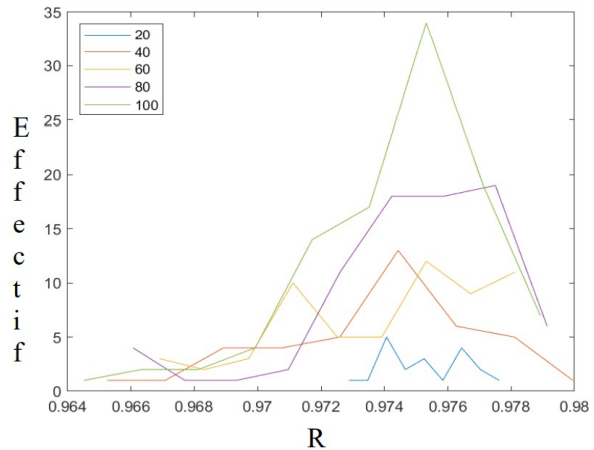


Fig. 6: Distribution du coefficient de corrélation en fonction du nombre de répétitions d'un même test pour le test de référence.

2.5.2. Significativité du coefficient R

Comme nous l'avons vu précédemment, nous utilisons le coefficient de corrélation R afin de juger de la corrélation entre nos flux estimés et nos flux mesurés.

Un test existe pour vérifier la significativité de nos coefficients R. Pour cela, on prend comme hypothèse H_0 qu'il n'existe pas de corrélation entre les flux estimés et les flux mesurés. Si nous rejetons cette hypothèse, alors l'hypothèse H_1 (H_1 : il existe une corrélation entre nos flux) est validée.

Pour que H_0 soit rejetée, on doit avoir :

$$\frac{|r_{XY}|}{\sqrt{1-r_{XY}^2}} \sqrt{n-2} > s(\alpha/2)$$

avec r_{XY} : la corrélation entre les échantillons X et Y (pour nous, les flux estimés et mesurés)

n : le nombre d'observations de X et Y (dans notre cas, $n = 576$)

$s(\alpha/2)$: la valeur critique, à lire dans la table de Student (pour nous, α représente le seuil d'erreur que nous choisissons à 5%, soit $\alpha = 5\%$)

Au seuil de 5%, $s(\alpha/2) = 1.965$ dans notre cas. Nous devons donc trouver une valeur supérieure à cette valeur critique pour conclure qu'il existe bien une corrélation significative entre les deux échantillons. Dans notre cas, nous choisissons plutôt de déterminer la valeur du

coefficient R à partir de laquelle nous obtenons une corrélation significative en résolvant l'inéquation :

$$\frac{|r_{XY}|}{\sqrt{1-r_{XY}^2}} \sqrt{576-2} > 1.965$$

Nous trouvons alors que notre coefficient est significatif à condition qu'il soit supérieur à $R = 0.082$. Une valeur du coefficient de corrélation supérieure à 0.082 suffit donc dans notre cas pour que la corrélation estimée soit significative à 5%.

3. RÉSULTATS

3.1. MOOR/CORN : tests pour la lande et le maïs sur une courte durée

3.1.1. Tests tsl pour la lande : sensibilité au jeu d'apprentissage

Le premier jeu de tests que nous avons réalisé est celui concernant le jeu d'apprentissage sur la lande. Comme évoqué précédemment, nous allons déterminer les différences entre un réseau entraîné avec des données choisies de façon aléatoire et un réseau entraîné avec des données imposées sur une ou plusieurs journées (**Fig. 4**). Les différents tests réalisés pour la lande sont synthétisés sur la **Fig. 7**.

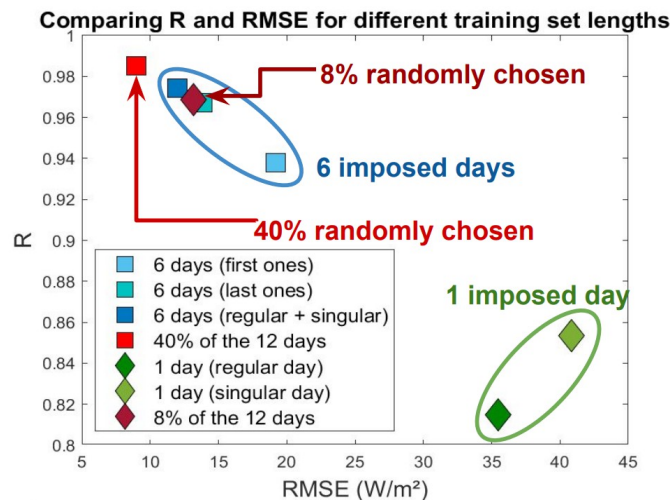


Fig. 7: Synthèse des différents tests réalisés sur le jeu d'apprentissage pour la chaleur sensible. Cette figure représente le coefficient de corrélation R en fonction de l'erreur quadratique $RMSE$. Les losanges verts représentent les résultats obtenus en imposant une journée. Les carrés bleus représentent les résultats obtenus en imposant six journées. Les marqueurs rouges représentent, pour le losange, 8% (équivalent à une journée) des données choisies aléatoirement et le carré rouge, 40% (équivalent à 6 journées).

Analysons cette figure en commençant par comparer ce que l'on obtient en imposant une journée comme entraînement et ce que l'on obtient en laissant le réseau choisir aléatoirement 8% de toutes les données disponibles sur les douze jours. Sur la **Fig. 7**, deux points correspondent à un entraînement sur une journée complète. Nous nous concentrons sur le point à $R = 0.82$ et $RMSE = 35 \text{ W/m}^2$ qui correspond au test étudié dans la partie « Critères de comparaison des différents tests ». Nous avons alors noté qu'un tel $RMSE$ était élevé pour des flux d'environ 100 W/m^2 et qu'un coefficient de corrélation de 0.82 n'était pas représentatif d'une bonne corrélation entre les flux ANN et EC.

Nous comparons ce point avec les résultats obtenus lorsque nous choisissons d'entraîner notre réseau avec 8% des données choisies aléatoirement parmi nos douze jours (le même nombre de points que pour un jour seul fixé). Ce point-ci est représenté sur la **Fig. 7** par un losange rouge, et lui sont attribués un $R = 0.97$ et une erreur $RMSE = 13 \text{ W/m}^2$. C'est une nette amélioration par rapport aux résultats obtenus en choisissant d'entraîner notre réseau sur une seule journée.

Ces résultats sont confirmés par les flux représentés sur la **Fig. 8**.

