RAPPORT TECHNIQUE

DU MÉMOIRE PRÉSENTÉ COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN TECHNOLOGIE DE L'INFORMATION

SUIVI ET MODÉLISATION DU POTENTIEL HYDRIQUE DU SOL DANS UN CONTEXTE DE STRESS HYDRIQUE : LE CAS D'UNE ÉRABLIÈRE À BOULEAU JAUNE À LA MARGE NORDIQUE DE SA DISTRIBUTION

BLANDINE COURCOT

Mai 2023

Table des matières

Table des figures

Tableaux

Introduction

Ce rapport technique est associé à un mémoire qui a été rédigé dans le cadre d'une maîtrise ès sciences (technologie de l'information) avec un profil science des données[1].

Dans le but d'alléger ce mémoire et d'en améliorer sa lisibilité, la démarche méthodologique qui a été mise en place au cours du projet de recherche est détaillée dans ce rapport. De nombreuses stratégies d'analyse ont été explorées afin, d'une part, d'avoir l'opportunité de nous confronter à une diversité d'approches en science des données, et en particulier à l'analyse des séries temporelles, et d'autre part, de pouvoir surmonter les limites de certaines approches.

Sept grandes phases d'analyse ont été définies :

1. Préparation des données ;
2. Sélection des variables prédictives ;
3. Analyse des changements d'état du potentiel hydrique du sol ;
4. Exploration des séries temporelles avec la méthode de Box-Jenkins ;
5. Test et validation de modèles prédictifs de type ARIMAX (*autoregressive integrated moving average model with exogenous variables*) ;
6. Exploration de modèles prédictifs de type NARX (*non linear autoregressive model with exogenous variables) ;*
7. Application à l'étude de périodes de sécheresse-flash à déficit hydrique (*dry drought*).

Ce rapport va ainsi détailler chacune des phases d'analyse. La dernière phase relative à l'étude des sécheresses-flash est détaillée dans le mémoire.

Nous rappelons la problématique centrale du projet, telle que définie au chapitre 2 du mémoire :

Dans un contexte de stress environnementaux, quelle serait l'évolution temporelle du potentiel hydrique du sol dans une future nouvelle normalité climatique des érablières à bouleau jaune ?

[1] Courcot, B. (2023). *Suivi et modélisation du potentiel hydrique du sol dans un contexte de stress hydrique : le cas d'une érablière à bouleau jaune à la marge nordique de sa distribution.* [Mémoire de maîtrise, Université TÉLUQ].

Cinq questions ont été formulées pour tenter de répondre à notre problématique.

1. Quelles sont les variables climatiques considérées dans notre étude qui ont des conséquences importantes sur l'évolution temporelle du potentiel hydrique du sol mesuré à la Station de biologie des Laurentides entre 2017 et 2020 ?

2. Pouvons-nous définir différents états de ce potentiel hydrique du sol et déterminer les fréquences et les probabilités associées aux changements d'état ainsi mis en évidence ?

3. Est-ce qu'un modèle prédictif de type ARIMAX, développé avec des données provenant de la SBL, permet de prédire l'évolution temporelle du potentiel hydrique du sol de ce même site, en considérant l'influence de variables climatiques exogènes ?

4. Pouvons-nous mettre en évidence des différences au niveau des peuplements étudiés, dans les variations du potentiel hydrique du sol dans un contexte spécifique de stress climatiques comme le stress hydrique ?

5. Dans ce contexte de stress hydrique, sommes-nous capables de prédire l'évolution temporelle du potentiel hydrique du sol du site d'intérêt ?

La Figure 1 résume la méthodologie qui a été mise en place, à partir des données brutes recueillies à la Station de biologie des Laurentides à Saint-Hippolyte (SBL).

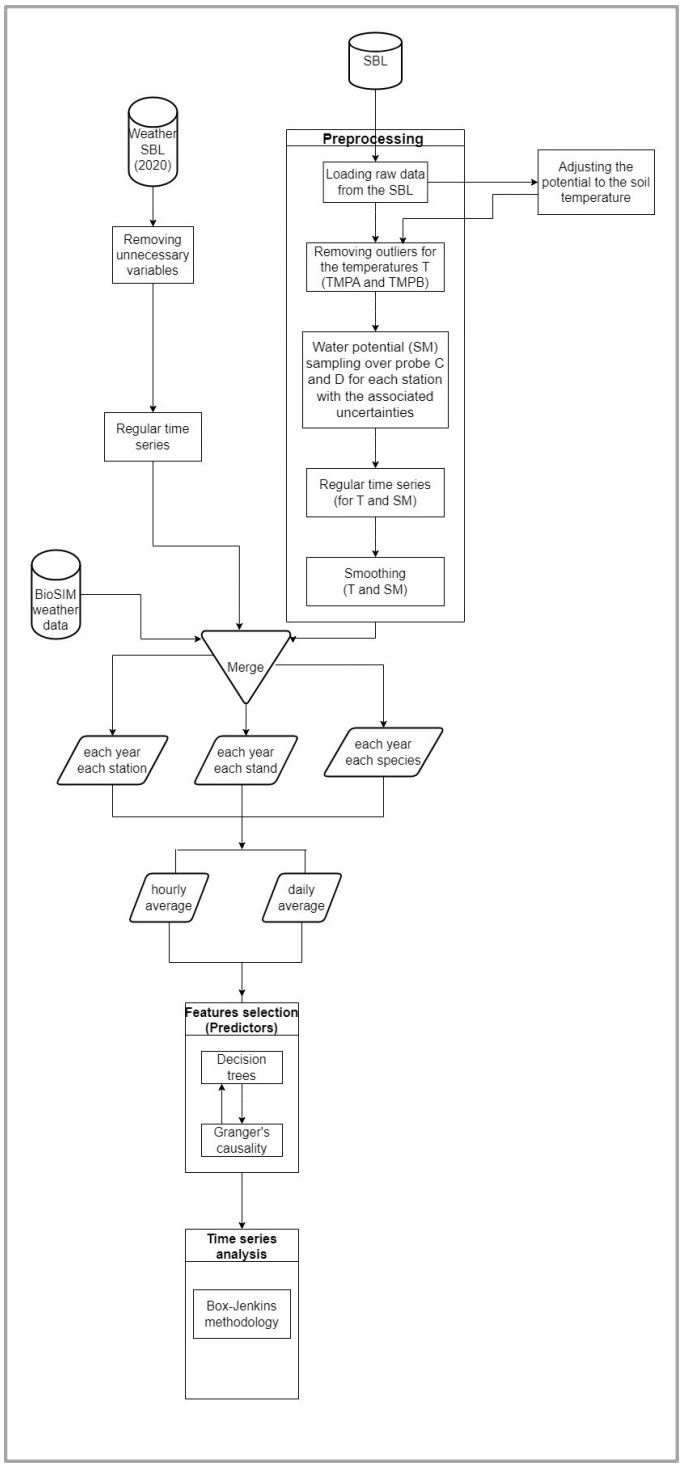


Figure 1 : Résumé des grandes étapes de l'analyse des données réalisée dans la cadre du projet de recherche.

Phase 1 : préparation des données

Description des données brutes

Les données brutes se composent de deux fichiers csv.

Le premier fichier provient des données de la Station de biologie des Laurentides (SBL) collectées directement sur le terrain considéré (data_sbl).

Le second fichier contient les données météorologiques, mesurées par la station météo de la SBL (data_meteo).

data_sbl contient 1 696 473 observations et 9 champs. Il y a 32 parcelles (ou station) considérées, qui sont réparties par bloc (stand). Ces blocs sont composés de différents peuplements d'arbres (species).

Trois *peuplements* ou *espèces* sont représentés : feuillus à dominance de bouleaux (hardwood), feuillus à dominance de hêtre (hardbeech) et feuillus mixtes (mixedwood). Elles sont abrégées par : HW, HB et HW, respectivement.

data_meteo contient 36 798 observations et 48 champs.

La dimension temporelle des données se présente ainsi :

- Fichier de la SBL : de mai 2017 à novembre 2020 / fréquence d'échantillonnage = 15min (0.001 Hz)
- Fichier météo : d'octobre 2019 à janvier 2022 / fréquence d'échantillonnage = 30min (0.0005 Hz)

Nous précisons que les noms des variables sont en grande majorité en anglais, ainsi que les légendes des graphiques.

Étapes de préparation des données

- Extraction des données pour les importer dans le logiciel Matlab R2022a
- Transformation des données brutes
    - Création de variables qualitatives de type *catégorie* :

      Stations (parcelle), Stands (blocs) et Species (espèces ou peuplements)

La répartition des stations selon les blocs et les peuplements se fait de la façon suivante :

```
% Pour les blocs (stands)
stand_01 ={'Station01','Station02','Station03','Station04'};
stand_02 ={'Station05','Station06','Station07','Station08'};
stand_03 = {'Station09','Station10','Station11','Station12'};
stand_04 = {'Station13','Station14','Station15','Station16'};
stand_05 = {'Station17','Station18','Station19','Station20'};
stand_06 = {'Station21','Station22','Station23','Station24'};
stand_07 = {'Station25','Station26','Station27','Station28'};
stand_08 = {'Station29','Station30','Station31','Station32'};
% Pour les espèces (species)

species_mixedwood={'Station01','Station02','Station03','Station04','Station13','S
tation14','Station15','Station16','Station25','Station26','Station27','Station28'
}
species_hardwood={'Station05','Station06','Station07','Station08','Station21','St
ation22','Station23','Station24'}
species_hardbeech={'Station09','Station10','Station11','Station12','Station17','S
tation18','Station19','Station20','Station29','Station30','Station31','Station32'
}
```

- Suppression des données de températures considérées comme aberrantes (inférieures à -20 °C et supérieures à +46 °C) pour les sondes A et B, de façon indépendante.

- Ajustement du potentiel hydrique à la température du sol, selon (Irmark, 2016)[2].

- Échantillonnage fréquentiel des données (potentiel hydrique et température) : moyennage sur 1 heure, pour les sondes similaires (A et B / C et D) et calcul de l'écart-type et de l'incertitude associés à la mesure du potentiel hydrique et de la température.

- Obtention de données régulières (une mesure toutes les heures) avec détermination des données manquantes grâce à une interpolation de type BSpline cubique.

[2] Irmak, S., Payero, J. O., VanDeWalle, B., Rees, S., & Zoubek, G. (2016). *Principles and operational characteristics of Watermark granular matrix sensor to measure soil water status and its practical applications for irrigation management in various soil textures*. University of Nebraska Lincoln Extension

- Lissage des données pour augmenter le rapport signal/bruit en appliquant un filtre de Savitzky-Golay (polynôme de degré 6, avec une fenêtre temporelle de 3 jours). Nous montrons que l'application d'une moyenne mobile seule (de 3 jours) a tendance à réduire les maxima locaux.

Nous avons exploité un avantage spécifique aux données recueillies sur le site expérimental de la SBL : une fréquence d'échantillonnage de 15 minutes, pour les données mesurées entre 2017 et 2020. Cela nous a permis de surmonter la variabilité des données, en échantillonnant les mesures toutes les heures.

Nous présentons dans cette partie une sélection de résultats intéressants, à l'issue de cette première phase d'analyse.

Avant le nettoyage des données, nous observons tout d'abord une répartition assez homogène du nombre d'observations par parcelle (Figure 2). La station01 comporte le plus grand nombre d'observations (54 827), tandis que la station32 en comporte le moins, avec 46 048.