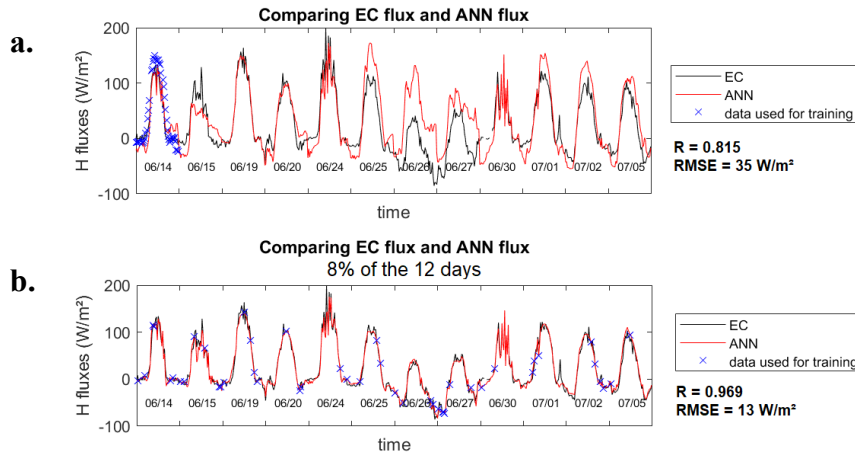


Fig. 8 : Représentation des flux de chaleur sensible mesurés (EC) et estimés (ANN) pour deux jeux d'apprentissage différents : **(a)** les points sont pris sur une journée complète que nous choisissons, **(b)** les points qui servent à l'entraînement sont aléatoirement sélectionnés dans notre jeu de données (8% de nos 12 journées).

Une première conclusion est donc que nous obtenons de meilleurs résultats en présentant en entrée un pourcentage de données choisi aléatoirement. Cette même conclusion peut être faite en comparant les résultats obtenus en entraînant le réseau avec 40% des données choisies aléatoirement et en l'entraînant avec 6 journées imposées (ce qui reviendrait à l'entraîner avec le même nombre de points, répartis différemment). Mais nous remarquons aussi sur la **Fig. 7** qu'en entraînant notre réseau avec six jours complets, soit la moitié de notre jeu d'apprentissage, nous obtenons des résultats aussi bons (voire moins bons) qu'en l'entraînant avec seulement 8% de données choisies aléatoirement. Comme attendu, le plus important semble être d'échantillonner le plus grand nombre de conditions différentes possibles afin de lui permettre de rencontrer un maximum de conditions différentes.

3.1.2. Tests tsl pour le maïs : sensibilité au jeu d'apprentissage

Nous avons réalisé les mêmes tests faits pour la lande maïs pour une végétation qui va croître durant la période étudiée : le maïs. Comme pour la lande, on note que le choix aléatoire du jeu d'apprentissage pour le maïs permet un meilleur entraînement du réseau de neurones et donne donc des meilleures corrélations et une plus faible RMSE (**Fig. 9**). Cependant, le maïs semble plus sensible au choix du jeu d'apprentissage. En effet, les corrélations obtenues lorsqu'une journée d'apprentissage est imposée sont bien meilleures pour la lande (0.82) que pour le maïs (0.35). Cela est probablement dû au fait que pendant la période étudiée, le maïs va grandir au contraire de la lande.

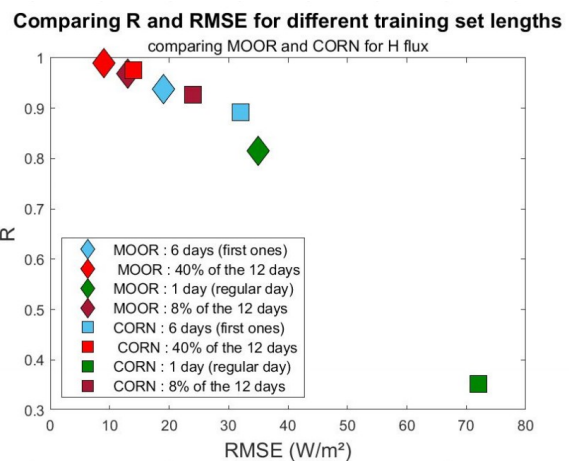


Fig. 9 : Comparaison des résultats obtenus pour le test 1 pour les deux végétations étudiées.

3.1.3. Tests ntt pour la lande : influence de la topographie du réseau de neurones

Le second jeu de tests concerne la topographie du réseau. Nous avons testé différentes structures afin d'obtenir un réseau optimisé pour la lande.

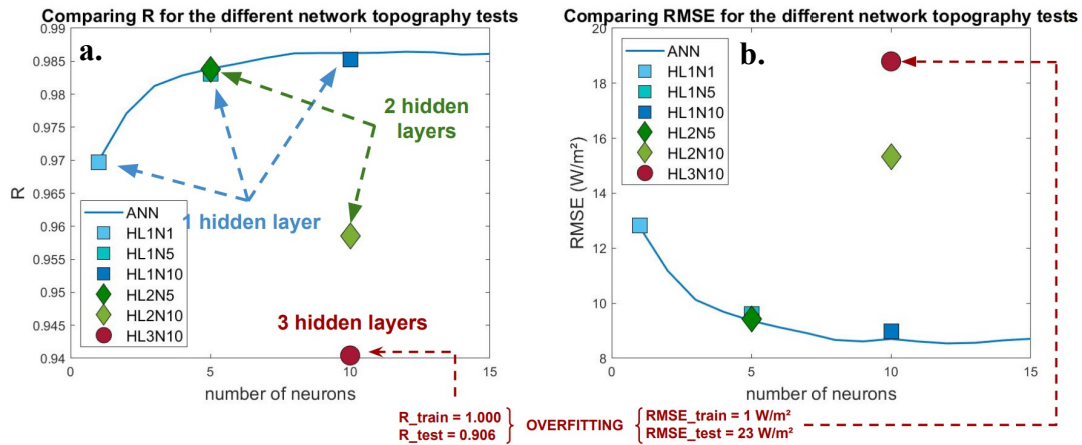


Fig. 10: Synthèse des résultats obtenus pour les tests de topographie (lande). **(a)** représente l'évolution du coefficient de corrélation R en fonction du nombre de neurones et **(b)** représente l'évolution de l'erreur RMSE en fonction du nombre de neurones pour la chaleur sensible.

Les différents résultats que nous avons obtenus pour les tests sur la topographie sont résumés sur la **Fig. 10**. Nous avons commencé par ajouter 1 (test HL2N10) puis 2 (test HL3N10) couches de 10 neurones par rapport au test de référence qui n'avait qu'une seule couche de 10 neurones (test HL1N10) et nous remarquons que plus il y a de couches, plus le coefficient de corrélation diminue (**Fig. 10a**) et plus l'erreur RMSE augmente (**Fig. 10b**). Nous en sommes donc rapidement venus à la conclusion qu'une seule couche de neurones était suffisante dans notre étude.

De plus, en augmentant le nombre de couches, nous remarquons que la qualité de notre coefficient de corrélation et de notre erreur RMSE pendant l'entraînement augmentent de façon importante mais une fois que nous appliquons notre réseau entraîné sur le reste de notre jeu de données, nous obtenons des résultats bien inférieurs. C'est par exemple le cas pour trois couches, comptant chacune dix neurones. Dans ce cas là, pendant l'entraînement, nous avons un $R = 1$ et une erreur $RMSE = 1 \text{ W/m}^2$ (**Fig. 10**). Cependant, lorsque l'on teste notre réseau sur des données qu'il n'a jamais rencontrées, on obtient un $R = 0.91$ et une erreur $RMSE = 19 \text{ W/m}^2$ ce qui est nettement moins satisfaisant. Nous avons ici observé un cas de sur-apprentissage. Rajouter des couches de neurones et créer un réseau plus complexe n'améliore donc pas systématiquement la performance du réseau.

Une fois la conclusion faite qu'une seule couche de neurones était suffisante, nous avons regardé quel était le nombre de neurones sur cette unique couche nous permettant d'obtenir un réseau optimal. Nous avons donc tracé l'évolution de R et de l'erreur RMSE en fonction du nombre de neurones sur cette couche (**Fig. 10**, ligne bleue). Nous obtenons rapidement des résultats convenables mais au delà de 8 neurones, notre R et notre RMSE semblent se stabiliser.

Pour la lande, il semblerait donc que le réseau optimal soit composé d'une seule couche cachée composée de huit neurones.

En répétant les mêmes tests sur le maïs, nous obtenons les mêmes conclusions (**Annexe C**).

3.1.4. Tests ivt pour la lande : sensibilité aux variables en entrée de notre réseau

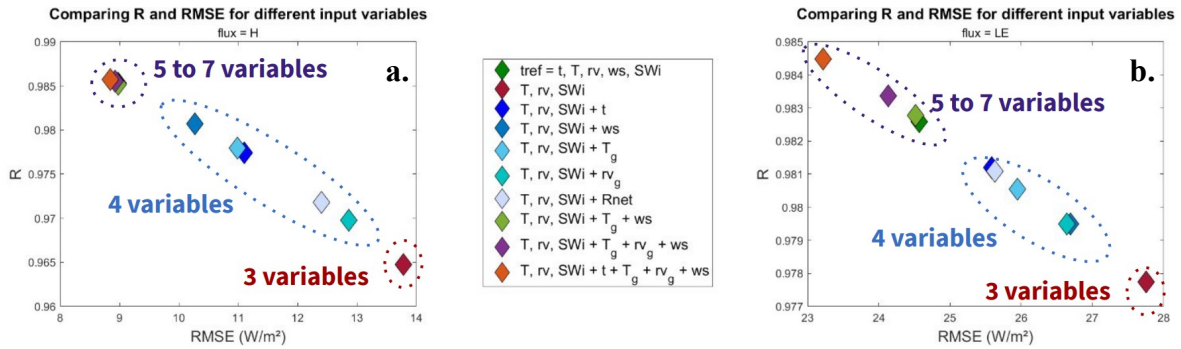


Fig. 11. Synthèse des résultats obtenus en entraînant notre réseau avec différentes variables en entrée. Évolution simultanée du coefficient R et de l'erreur RMSE en fonction des variables choisies en entrée (a) pour le flux de chaleur sensible (b) pour le flux de chaleur latente. Les variables d'entrée présentées dans la légende sont la température (T), le temps (t), le rapport de mélange (r_v), le rayonnement net (Rnet), le module de la vitesse du vent (w_s), le rayonnement solaire incident (Sw_i), la température du sol (T_g) et l'humidité du sol (H_g).

Enfin, nous avons réalisé un test concernant les variables mises en entrée de notre réseau. Ce test est réalisé dans un premier temps pour la lande et les résultats obtenus sont synthétisés dans la **Fig. 11**. Attention cependant aux échelles des ordonnées lors de l'analyse de ces résultats : nous comparons des valeurs qui dans l'ensemble sont relativement bonnes.

Nous avons commencé par entraîner notre réseau avec les trois variables physiques suivantes : la température, l'humidité et le rayonnement solaire incident. Les résultats obtenus sont moins bons que ceux observés avec le test de référence qui lui, en plus de ces variables-ci, comptait aussi le temps et le module de la vitesse du vent. Par la suite, nous avons ajouté à ces trois variables physiques un quatrième paramètre parmi ceux à notre disposition qui nous semblaient les plus pertinents. Nous remarquons sur la **Fig. 11a** que pour le flux de chaleur sensible, nous obtenons les meilleurs résultats à 4 variables en ajoutant le module de la vitesse du vent. Ce résultat nous semble tout à fait raisonnable puisque les transferts entre le sol et l'atmosphère sont facilités par la turbulence dynamique et donc dépendants du vent. Sur la **Fig. 11b**, soit pour la chaleur latente, les meilleurs résultats à quatre variables sont obtenus en utilisant le temps. Ce résultat, bien que déjà mis en évidence par des études antérieures, ne nous semble pas facilement explicable d'un point de vue de la physique de l'atmosphère. Concernant les paramètres du sol (température et contenu en eau), les flux étant proportionnels à un gradient vertical de température (flux de chaleur sensible) et d'humidité (flux de chaleur latente), nous pourrions nous attendre à une plus nette amélioration de R et RMSE lorsque ceux-ci sont ajoutés aux trois paramètres d'entrée initiaux. L'amélioration est notable lors de l'ajout de la température du sol mais négligeable pour l'humidité du sol. Ceci est sans doute dû à un déphasage des fluctuations d'humidité du sol par rapport aux flux turbulents alors que la température du sol varie de façon quasi-instantanée avec les variations atmosphériques.

De façon plus générale, que ce soit pour la chaleur sensible ou pour la chaleur latente, nous remarquons que nous obtenons de très bons résultats en utilisant entre cinq et sept variables.

Par la suite, nous avons réalisé un test supplémentaire mêlant topographie et variables d'entrée. Pour cela, nous observons l'évolution du nombre de neurones en utilisant une seule couche cachée pour différentes variables en entrée de notre réseau (**Fig. 12**). Nous testons cela pour trois (température, humidité et rayonnement solaire), cinq (température, humidité, rayonnement solaire, temps, module de la vitesse du vent) et sept variables (température,

humidité, rayonnement solaire, temps, module de la vitesse du vent, température du sol et humidité du sol).

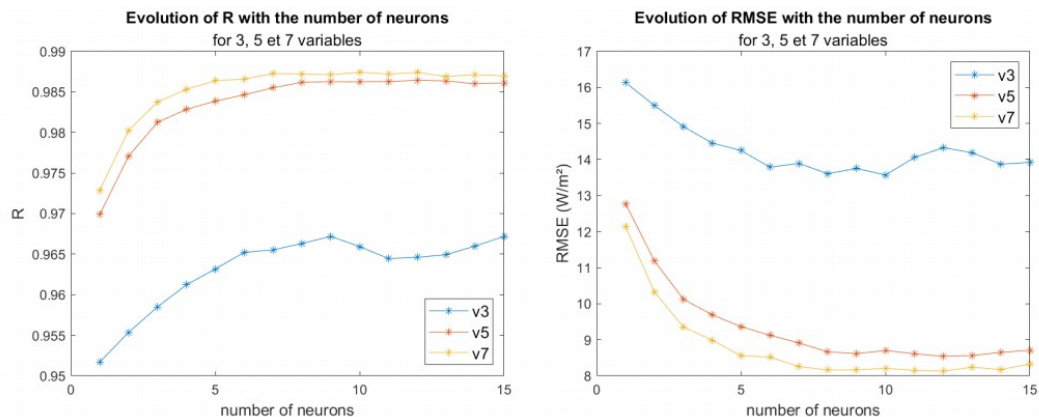


Fig. 12: Évolution du coefficient de corrélation (à gauche) et de l'erreur RMSE (à droite) en fonction du nombre de neurones pour différentes variables en entrée. Ces tests sont réalisés pour le flux de chaleur sensible.

Nous remarquons deux choses. Premièrement, nous confirmons le fait que d'utiliser trois variables fonctionne moins bien que d'en utiliser cinq ou sept. Nous notons aussi, au passage, que la différence entre cinq et sept variables n'est encore une fois pas flagrante. Deuxièmement, nous remarquons que dans les trois cas, nous obtenons une structure de notre réseau qui semble optimale pour huit neurones, quelque soit le nombre de variables en entrée.