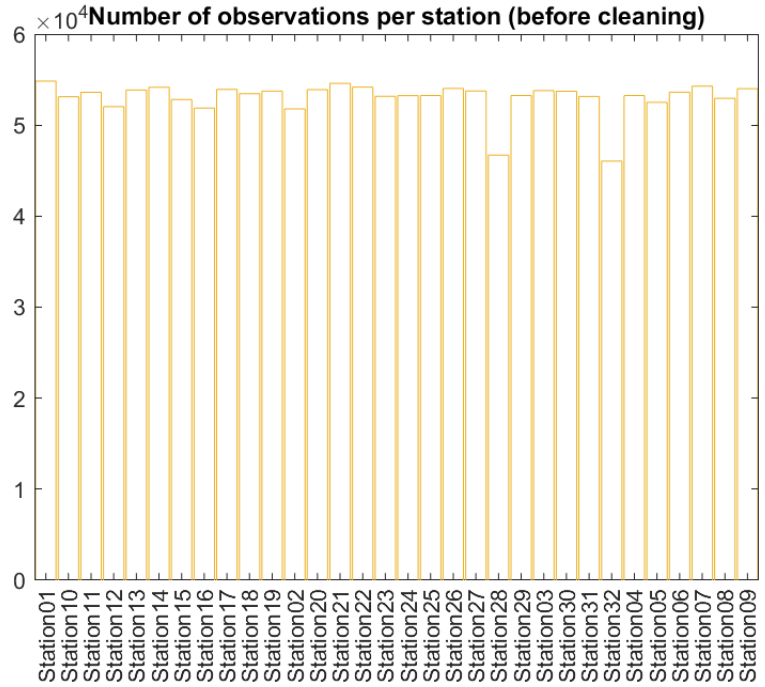


Figure 2 : Nombre d'observations par station.

Nous considérons également la répartition des observations, par bloc (Figure 3), puis par peuplement (Figure 4). L'espèce des feuillus (HW) ne compte que deux blocs, les blocs 2 et 6, ce qui explique un nombre plus faible d'observations.

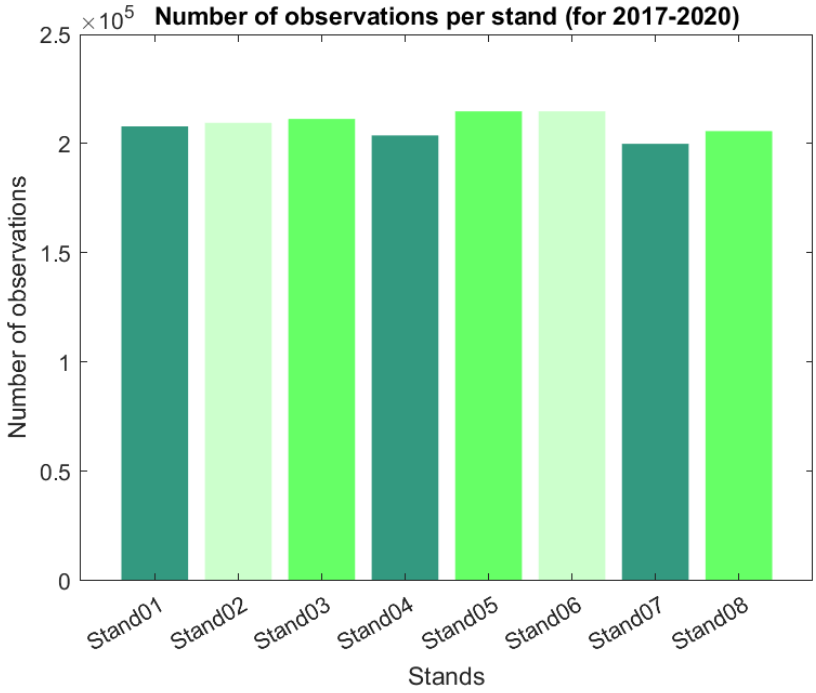


Figure 3 : Nombre d'observations par bloc.

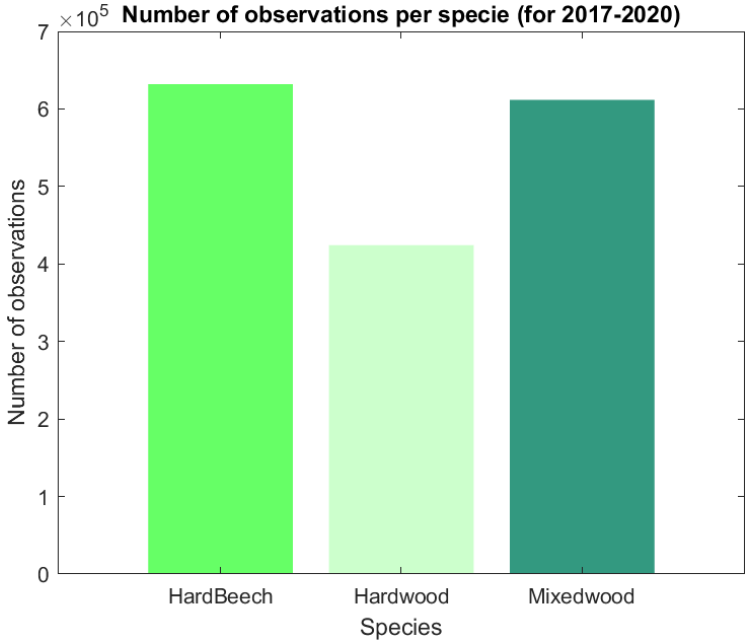


Figure 4 : Nombre d'observations par peuplement.

Après cette première exploration des données, nous allons nous intéresser aux données dites aberrantes (*outliers*). Nous constatons que certaines mesures de température ne sont pas cohérentes et ne peuvent pas exister physiquement (Figure 5). Nous allons les supprimer, et nous faisons le choix de ne pas les remplacer par de nouvelles valeurs qui pourraient être interpolées.

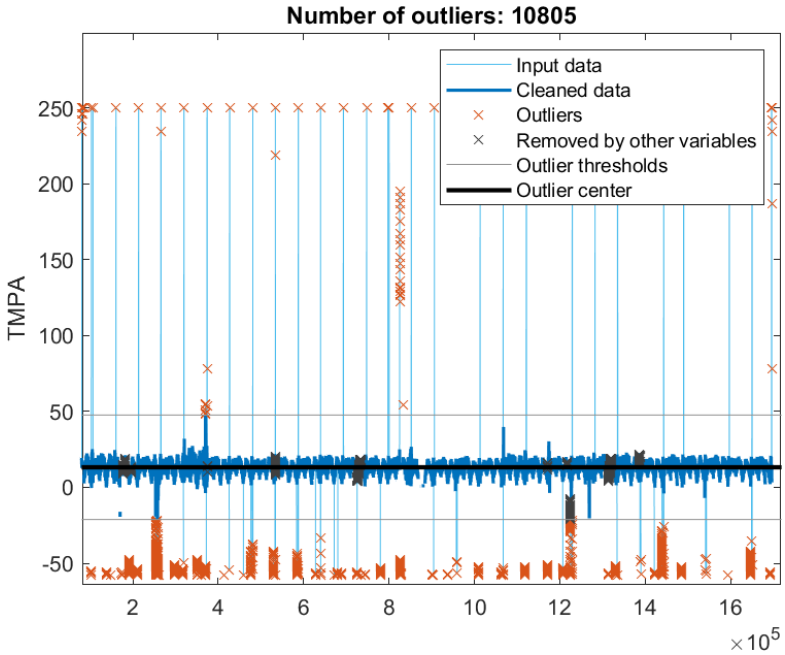


Figure 5 : Étude des mesures de température aberrantes.

Les deux droites en gris sur la Figure 5 représentent les valeurs seuils définies comme critères de sélection des valeurs de température aberrantes (températures inférieures à -20 °C et supérieures à +46 °C). Les points en bleu sont les données d'entrée et les points en orange représentent les observations supprimées.

Il est intéressant de regarder plus en détail la proportion d'observations qui a été retirée à la suite de cette étape (Figure 6).

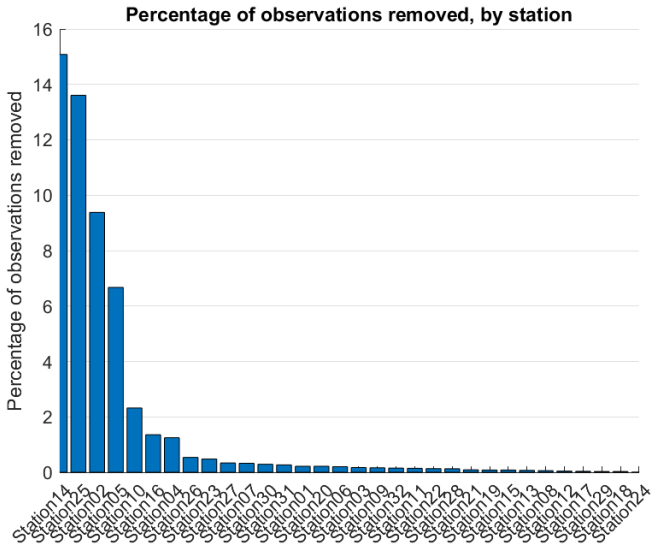


Figure 6 : Proportions des observations enlevées après considération des valeurs de température aberrantes (par station).

Les stations 14, 25, 02 ont entre 9% et 15 % de leurs observations qui ont été considérées comme aberrantes (selon les critères définis précédemment). Pour la grande majorité des stations, moins de 1% des observations ont été supprimées.

Nous allons également vérifier la répartition des mesures, pour les températures et pour le potentiel hydrique. En particulier, nous souhaitons avoir une idée du type de répartition (gaussienne ou non), mais aussi nous assurer que les deux sondes de température et de potentiel donnent des mesures similaires.

Pour le potentiel hydrique, la répartition des observations semble non normale (Figure 7) avec des profils de distributions des données comparables entre la sonde C et la sonde D.

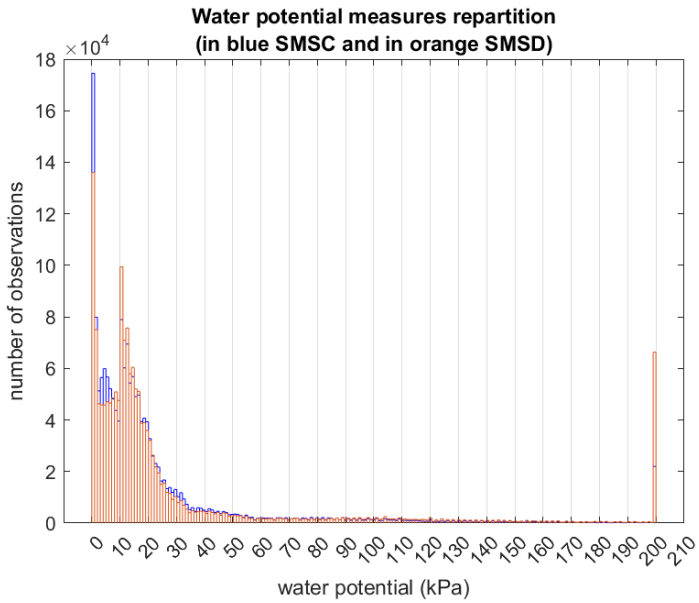


Figure 7 : Répartition des mesures de potentiel hydrique du sol, pour chaque sonde utilisée. Valeurs mesurées par la sonde C en bleu et valeurs mesurées par la sonde D en orange.

La sonde C a tendance à donner davantage de mesures vers les valeurs basses, tandis que la sonde D a une densité de mesures plus grande vers les valeurs plus hautes.

Pour la température, les profils sont très similaires entre la sonde A et B, avec une répartition normale des données (Figure 8).