3.1.5. Tests ivt pour le maïs : sensibilité aux variables en entrée de notre réseau

Nous répétons les mêmes tests sur les variables d'entrée mais pour le second type de végétation étudié : le maïs. Les résultats que nous avons obtenu pour ce couvert sont synthétisés dans la **Fig. 13**. (attention, notamment pour le flux de chaleur latente, nous comparons des résultats qui dans l'ensemble sont bons, voir les échelles des ordonnées).

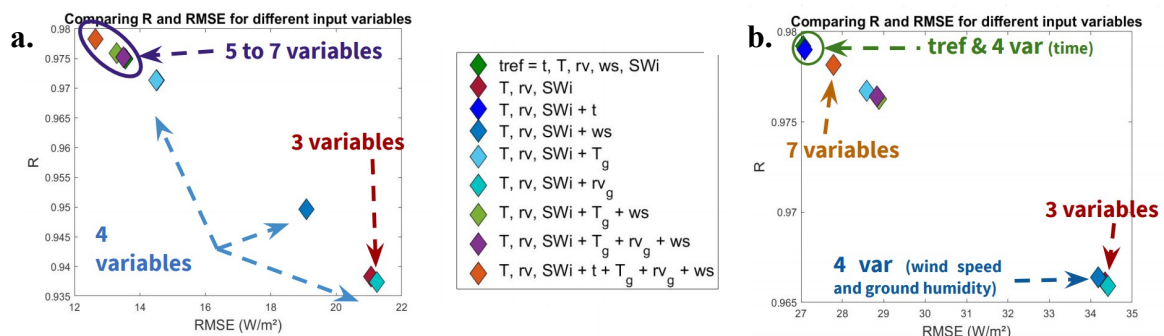


Fig. 13: Synthèse des résultats obtenus en entraînant notre réseau avec différentes variables en entrée (pour le maïs). Est représentée ici l'évolution simultanée du coefficient R et de l'erreur RMSE en fonction des variables choisies en entrée (a) pour le flux de chaleur sensible (b) pour le flux de chaleur latente.

Nous remarquons que pour le flux de chaleur sensible (**Fig. 13a**) nous obtenons des résultats globalement semblables à ceux que l'on obtenait pour la lande. Nous pouvons faire les mêmes conclusions, à savoir que l'humidité du sol n'améliore pas nos résultats dans le cas où nous présentons quatre variables et qu'il n'y a que très peu de différences entre les résultats obtenus avec un nombre de variables compris entre cinq et sept.

Par contre, pour le flux de chaleur latente (**Fig. 13b**) nous obtenons des résultats quelque peu différents. En effet, nous trouvons que l'humidité du sol n'est pas une variable qui

permet une amélioration de nos résultats comme précédemment, mais en plus ici nous remarquons que le module de la vitesse du vent ne permet pas non plus d'obtenir de meilleurs résultats. La variable prédominante dans ce cas-ci semble être le temps. En effet, les trois meilleurs tests observés sont ceux qui comptent le temps parmi leurs variables. Il semblerait même que de n'utiliser que la température, l'humidité, le rayonnement solaire incident et le temps (quatre variables) nous permette d'obtenir de meilleurs résultats qu'en utilisant les sept variables (sachant que dans ces sept variables apparaissent le module de la vitesse du vent et l'humidité du sol).

3.2. Étude d'un jeu de données sur une année

3.2.1. Test tsl pour un an de données : sensibilité au jeu d'apprentissage

Afin d'aller plus loin dans cette étude, nous allons tester la sensibilité du réseau au jeu d'apprentissage pour un an de données. Cela permet de nous rapprocher de l'objectif à long terme qui serait de déployer une telle méthode sur le site de Lannemezan afin d'estimer des flux de façon continue. Puisque ce nouveau jeu contient plus de données que le précédent afin de simuler le déploiement de cette méthode, les tests réalisés vont être légèrement différents mais nous gardons le même objectif : déterminer l'impact de la longueur du jeu d'apprentissage et de la sélection des données qui composent ce dernier.

Nous travaillons dans des conditions similaires aux tests tsl précédents, à savoir : un réseau composé d'une couche cachée et de dix neurones, ainsi que le temps, la température de l'air, le rapport de mélange, le module de la vitesse du vent et le rayonnement net en paramètres d'entrée.

Cinq tests vont être réalisés. Nous reprenons le test de référence précédemment effectué (à savoir une sélection aléatoire pour l'apprentissage de 40% des données disponibles) et nous introduisons quatre nouveaux tests pour ce jeu, présentés ci-dessous :

- entraîner le réseau avec les cinq premiers jours de chaque mois, pendant une année, sans prendre en compte la qualité de ces cinq premiers jours. Il serait possible de mettre en place un tel apprentissage sur le plateau de Lannemezan. Par exemple, nous pouvons envisager d'entraîner le réseau de neurones à partir d'une station d'eddy-covariance pendant 5 jours sur une prairie avant de déplacer la station sur un couvert végétal différent pendant 5 jours,... etc pendant un mois. Puis, recommencer cette rotation au début de chaque mois.
- entraîner le réseau avec cinq journées par mois sélectionnées parmi les journées pour lesquelles les flux et les données météorologiques semblent corrects. Cette méthode est difficilement applicable, puisqu'elle nécessiterait de connaître à l'avance les conditions météorologiques et turbulentes. Mais il est intéressant de comparer ce test au précédent utilisant la même méthode d'apprentissage mais sans tenir compte des conditions rencontrées.
- entraîner le réseau avec 16% du jeu données choisis aléatoirement. Cela revient à utiliser le même nombre de points pour l'entraînement qu'en utilisant 5 jours tous les mois. De ce fait, nous pourrions comparer les résultats obtenus en imposant 5 jours tous les mois avec ceux obtenus en laissant le réseau choisir aléatoirement l'équivalent de cinq jours/mois sur toute l'année.

- entraîner le réseau avec 3 semaines par saison (3 semaines en septembre, 3 semaines en décembre, 3 semaines en mars et 3 semaines en juin). En terme de logistique, cette méthode serait la plus facile à mettre en place car elle nécessite moins de déplacements de couvert végétal en couvert végétal de la station de mesures de flux.

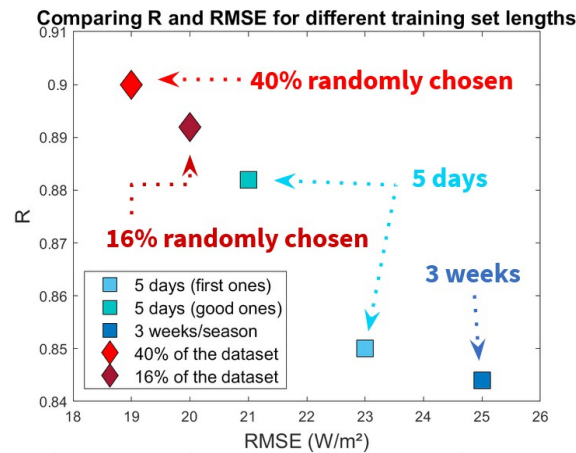


Fig. 14: Synthèse des différents tests réalisés sur le jeu d'apprentissage d'une année pour la chaleur sensible. Cette figure représente le coefficient de corrélation R en fonction de l'erreur quadratique RMSE. Les carrés bleus représentent les résultats obtenus en imposant certaines journées ou certaines semaines. Les marqueurs rouges représentent les résultats obtenus en laissant le réseau choisir aléatoirement un certain pourcentage de points parmi toutes les données disponibles.

Les résultats obtenus pour ces différents tests (pour la chaleur sensible) sont synthétisés dans la **Fig. 14**. De façon générale, nous remarquons que comparés à l'étude précédente sur douze jours, les coefficient R et les erreurs RMSE sont moins bons. Cela est compréhensible, vu que nous étudions ici un panel de données plus large et plus important. Comme lors de notre étude sur le précédent jeu de données, nous remarquons que les meilleurs résultats sont obtenus pour un réseau de neurones qui apprend avec 40% de données aléatoirement sélectionnées. Nous remarquons aussi qu'en choisissant 5 journées par mois, nous obtenons de moins bons résultats qu'en sélectionnant aléatoirement 16% des données. En quantité de points utilisés pour l'entraînement, utiliser 5 jours ou 16% des données revient au même, ce qui vient confirmer les observations faites lors de l'étude du jeu précédent à savoir : choisir aléatoirement un certain pourcentage de données pour apprendre au réseau permet d'obtenir de meilleurs résultats qu'en sélectionnant et imposant certaines journées complètes au réseau. De même, de meilleurs résultats sont obtenus lorsque nous choisissons 5 bonnes journées au lieu de prendre à chaque fois les 5 premières journées de chaque mois (sans tenir compte des conditions de ces premières journées). Les résultats les moins bons sont obtenus pour un entraînement de 3 semaines/saison. Cette méthode, bien que plus facile à mettre en place, ne semble pas être la plus efficace.

Pour la chaleur latente, les mêmes conclusions peuvent être faites.

3.2.2. Test ntt pour un an de données : influence de la topographie du réseau de neurones

Lors des précédents tests sur un jeu de douze journées de beau temps, nous avons remarqué que la topographie du réseau la plus optimale était atteinte pour une couche cachée et huit neurones. Nous avons aussi vu que rajouter des couches cachées n'améliorait pas les résultats obtenus, au contraire. Nous allons ici refaire les mêmes tests mais pour une année de données. Nous travaillons avec un jeu d'apprentissage de 40% de données aléatoirement sélectionnées parmi toutes les données disponibles et avec les variables d'entrée suivantes : le temps, la température de l'air, le rapport de mélange, le module de la vitesse du vent et le

rayonnement net.

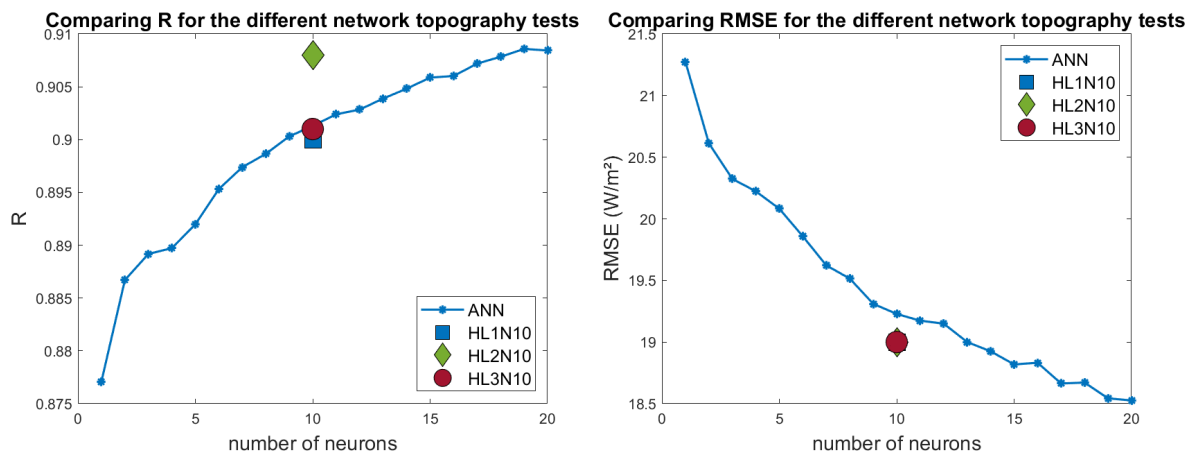


Fig. 15: Synthèse des résultats obtenus pour les tests de topographie (jeu de données sur une année, chaleur sensible). **(a)** représente l'évolution du coefficient de corrélation R en fonction du nombre de neurones et **(b)** représente l'évolution de l'erreur RMSE en fonction du nombre de neurones pour la chaleur sensible.

Contrairement au jeu précédent, nous remarquons cette fois-ci qu'augmenter le nombre de couches n'altère pas significativement les critères de comparaison du réseau. Une légère augmentation du coefficient de corrélation peut même être observée en comparant les résultats obtenus pour une couche à 10 neurones ($R \sim 0.90$) et deux couches à 10 neurones ($R \sim 0.91$) (**Fig. 15a**). De même, nous remarquons que l'erreur RMSE (**Fig. 15b**) est la même que le réseau soit composé d'une, de deux ou de trois couches ($RMSE = 19W/m^2$).

Les courbes bleues (**Fig. 15**) représentent l'évolution du coefficient de corrélation et de l'erreur RMSE en fonction du nombre de neurones sur une seule couche cachée. En utilisant une structure à une seule couche et en augmentant progressivement le nombre de neurones, nous améliorons légèrement l'erreur commise entre les flux estimés ANN et les flux mesurés EC. L'erreur passe de $19W/m^2$ pour 10 neurones sur une couche cachée à $18.5W/m^2$ pour 20 neurones sur une couche cachée. Cette différence n'est pas significative pour notre étude mais elle permet de mettre en avant le fait que, contrairement aux résultats obtenus pour le jeu de données BLLAST, ajouter des neurones semble ici améliorer l'erreur. Pour le coefficient de corrélation, le réseau parvient, avec une couche et 19 neurones à obtenir une corrélation aussi bonne qu'en utilisant deux couches et 10 neurones.

4. DISCUSSION

4.1. CORN/MOOR : discussion des résultats

Les tests concernant le jeu d'apprentissage ont permis de mettre en avant l'importance d'échantillonner un maximum de conditions différentes dans ce jeu. Le choix aléatoire des points d'entraînement permet cet échantillonnage, et de ce fait une meilleure estimation des flux de chaleur sensible et de chaleur latente. Cette sensibilité au jeu d'apprentissage avait déjà été relevée dans les précédentes études et reste un point important à prendre en compte dans l'utilisation d'une telle méthode.

De façon générale, notre réseau de neurones semble moins sensible à la topographie du réseau qu'au jeu d'apprentissage. Cependant, quelques structures semblent plus optimales que d'autres. C'est notamment ce que nous avons pu remarquer en utilisant une couche cachée et huit neurones. De précédentes recherches [2] ont mis en avant le fait que le choix de neurones optimal été un nombre ne dépassant pas deux fois le nombre de variables, soit dix dans notre cas. Le nombre de huit neurones obéit donc bien à cette loi.