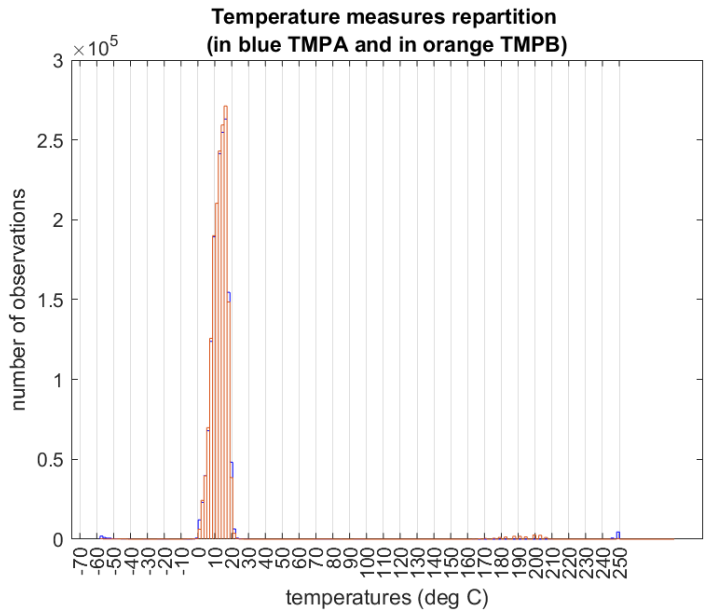
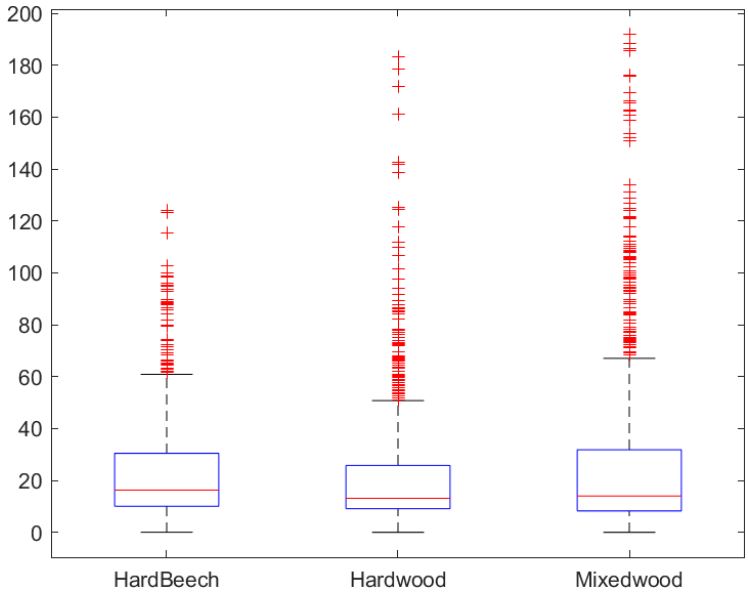
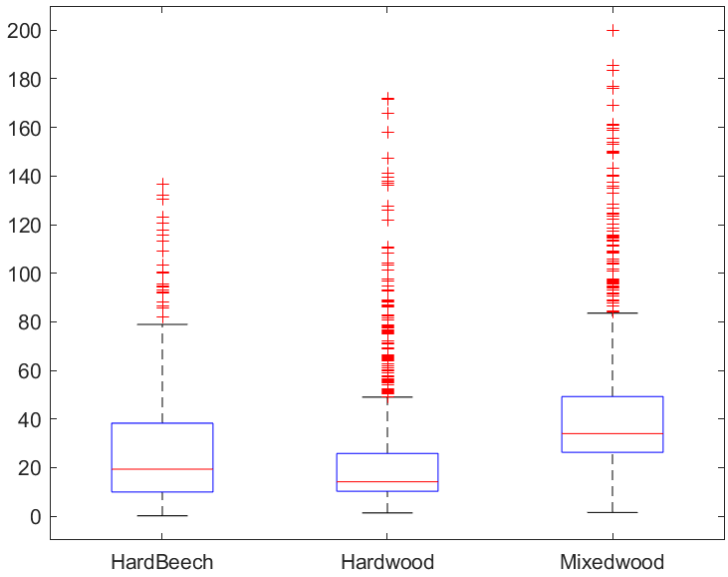


Figure 8 : Répartition des mesures de température du sol, pour chaque sonde utilisée. Valeurs mesurées par la sonde A en bleu et valeurs mesurées par la sonde B en orange.

Nous allons tracer les diagrammes à boîte du potentiel hydrique, pour toutes les années considérées (de 2017 à 2020), pour chaque sonde C et D (Figure 9). Ces diagrammes nous permettent une première analyse statistique, descriptive.



(a) Potentiel hydrique du sol : mesures de la sonde C.



(b) Potentiel hydrique du sol : mesures de la sonde D

Figure 9 : Diagrammes à boîtes du potentiel hydrique du sol, pour toutes les années considérées (de 2017 à 2020), pour (a) la sonde C et (b) la sonde D.

Les valeurs médianes sont représentées par une droite en rouge, le premier et le troisième quartile sont respectivement les bords inférieur et supérieur de la boîte, tandis que les valeurs minimales et maximales sont représentées par les droites en noir.

Les valeurs considérées par défaut comme aberrantes (supérieures à 3.5*écart-type) sont représentées par des croix rouges. Il est important de souligner que ces valeurs ne sont pas aberrantes, elles reflètent les variations extrêmes du potentiel. L'espèce mixte (MW) présente une répartition du potentiel hydrique davantage concentrée autour des valeurs extrêmes, ce qui est en accord avec sa spécificité phénologique. Les valeurs médianes sont quant à elles très similaires entre les trois peuplements.

En traçant le potentiel hydrique du sol, noté Ψ, pour 2020 (Figure 10), mesuré par la sonde C (SMSC en bleu) et mesuré par la sonde D (SMSD en orange), avec une échelle identique, nous pouvons mettre en évidence les variabilités qui existent entre les deux sondes. Pour cet exemple, nous avons choisi la station 01.

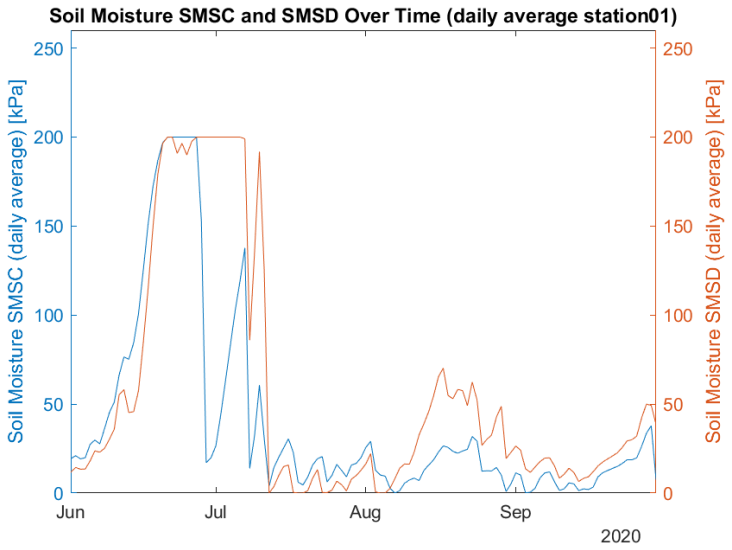


Figure 10 : Représentation du potentiel hydrique du sol mesuré par la sonde C (en bleu) et mesuré par la sonde D (en orange). Station 01, année 2020.

Le profil des deux courbes est similaire et le potentiel mesuré évolue de manière comparable, mais certains extrema varient d'une sonde à l'autre.

Nous allons tracer la différence entre Ψ_C et Ψ_D pour la même période (2020) pour permettre une meilleure visualisation des différences observées (Figure 11).

Nous avons filtré dans un premier temps les différences inférieures à 10 kPa. La Figure 10 met en évidence la diminution du potentiel Ψ_C au début du mois de juillet, en comparaison aux valeurs mesurées par la sonde D. Celles-ci deviennent ensuite inférieures à Ψ_C vers l'autre moitié du mois de juillet. Au mois d'août, nous observons la tendance inverse, et ce jusqu'à octobre.

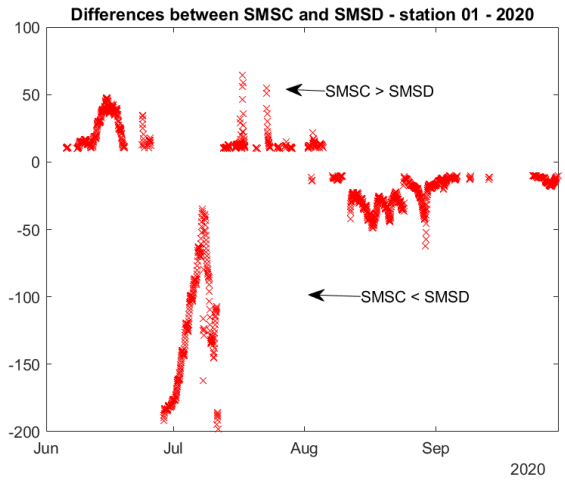


Figure 11 : Différence entre le potentiel hydrique du sol mesuré par la sonde C et le potentiel mesuré par la sonde D. Station 01, année 2020.

De la même manière, nous pouvons visualiser les différences pour la période 2017 – 2020, pour la station 01 (Figure 12). Certains écarts de l'ordre de 200 kPa illustrent un dysfonctionnement d'une des deux sondes.

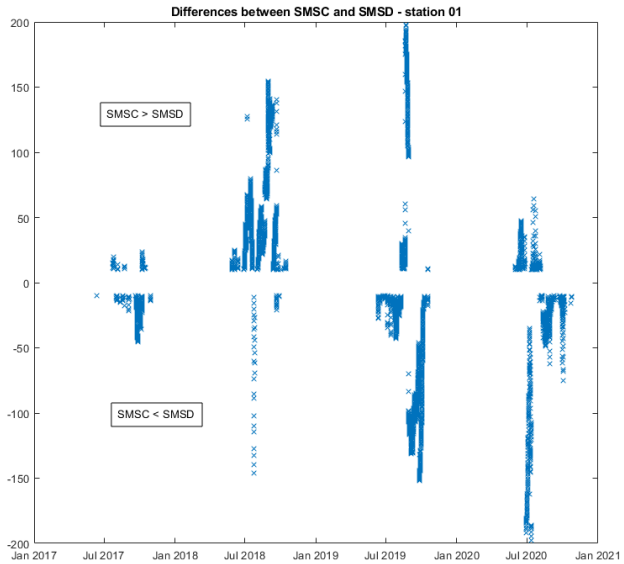


Figure 12 : Différence entre le potentiel hydrique du sol mesuré par la sonde C et le potentiel mesuré par la sonde D. Station 01, années 2017 à 2020.

Après avoir supprimé les données de température aberrantes, pour la station 01, il reste 54 709 observations. Sur ces données, 18 350 présentent une différence ΔΨ > 10 kPa (33,5% des données restantes).

Pour les températures, après avoir enlevé les données aberrantes, nous avons également filtré les données telles que (TMPA – TMPB) > 5°C.

Pour la station 01, sur la période 2017-2020, il n'y a que 5 observations qui répondent à ce critère (Figure 13).