Lorsque nous réalisons les tests sur les variables d'entrée, nous remarquons que nous obtenons des résultats moins bons en entraînant notre réseau avec les trois variables physiques (température, humidité et rayonnement incident) et l'humidité du sol. Ceci pourrait s'expliquer par le fait que l'humidité du sol présente un décalage temporel avec les autres variables. Ce constat a déjà été fait par Jason Kelley et Eric Pardyjak lors de leurs études sur l'estimation de l'évapotranspiration par intelligence artificielle. Il est donc rassurant d'observer que nos constats rejoignent les leurs.

De même, nous remarquons pour les flux de chaleur latente que la variable temps semble avoir une forte influence sur la bonne estimation des flux. Cela semble notamment être dû au fait que les flux de chaleur latente évoluent peu pendant la période étudiée, et que le temps permet donc une bonne représentation du cycle diurne et une certaine périodicité.

Pour la chaleur sensible, la température du sol semble importante pour le maïs et le module de la vitesse du vent semble important pour la lande. Mais comme nos jeux ne sont composés que de quelques jours avec très peu (voire pas du tout) de vent, il est difficile de conclure sur l'importance de celui-ci. D'autant plus que pour le maïs, si l'on regarde la chaleur sensible, le module de la vitesse du vent semble diminuer les performances du réseau...

De façon plus générale, nous pouvons dire qu'ajouter des variables semble améliorer notre réseau mais qu'entre cinq variables bien choisies et sept variables, les différences ne sont pas suffisamment importantes pour justifier la nécessité d'utiliser sept variables. Mais notre jeu est limité car l'humidité du sol et les conditions de vents changent peu, d'où l'intérêt de l'utilisation d'un plus grand jeu de données.

4.2. Jeu de données sur une année : discussion des résultats

Comme lors des précédents tests réalisés sur douze jours, les tests concernant le jeu d'apprentissage ont permis de mettre en avant l'importance d'échantillonner un maximum de conditions du jeu de données. Pour cela, la méthode la plus efficace semble être de laisser le réseau choisir un certain pourcentage de données de façon aléatoire. Cependant, ce n'est pas la méthode la plus évidente à mettre en place. D'un autre côté, nous avons aussi pu remarquer qu'il semblerait plus efficace d'entraîner le réseau de neurones sur un temps plutôt court (par exemple 5 jours) tous les mois, plutôt que de l'entraîner sur une longue période (par exemple

3 semaines) toutes les saisons. Cela permettrait un meilleur échantillonnage des différentes conditions, mais probablement aussi une meilleure représentation de l'évolution de la végétation au cours de l'année. Des recherches plus poussées pourraient permettre de mettre en avant de nouvelles façons de déployer la station d'eddy-covariance de façon à estimer des flux le plus justes possibles.

Pour la topographie du réseau, précédemment, nous nous retrouvions dans un cas de sur-apprentissage en ajoutant des couches de neurones. L'erreur RMSE et le coefficient de corrélation R devenaient moins bons, ce qui nous permettait de conclure que nous n'avions pas besoin de plus d'une couche cachée afin d'obtenir un réseau optimisé. Cependant, nous remarquons que dans le cas d'un jeu plus long, ajouter des couches cachées améliore légèrement le réseau même si cette amélioration n'est pas significative pour notre étude. Utiliser une seule couche cachée et y ajouter progressivement des neurones semble aussi améliorer les résultats obtenus, alors que précédemment, lorsque nous atteignons 8 neurones, l'erreur RMSE et le coefficient R stagnaient et n'évoluaient plus. Il faudrait donc chercher à trouver un entre-deux entre optimisation de la structure et optimisation du temps d'apprentissage, puisque dans le cas du jeu long, ajouter des neurones ou des couches induit un temps d'apprentissage plus important pour une amélioration des résultats peu significative.

Nous n'avons pas eu le temps d'effectuer les tests concernant les variables d'entrée pour ce jeu-ci. Cependant, nous avons pu observé pendant la phase de traitement du jeu de données d'une année une prédisposition du réseau à s'appuyer sur le rayonnement net et non sur le temps afin de reproduire le cycle diurne des flux. De futures recherches permettraient peut-être de mettre en avant l'importance des différentes variables utilisées.

5. CONCLUSION

L'estimation de flux en surface par réseaux de neurones fonctionne : c'est la conclusion majeure que nous pouvons faire suite à nos différentes expériences. Nous avons pu voir que le jeu d'apprentissage semble posséder une influence bien plus importante sur l'estimation de ces flux en surface que la topographie de notre réseau ou même encore que les variables mises en entrée. Mais il reste encore certains tests à effectuer (optimisation du déploiement des stations, de la topographie et étude de l'importance du choix des variables pour un apprentissage d'une année complète) afin de conclure définitivement sur l'importance du choix de ce jeu d'apprentissage pour l'utilisation d'une telle méthode.

Il pourrait aussi être intéressant d'étudier les flux estimés à une échelle plus petite, comme à la journée. En effet, nous avons calculé les coefficients de corrélation et les erreurs RMSE à l'échelle du jeu de données pour avoir une vision globale de l'estimation des flux. Mais les résultats trouvés varient à l'échelle journalière. Comme nous avons pu le voir, certaines journées enregistrent des flux relativement bien estimés alors que d'autres journées proposent des flux qui semblent bien moins correctes. Il serait donc intéressant d'effectuer une telle étude et de confronter ces flux journaliers aux conditions météorologiques à la journée (présence de pluie, grosses chaleurs, vents importants,...).

Dans un futur proche, il serait pertinent de parvenir à appliquer cette méthode sur le site de Lannemezan. Des stations météorologiques classiques et une station d'eddy-covariance devraient être achetées avec pour objectif d'entraîner ces stations classiques pendant une année complète sur différents couverts. Mais avant cela, des recherches supplémentaires seront nécessaires afin de déterminer un certain nombre de consignes pour que cette approche fonctionne. Les tests réalisés sur le jeu de données d'une année et les premières conclusions que nous avons pu en tirer peuvent être un bon point de départ et auront permis de soulever de nouvelles questions, comme par exemple combien de temps une surface doit-elle être entraînée pour estimer de façon convenable des flux tout au long de l'année ? Quelles variables utiliser pour cet apprentissage ? Ou encore : comment organiser l'entraînement de plusieurs stations météorologiques classiques déployées sur différents couverts à l'aide d'une seule station d'eddy-covariance ?

6. CONCLUSION PERSONNELLE

Durant ces six mois de stage, j'ai pu découvrir et travailler sur une nouvelle thématique : l'intelligence artificielle. Et même plus particulièrement sur les réseaux de neurones. J'ai pu appliquer les connaissances acquises durant ces deux années de master à un domaine qui m'était jusqu'à présent inconnu et j'ai vraiment trouvé cela intéressant, enrichissant et avant tout formateur. Je pense avoir gagné en autonomie, mais aussi en confiance et en communication. Des réunions ont été organisées avec des chercheurs américains et parisiens travaillant ou étant intéressés par des recherches dans l'utilisation de l'intelligence artificielle pour l'estimation des flux. J'ai beaucoup apprécié partager nos résultats avec eux, et échanger sur certains points. Leurs avis et leurs conseils ont été d'une grande aide dans l'évolution de notre travail. C'était encore une fois une expérience formatrice qui donne, je pense, un premier aperçu du quotidien d'un chercheur.