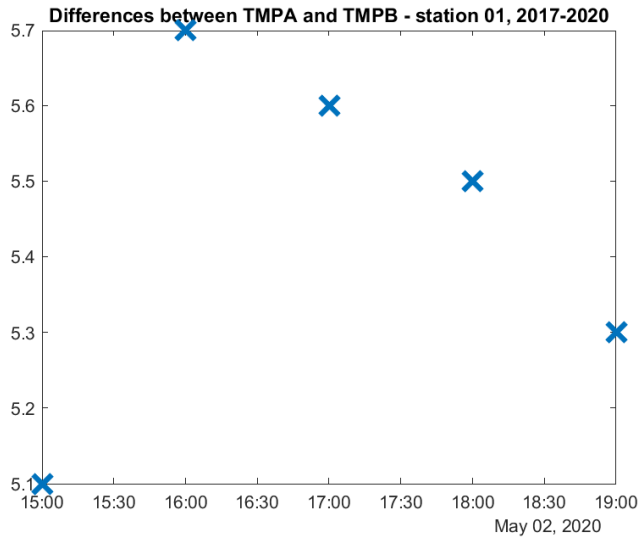


Figure 13 : Différence entre les mesures de température du sol de la sonde A et de la sonde B. Station 01, années 2017 à 2020

Dans une première approche, nous avons récupéré les précipitations relevées à Saint-Hippolyte en 2020 (d'après environnement Québec), et nous les avons représentées avec les moyennes journalières du potentiel hydrique mesuré par la sonde C (Figure 14). Nous pouvons mettre en évidence que des précipitations plus élevées, généralement supérieures à 10mm, sont suivies par une diminution du potentiel hydrique du sol. Cette tendance est plus marquée quand le potentiel hydrique est élevé, par exemple à la fin du mois du juin 2020 et au début du mois de juillet 2020.

Ce graphique exploratoire permet de s'assurer de la cohérence des données de potentiel hydrique du sol.

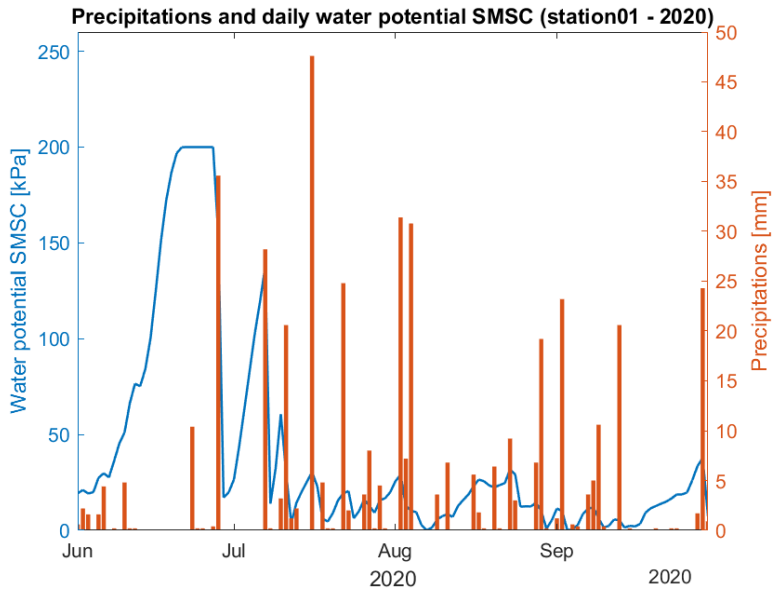


Figure 14 : Représentation du potentiel hydrique du sol mesuré par la sonde C [kPa] et des précipitations [mm], pour la station 01, année 2020.

Nous faisons de même avec les températures du sol, mesurées par la sonde A (Figure 15).

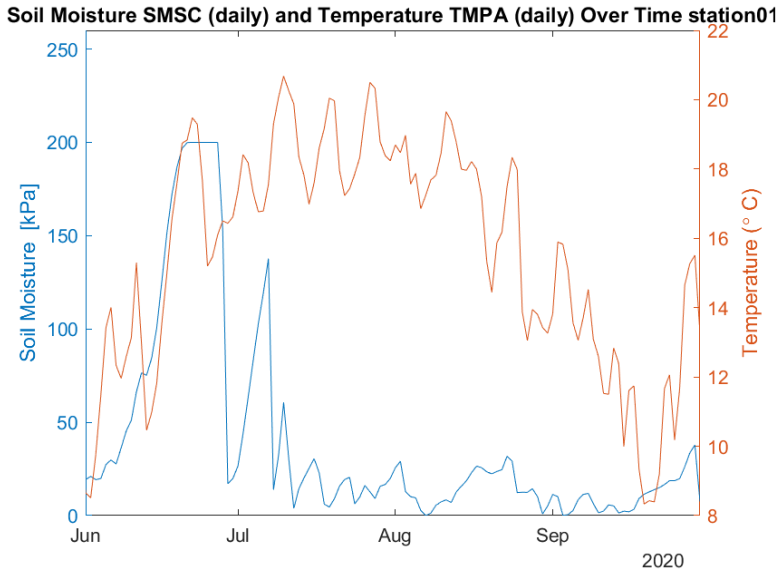


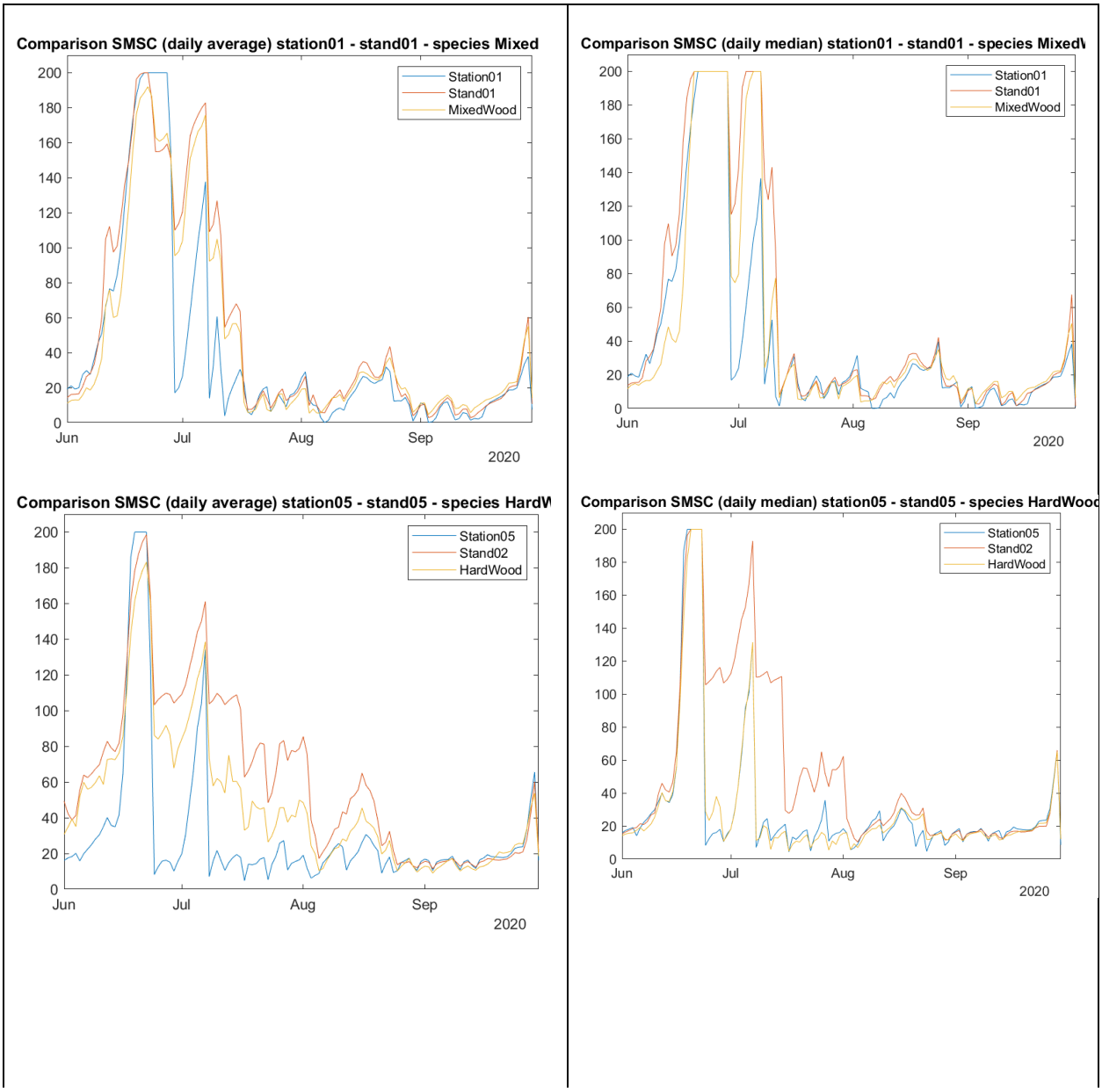
Figure 15 : Représentation du potentiel hydrique du sol mesuré par la sonde C [kPa] et de la température du sol mesurée par la sonde A [°C], pour la station 01, année 2020.

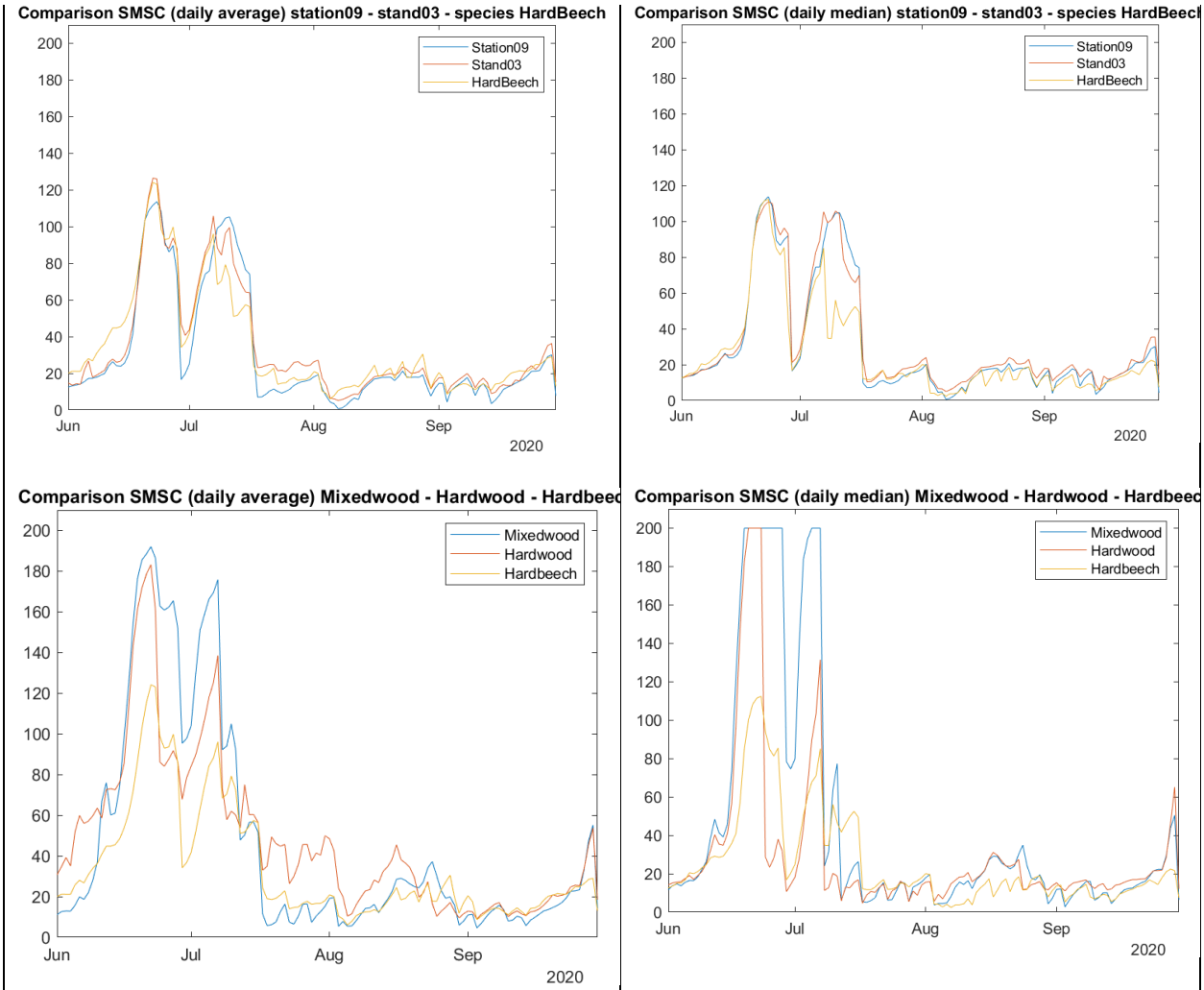
Les températures augmentent au cours du mois de juin 2020, correspondant à une augmentation concomitante du potentiel. Elles restent relativement élevées jusqu'à la fin du mois d'août, pour chuter au cours du mois de septembre.