J'ai aussi eu l'occasion de participer à une campagne de terrain de mesures aéroportées dans le cadre de la campagne HILIAISE ce qui a été une superbe expérience. Ayant travaillé pendant ces six mois sur des données récoltées pendant de précédentes campagnes, il a été intéressant de découvrir comment ces dites campagnes sont organisées, comment les données peuvent être récoltées. Je remercie de nouveau mes maîtres de stage, Fabienne et Marie, pour m'avoir offert l'opportunité de vivre cette expérience.

De façon générale, ce stage fut une expérience professionnelle valorisante et encourageante qui m'a conforté dans mon avenir professionnel.

BIBLIOGRAPHIE

- [1] **Carolyn Reynolds, Keith Williams, Ayrton Zadra.**
WGNE Systematic Error Survey Results Summary.
February 2019.

- [2] **M. Kumar, N.S. Raghuwanshi, R. Singh.**
Artificial neural networks approach in evapotranspiration modeling : a review.
Review, Irrig. Sci. 2011, 29, 11-25.

- [3] **Jason Kelley, Eric Pardyjak.**
Using Neural Networks To Estimate Site-Specific Crop Evapotranspiration with Low-Cost Sensors.
Article, 23 February 2019.

- [4] **Jason Kelley et al.**
Using Machine Learning to Integrate On-Farm Sensors and Agro-Meteorology Networks Into Site-Specific Decision Support.
Article, *Transactions of the ASABE* 63, no.5 (2020).

- [5] **Howard Demuth, Mark Beale.**
Neural Network Toolbox, For Use With MATLAB®.

- [6] **Campagne BLLAST.**
<https://www7.obs-mip.fr/bllast/>

- [7] **Lothon et al.**
The BLLAST field experiment : Boundary-Layer Late Afternoon and Sunset Turbulence.
Atmos. Chem. Phys., 14, 10931-10960, 2014.

ANNEXES

Annexe A : Fiche résultat type réalisée pour chaque test.

1 MOOR : Tests sur la longueur du jeu d'apprentissage (tsl : training set length)

Pour les tests suivants (nommés tsl pour training set length), nous ne considérons que les iop pour la lande. Ils sont au nombre de douze et sont répartis entre le mois de juin et le mois de juillet 2011 (campagne BLLAST). Voici le détail des IOP utilisés :

14/06, 15/06 ,19/06, 20/06, 24/06, 25/06,

26/06, 27/06, 30/06, 01/07, 02/07, 05/07

Afin de simuler un comportement global de notre réseau, nous le lançons 100 fois (tests préalablement réalisés pour estimer le nombre d'itérations nécessaires) et utilisons les statistiques obtenues afin de l'étudier (ici sont utilisées les médianes).

Tests sur la longueur du jeu d'entraînement (tsl) :			
Nom	Jours	% train	% test
tref	tous les iop	40	60
tsl_6.0_90	14, 15, 19, 20, 24, 25 juin	90	10
tsl_6.1_90	26, 27, 30 juin + 1, 2, 5 juillet	90	10
tsl_6.2_90	14, 15, 26, 27 juin + 1, 2 juillet	90	10
tsl_1.0_90	19 juin	90	10
tsl_1.1_90	26 juin	90	10
tsl_12.8	tous les iop	8	92

1.1 Résultats généraux

Les tableaux sont complétés en utilisant les médianes des 100 itérations. Les RMSE sont en W/m².

1.1.1 Pour la chaleur sensible (H) :

Résultats tests tsl pour la chaleur sensible (H) :						
Nom	RMSE train	RMSE test	RMSE ann	R train	R test	R ann
tref_H	7	10	9	0.990	0.985	0.989
tsl_6.0_90_H	7	9	19	0.992	0.987	0.938
tsl_6.1_90_H	7	9	14	0.991	0.983	0.967
tsl_6.2_90_H	6	9	12	0.993	0.983	0.974
tsl_1.0_90_H	7	11	35	0.993	0.994	0.815
tsl_1.1_90_H	5	8	41	0.992	0.981	0.853
tsl_12.8_H	6	14	13	0.994	0.967	0.969

1.1.2 Pour la chaleur latente (LE) :

Résultats tests tsl pour la chaleur latente (LE) :						
Nom	RMSE train	RMSE test	RMSE ann	R train	R test	R ANN
tref.LE	21	27	25	0.989	0.980	0.983
tsl_6.0_90.LE	22	24	30	0.983	0.982	0.975
tsl_6.1_90.LE	20	23	28	0.990	0.987	0.980
tsl_6.2_90.LE	21	27	27	0.988	0.981	0.980
tsl_1.0_90.LE	22	21	32	0.985	0.993	0.971
tsl_1.1_90.LE	15	22	76	0.993	0.996	0.906
tsl_12.8.LE	19	31	30	0.989	0.973	0.974

1.2 Résultats individuels

1.2.1 tref_H : chaleur sensible (H)

Test sur la longueur du jeu d'apprentissage tref pour la chaleur sensible (H). Nous considérons ici les 12 jours iop (14/06, 15/06, 19/06, 20/06, 24/06, 25/06, 26/06, 27/06, 30/06, 01/07, 02/07, 05/07) et un réseau de neurones à 1 couche cachée contenant 10 neurones. 40% des données sont utilisées pour l'entraînement et 60% pour le test.

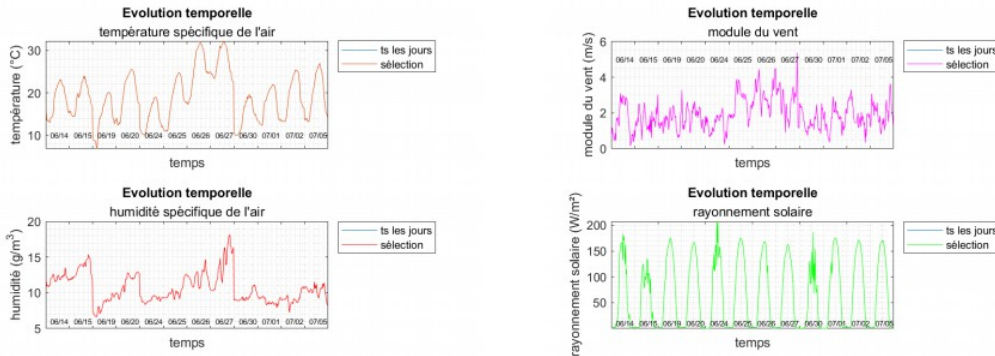


Figure 2: Données météorologiques des journées concernées. Sont surlignées les données qui ont servi à l'entraînement du réseau.