Nous allons procéder à une analyse hiérarchique en comparant l'évolution du potentiel hydrique mesuré par la sonde C et moyenné sur une journée à différentes échelles : moyenne journalière réalisée à l'échelle de la parcelle (station), l'échelle du bloc (stand), puis à l'échelle du peuplement (species).

La colonne de gauche du Tableau 1 représente les valeurs moyennes du potentiel, la colonne de droite, les valeurs médianes.

Tableau 1 : (A) valeurs moyennes et (B) valeurs médianes du potentiel hydrique du sol mesuré par la sonde C, sur une journée (évolution sur l'année 2020). Trois échelles sont présentées : station, bloc et peuplement.





(A) potentiel hydrique [kPa] journalier moyen pour 2020, à l'échelle des peuplements

(B) potentiel hydrique [kPa] journalier médian pour 2020, à l'échelle des peuplements

Les trois peuplements sont comparés. Les valeurs médianes sont plus élevées que les valeurs moyennes, ce qui s'explique par une densité plus élevée des valeurs mesurées autour des extrema du potentiel (surtout pour l'espèce mixte MW et de feuillus HW).

Sur la Figure 16, nous pouvons comparer le graphique du potentiel hydrique moyen de 2020 dans le Tableau 1 (A) avec le résultat présenté dans (Bélanger, 2021)[3].

[3] Bélanger, N., Collin, A., Khlifa, R. et Lebel-Desrosiers, S. (2021). *Balsam Fir and American Beech influence soil respiration rates in opposite directions in a Sugar Maple Forest near its northern range limit*. Frontiers in Forests and Global Change, 4 :664584. http://dx.doi.org/10.3389/ffgc.2021.664584

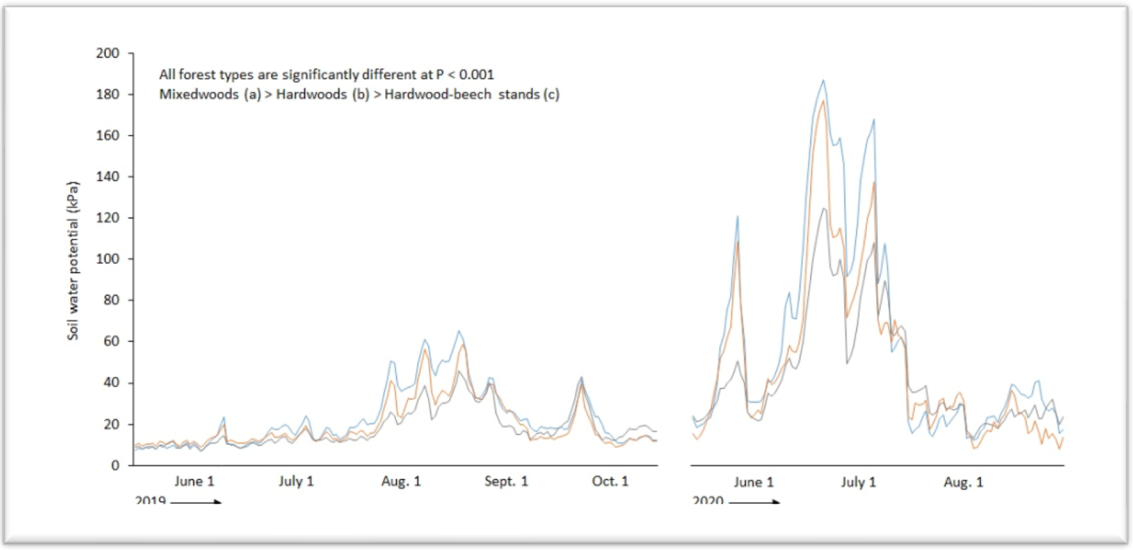


Figure 16 : Matériel supplémentaire de (Bélanger, 2021). Évolution du potentiel hydrique (moyenne journalière) pour les trois peuplements considérés (année 2019 et 2020) : MW (en bleu), HW (en orange) et HB (en gris).

Après avoir échantillonné le potentiel hydrique du sol sur les deux sondes C et D, en considérant une période d'échantillonnage de 1 heure, nous pouvons tracer le potentiel de la station 01, pour 2020, avec l'incertitude associée (Figure 17), voir l'Annexe I pour les détails du calcul.

Pour la station 01, nous pouvons identifier deux périodes temporelles (en 2020) qui semblent présenter une plus grande incertitude du potentiel hydrique. Ces périodes sont repérées par des rectangles verts sur la Figure 17.

Nous pouvons tracer ce même graphique pour les 31 autres stations pour identifier les périodes qui peuvent être plus problématiques, en termes d'incertitudes sur le potentiel hydrique mesuré. Cette approche à l'échelle de chaque station permet d'identifier les sondes problématiques qui nécessitent un suivi technique.

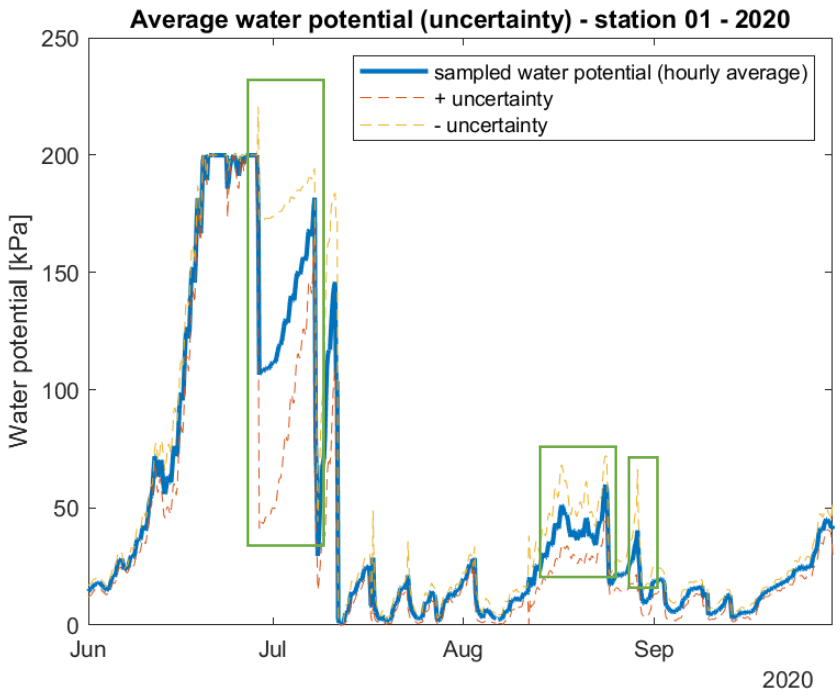


Figure 17 : Potentiel hydrique du sol moyenné sur la sonde C et la sonde D, avec les incertitudes associées (station01, année 2020).

En traçant Ψ_C et Ψ_D mesurés en 2020 (Figure 18), les différences sont mises en avant, ce qui explique alors les incertitudes plus grandes pour certaines périodes.

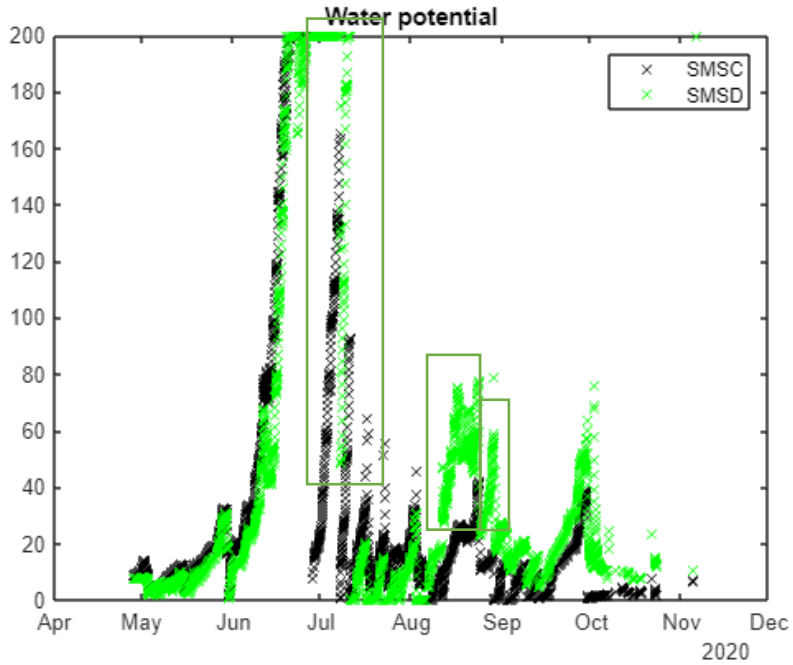


Figure 18 : Potentiel hydrique du sol mesuré par la sonde C (en noir) et la sonde D (en vert), station01, année 2020.

Nous procédons ensuite à une interpolation de type BSpline cubique, pour interpoler les données temporelles manquantes (se référer à l'Annexe II). L'objectif est *d'obtenir des données espacées régulièrement dans le temps, ce qui va permettre l'analyse des séries temporelles*. Nous souhaitons obtenir un point temporel, toutes les heures, sur une période contenue dans l'intervalle de mai à novembre.

Sur la Figure 19 et la Figure 20 sont représentées en bleu les données originales, échantillonnées sur les deux sondes de mesure respectives, sur une heure. Nous constatons, pour la température (Figure 19), et le potentiel hydrique du sol (Figure 20) que les données originales comptent environ 1000 points temporels : des jours sont manquants. Après l'interpolation, nous obtenons 4227 points temporels, représentés en orange. Nous constatons que l'évolution des variables au cours du temps est reproduite après l'interpolation.

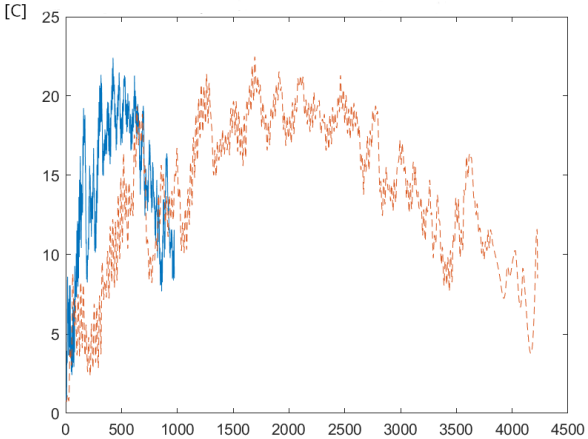


Figure 19 : Mesures de la température du sol [°C] avant régularisation (en bleu) et après régularisation (en orange, trait en pointillé). Données échantillonnées sur les deux sondes, sur une heure. Station 01, année 2020.

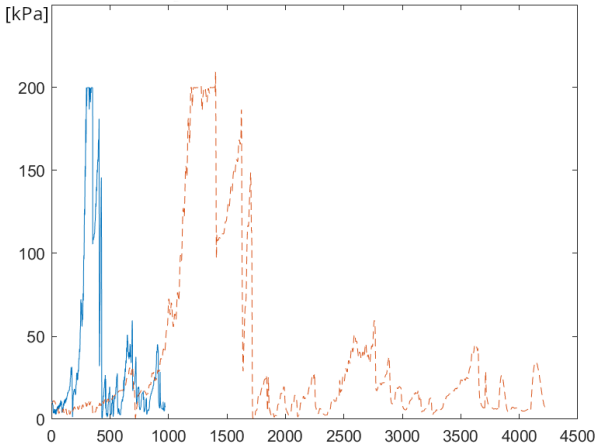


Figure 20 : Mesures du potentiel hydrique du sol [kPa] avant régularisation (en bleu) et après régularisation (en orange, trait en pointillé). Données échantillonnées sur les deux sondes, sur une heure. Station 01, année 2020.

La prochaine étape consiste à appliquer un filtre qui permettra de lisser les données, c'est-à-dire d'augmenter le rapport signal sur bruit. Il s'agit ici de ne pas introduire de distorsion dans les données, mais de permettre une bonne approximation du signal d'origine, qui est déjà moyenné et interpolé.

Nous avons testé plusieurs filtres, et le filtre Savitzky-Golay (SG) nous semble le plus performant (se référer à l'Annexe III). En effet, associé à une moyenne mobile sur 3 jours, un filtre SG s'appuyant sur un polynôme de degré 6, permet de ne pas tronquer les pics du potentiel hydrique (Figure 21).

Nous pouvons comparer le filtre SG avec uniquement l'application d'une moyenne mobile (sur 3 jours, centrée), tel que représenté sur la Figure 22 : les maxima sont tronqués dans le cas de la moyenne mobile seule.

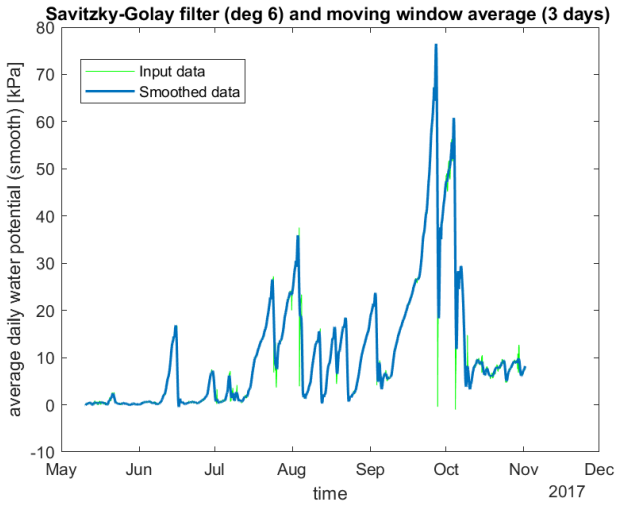


Figure 21 : Application d'un filtre de Savitzky-Golay. Comparaison entre les données avant l'application du filtre (en vert) et après (en bleu). Année 2017.

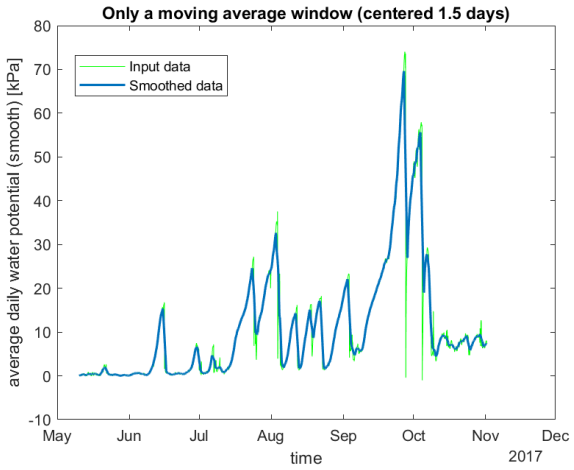


Figure 22 : Filtrage uniquement par moyenne mobile. Comparaison entre les données avant l'application du filtre (en vert) et après (en bleu). Année 2017.

Dernière exploration : nous avons voulu vérifier la variabilité journalière du potentiel hydrique. En effet, nous voulions avoir une idée de l'évolution journalière du potentiel dans notre contexte d'étude.

Nous avons pour cela considéré le potentiel moyenné sur 1 heure (échantillonné), rendu régulier, puis lissé. Nous avons extrait le potentiel à différentes heures de la journée sur toute la période observée de 2020, pour trois stations différentes, représentant chacune un peuplement différent : station 01 pour le peuplement mixte (MW), station 05 pour les feuillus (HW) et station 09 pour les feuillus à dominance de hêtres (HB).

Pour les trois peuplements, le potentiel hydrique semble ne pas varier de façon significative au cours d'une même journée (Figure 23).

Notons cependant une augmentation du potentiel à 17h, le 17 juillet (Figure 23, courbe verte), observée à la station 01 et 05, et légèrement à la station 09. Si nous nous référons aux données des précipitations, le 17 juillet 2020 correspond au lendemain d'une journée de pluie plus abondante.

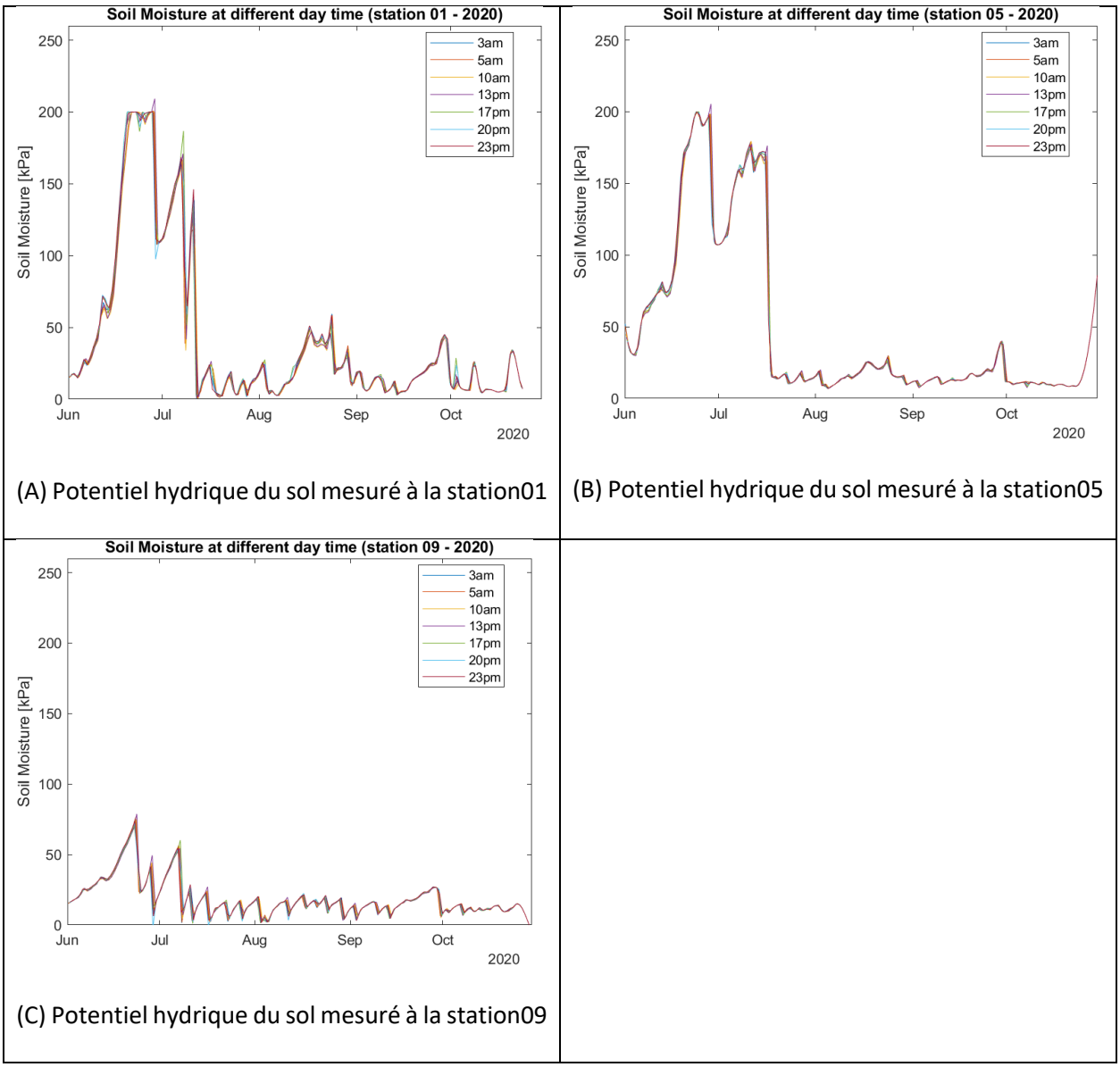


Figure 23 : Potentiel hydrique du sol mesuré à différentes heures de la journée : (A) à l'échelle de la station 01, (B) à l'échelle de la station 05 et (C) à l'échelle de la station 09 (année 2020).

Ajustement du potentiel hydrique en fonction de la température du sol

Faisant écho aux recommandations du constructeur, Irmak et *coll.* (2016)[4] ont en effet mis en évidence la nécessité d'ajuster le potentiel hydrique aux températures du sol, en l'augmentant de 1% pour chaque degré inférieur à 70 ^{o}F (ce qui correspond à 21 ^{o}C), ou en le diminuant de 1% pour chaque degré supérieur à 70 ^{o}F.

Ils proposent l'équation suivante, pour l'ajustement du potentiel hydrique mesuré Ψ:

$$\Psi_{adj} = \Psi - (T_{sol} - 70^{0}F)\ x\ 0.01\ x\ \Psi$$

La température du sol T_{sol} s'exprime en [^{o}F] et Ψ, Ψ_{adj} en [kPa].

Entre mai et novembre, les températures du sol mesurées à la SBL se situent dans l'intervalle [5;15] ^{o}C, en dessous de la température de 21 ^{o}C suggérée par Irmak et *coll.* (2016).

Ainsi, nous avons opté pour un ajustement de chaque valeur du potentiel hydrique par rapport à la température du sol mesurée au même moment, en nous appuyant sur l'équation proposée par (Irmak, 2016).

Nous pouvons alors comparer les valeurs sans et avec ajustement, pour les trois peuplements étudiés : les valeurs obtenues sont plus élevées que les valeurs non ajustées (Figure 24, Figure 25 et Figure 26).

Conclusion : *La suite des analyses se fera avec le potentiel hydrique du sol ajusté.*

Note importante :

Les sondes Watermark permettent d'obtenir une mesure positive du potentiel hydrique du sol, comprise entre 0 et 200 kPa.

Ces mesures sont reportées telles quelles par l'appareil de mesure. Ainsi, une valeur basse du potentiel hydrique tendra vers 0kPa et correspondra à un sol saturé en eau. À l'opposé, une valeur élevée du potentiel hydrique tendra vers 200kPa et correspondra à un sol sec. Cette valeur correspond à une

[4] Irmak, S., Payero, J., VanDeWalle, B., Rees, J., Zoubek, G., Martin, D., Kranz, W., Eisenhauer, D. et Leininger, D. (2016). Principles and operational characteristics of watermark granular matrix sensor to measure soil water status and its practical applications for irrigation management in various soil textures. EC783, University of Nebraska Lincoln Extension.

pression négative de -200 kPa, qui s'apparente à une tension. En d'autres termes, pour l'arbre, la force nécessaire pour extraire l'eau du sol est importante.

Par convention, et par souci de clarté, nous opterons pour une représentation positive des valeurs du potentiel hydrique, pour garder la cohérence des mesures données par les sondes Watermark.