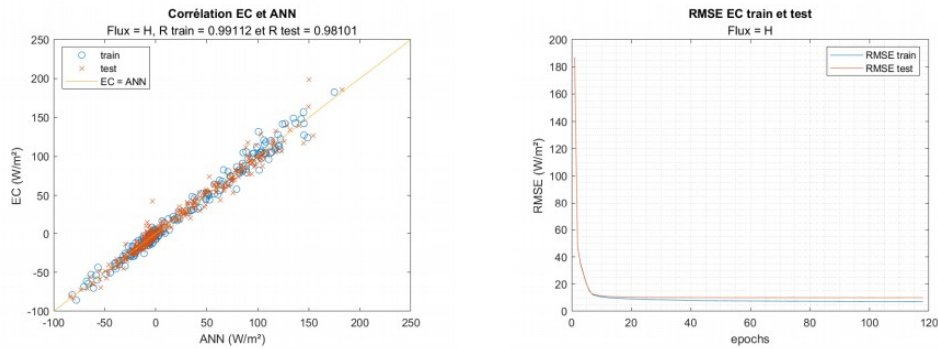


Figure 3: Résultats de l'entraînement du réseau. A gauche, corrélation entre les flux mesurés avec l'eddy-covariance (EC) et ceux mesurés à partir du réseau de neurones (ANN). A droite, évolution du RMSE pendant le training et le test. Cette erreur est calculée en prenant les flux d'eddy-covariance comme référence.

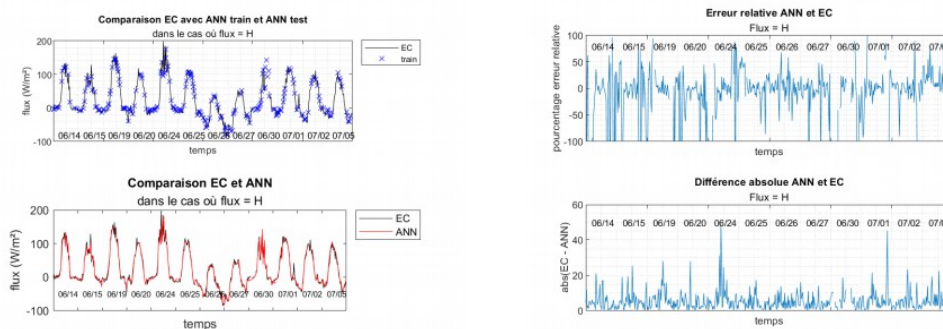


Figure 4: Comparisons des résultats obtenus avec le réseau et l'eddy-covariance. A gauche, la mise en avant des données utilisées pour le training (graphique du haut) et la superposition de l'évolution des flux mesurés par eddy-covariance et ceux estimés avec notre réseau (en bas). A droite, l'erreur relative (en haut) et la différence absolue (en bas) des flux estimés et mesurés.

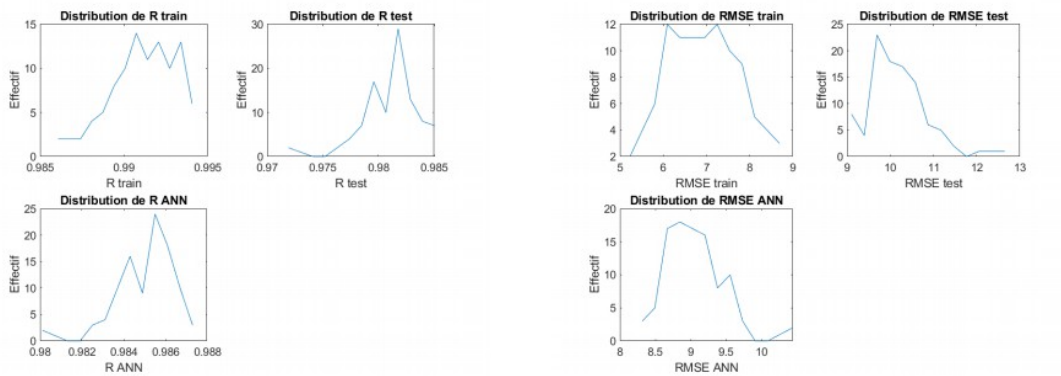


Figure 5: Comparaisons des distributions. A gauche, distributions des coefficients de corrélation R pour le train, le test et le flux ANN estimé. A droite, distributions des erreurs RMSE pour le train, le test et le flux ANN estimé. La référence prise pour le calcul de ces erreurs est le flux mesuré par eddy-covariance.

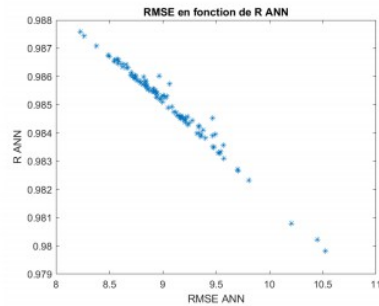


Figure 6: Corrélation entre R ANN et RMSE ANN.

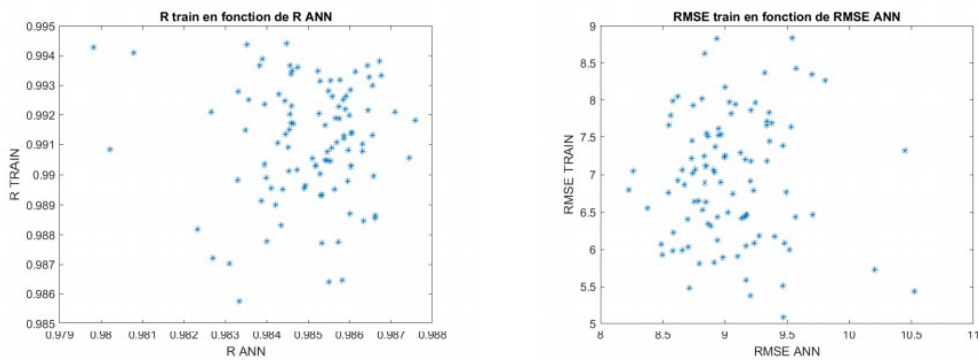


Figure 7: A gauche, corrélation entre R train et R ANN. A droite, corrélation entre RMSE train et RMSE ANN (les erreurs sont calculées par rapport aux flux mesurés par eddy-covariance).

Annexe B : Tableaux récapitulatifs des différents jeu de tests réalisés.

Name	Days	% train	% test
tref	12 days	40	60
tsl_6.0_90	6 days (first ones)	90	10
tsl_6.1_90	6 days (last ones)	90	10
tsl_6.2_90	6 days (regular + singular)	90	10
tsl_1.0_90	regular day	90	10
tsl_1.1_90	singular day	90	10
tsl_12_8	12 days	8	92

Name	Nb of layers	Nb of neurons
tref	1	10
ntt_hl1_n5	1	5
ntt_hl2_n10	2	10
ntt_hl2_n5	2	5
ntt_hl1_n1	1	1
ntt_hl3_n10	3	10

Name	Variables
tref	T, t, rv, ws, SWi
ivt_v3	T, rv, SWi
ivt_v4_tps	T, t, rv, SWi
ivt_v4_ws	T, rv, ws, SWi
ivt_v4_Tg	T, Tg, rv, SWi
ivt_v4_rvg	T, rv, rvg, SWi
ivt_v4_Rnet	T, rv, SWi, Rnet
ivt_v5_Tg+ws	T, Tg, rv, ws, SWi
ivt_v6_Tg+rvg+ws	T, Tg, rv, rvg, ws, SWi
ivt_v7_t+Tg+rvg+ws	T, Tg, t, rv, rvg, ws, SWi

Annexe C : Récapitulatif des résultats obtenus pour le jeu de tests ntt, pour le maïs (flux de chaleur sensible).