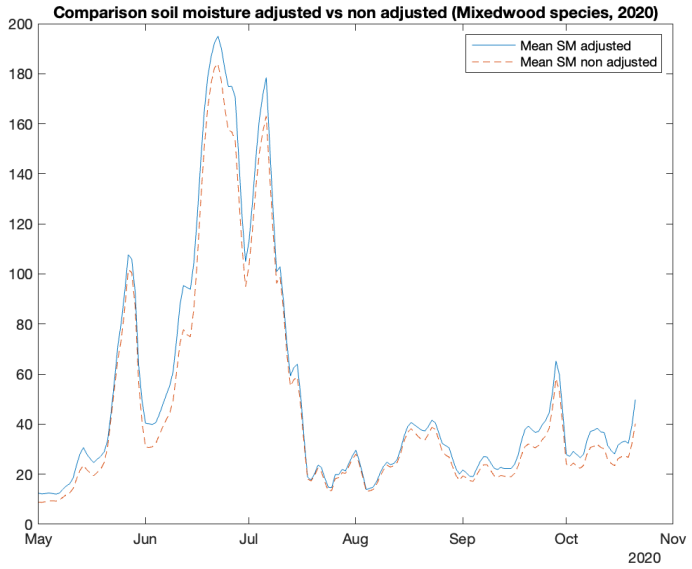


Figure 24 : Comparaison entre le potentiel hydrique du sol ajusté (en bleu) et non ajusté (en orange, trait en pointillé). Peuplement MW, année 2020.

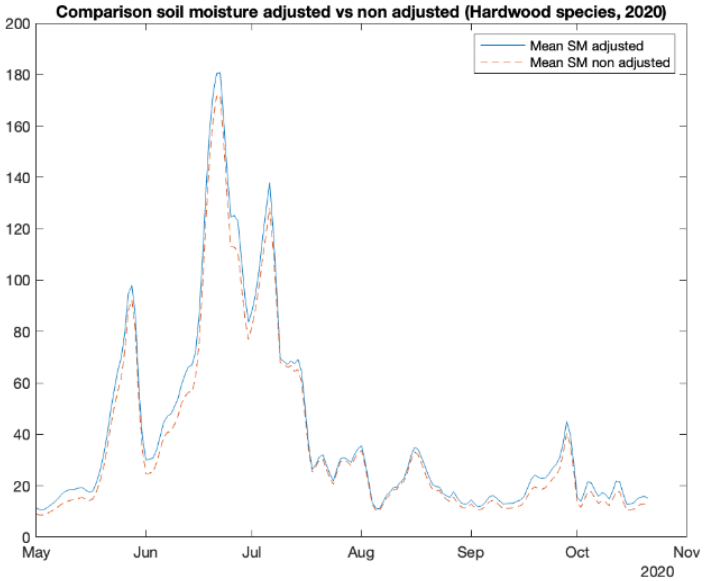


Figure 25 : Comparaison entre le potentiel hydrique du sol ajusté (en bleu) et non ajusté (en orange, trait en pointillé). Peuplement HW, année 2020.

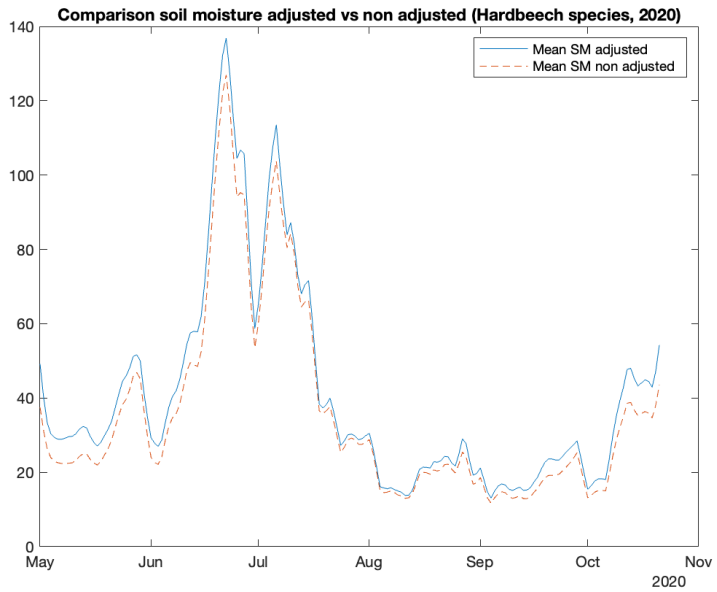
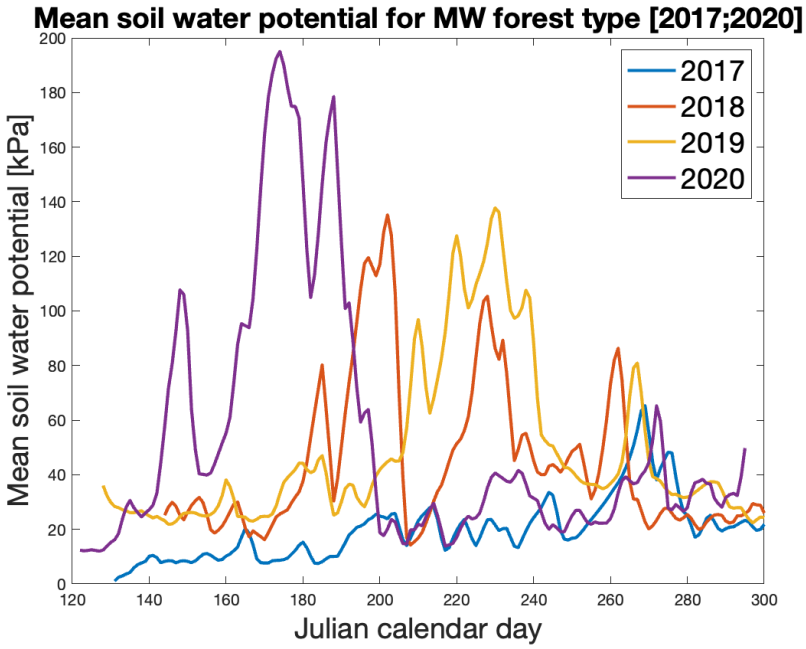
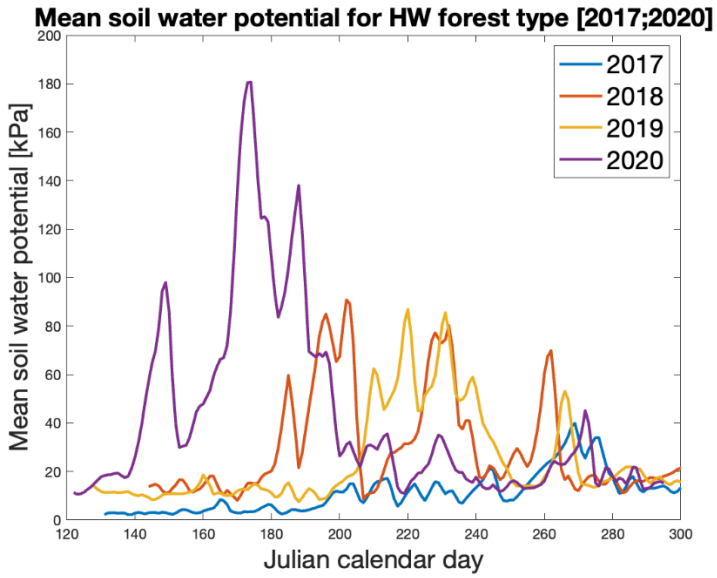


Figure 26 : Comparaison entre le potentiel hydrique du sol ajusté (en bleu) et non ajusté (en orange, trait en pointillé).

Peuplement HB, année 2020.

Nous obtenons ainsi, pour les trois peuplements considérés, entre 2017 et 2020, les courbes de potentiel hydrique ajusté représentées sur la Figure 27.





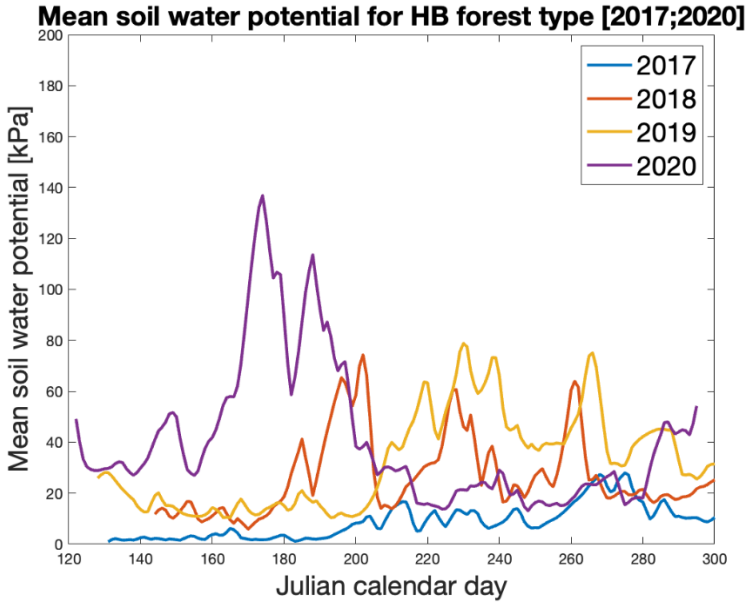


Figure 27 : Évolution du potentiel hydrique du sol entre 2017 et 2020, pour les trois peuplements étudiés (MW, HW et HB).

Pour chacun des trois peuplements, le potentiel hydrique mesuré en 2020 présente des valeurs plus élevées entre mai et août. En 2017, le potentiel présente un potentiel hydrique qui ne dépasse pas les 60kPa. L'année 2017 est la première année de l'expérimentation *SmartForests,* les données sont peut-être à considérer avec prudence, les sondes utilisées étant en période de rodage.

Pour mieux appréhender les différences observées entre les quatre années, nous allons étudier plus précisément les variables météorologiques accessibles.

Préparation et exploration des données météorologiques

Données générées par Biosim (de 2017 à 2020)

Les données météorologiques mesurées directement à la SBL ne sont disponibles que pour l'année 2020. Ainsi, il a été nécessaire de générer les données à partir du logiciel BioSIM[5]. Les détails de cette étape sont précisés à l'Annexe IV.

Pour l'année 2020, nous pouvons alors tracer l'évolution de l'humidité relative et de l'irradiance solaire (Figure 28).

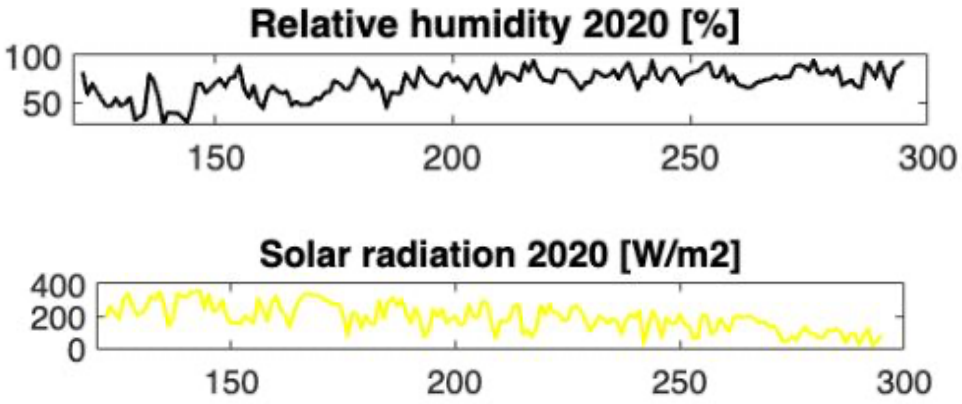


Figure 28 : Évolution de l'humidité relative et de l'irradiance solaire (2020).

Le diagramme ombrothermique de l'année 2020 permet également de visualiser l'évolution de la température de l'air (journalière) en mettant en perspective les précipitations (Figure 29).

[5] Régnière, J. et Saint-Amand, R. (2017). BioSIM 11. Rapport d'information LAU-X-129, Ressources naturelles Canada, Service canadien des forêts, Centre de foresterie des Laurentides

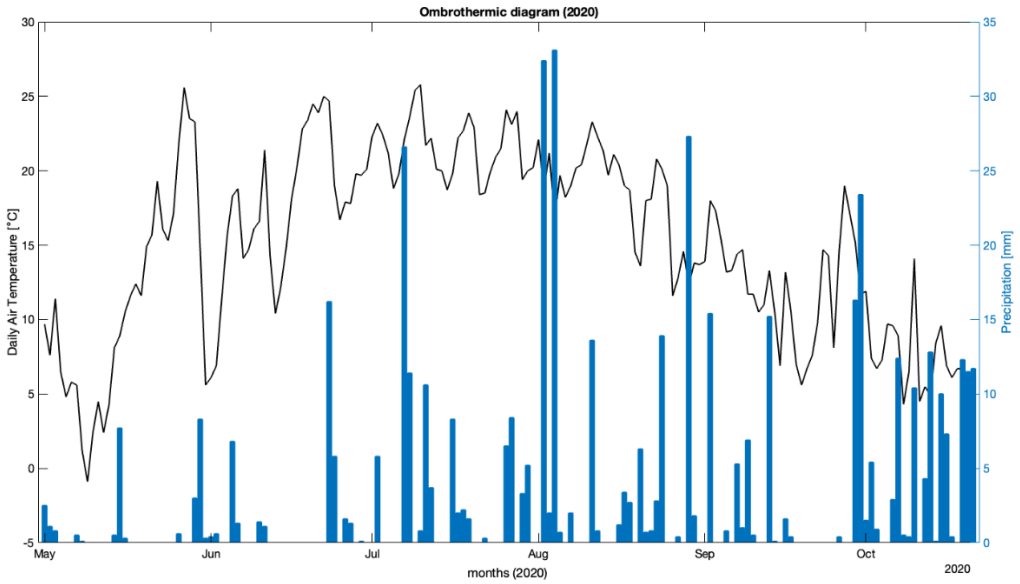


Figure 29 : Diagramme ombrothermique (2020).

Nous allons également représenter les précipitations mensuelles, pour les 4 années étudiées, mettant ainsi en évidence que l'année 2020 a été l'année avec le moins de précipitations, pour la période allant de début mai à fin octobre (Figure 30).

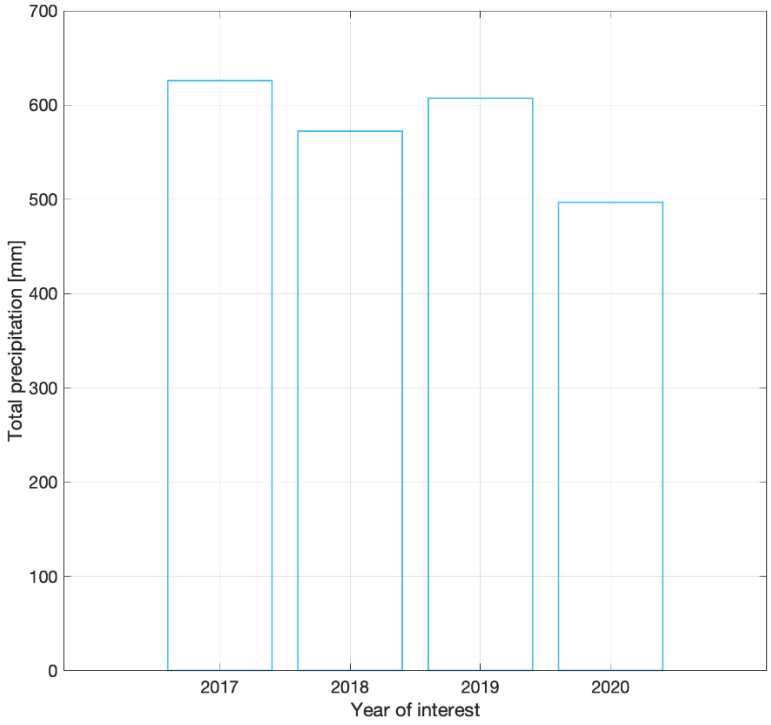


Figure 30 : Précipitations mensuelles cumulées entre mai et octobre (2017 - 2020)

Nous pouvons également comparer la quantité de précipitations par mois, pour ces mêmes années (Figure 31).

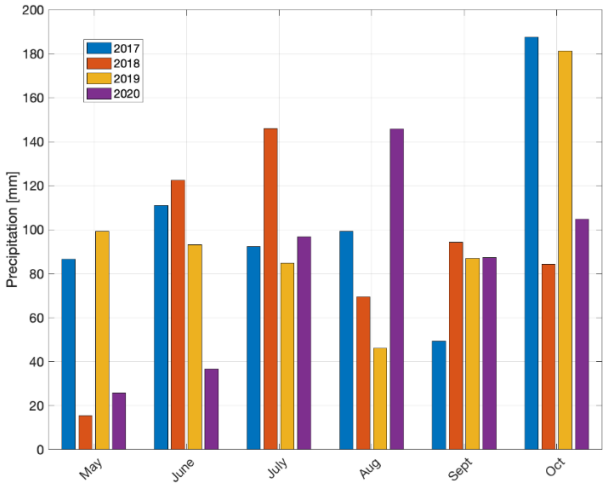


Figure 31 : Précipitations mensuelles (2017 -2020).

Données de la station météo de la SBL pour l'année 2020

Les données accessibles de la station météo de la SBL permettent d'obtenir la direction et la vitesse journalière du vent.

Nous pouvons tracer la rose des vents correspondante, pour l'année 2020, et mettre en évidence la direction dominante des vents (SE) (Figure 32).

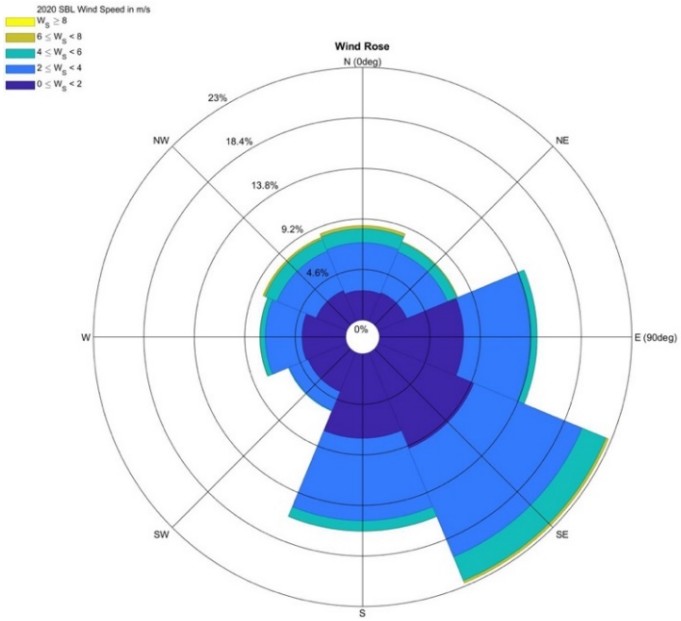


Figure 32 : Rose des vents et direction dominante des vents en 2020.

Analyse de cette première étape de préparation des données

Constat

Ce qui ressort de cette première étape exploratoire, c'est une grande variabilité des données initiales. Ce constat est un enjeu qui ressort en science des données, et en particulier en sciences de l'environnement[6].

Dans notre contexte, nous avons mis en évidence des variabilités entre les mesures effectuées par les différentes sondes, de température et de potentiel hydrique.

Pour pallier cette variabilité dans les données, notre stratégie d'analyse repose sur deux spécificités expérimentales :

- une fréquence d'échantillonnage très haute qui correspond à une période de 15 minutes
- la possibilité d'augmenter l'échelle d'analyse : de la parcelle au peuplement

Nous avons développé notre pipeline d'analyse pour pouvoir mener facilement les analyses aux trois échelles accessibles : au niveau de la parcelle, du bloc et du peuplement.

Les choix que nous avons faits

La moyenne journalière du potentiel hydrique sera utilisée. Nous avons en effet montré que l'heure du jour n'implique pas de grande variabilité dans la mesure. Néanmoins, les valeurs échantillonnées sur une heure sont également disponibles et permettront une analyse plus fine des variations du potentiel hydrique du sol.

Nous avons de plus préparé les données initiales afin d'accéder à trois échelles d'analyse.

Enfin, nous avons fait le choix de régulariser les données en remplaçant les données manquantes par des valeurs interpolées**. L'obtention de séries temporelles dites *régularisées* est une condition importante pour la suite de l'analyse.**

[6] Blair, G.S, Henrys, P., Leeson A., Watkins, J., Eastoe, E., Jarvis, S., Young, P.J. (2019) Data science of the natural environment: a research roadmap. *Frontier in Environmental Science, 7*, 121.

Les limites de cette étape de préparation des données

Une des limites principales vient de notre choix de moyenner les données pour considérer l'échelle du bloc et du peuplement. Or, nous avons également déterminé les incertitudes associées aux mesures du potentiel hydrique du sol. Celles-ci ont été calculées pour chaque station.

Les outils développés et disponibles

La préparation des données constitue une étape importante qui permet une bonne connaissance initiale des données qui seront exploitées par la suite. En nous assurant de la validité et de la cohérence de celles-ci, nous pourrons déterminer des modèles prédictifs plus fiables. La visualisation de données est centrale dans cette compréhension. C'est pourquoi nous avons proposé de nombreuses figures qui ont permis de mettre en lumière certaines spécificités des données sur lesquelles repose notre étude.

Mais cela a permis également de baliser de futures études qui vont exploiter les mêmes données brutes.

Un pipeline a été développé sous Matlab. Il permet de préparer les données issues de la SBL, en gardant leur particularité temporelle.

En effet, notre objectif est de tester et valider un ou plusieurs modèles prédictifs du potentiel hydrique du sol. Pour y arriver, il nous faut une seconde étape qui repose sur l'exploration d'approches dédiées aux séries temporelles.

Validation des données météorologiques générées par BioSIM

Une validation des données météo générées par le logiciel BioSIM a été réalisée grâce à une comparaison avec les données recueillies par la station météo de la SBL en 2020. Cinq constations ont été faites :

(1) Les précipitations ne sont pas fournies par les données de la station météo de la SBL.

(2) L'humidité relative de l'air générée par BioSIM est plus basse que celle relevée à la SBL, mais les mêmes tendances sont observées.

(3) La température de l'air générée par BioSIM suit les mêmes tendances que les mesures faites à la SBL, mais elle est toujours plus élevée.

(4) La température du sol moyennée sur les deux sondes de mesures, pour chaque peuplement considéré (MW, HW, HB), suit la tendance des températures mesurées à la station météo de la SBL. Cela permet de s'assurer de la validité des mesures réalisées sur chaque parcelle.

(5) Le graphique montrant l'évolution de la quantité d'eau dans le sol mesurée à différentes profondeurs (à la station météo de la SBL) avec sur le même graphique les précipitations obtenues avec BioSIM, permet de valider la pertinence des précipitations. En effet, nous constatons qu'après chaque pluie, la quantité d'eau augmente, parfois faiblement, mais un pic est observable après chaque événement.

Phase 2 : évaluation des prédicteurs pertinents (exploratoire)

Cette deuxième phase a pour objectif d'évaluer l'influence des variables exogènes sur le potentiel hydrique moyen, plus précisément appelé, le potentiel matriciel du sol. Les deux noms seront utilisés.

Nous exploiterons les données relatives aux trois peuplements étudiés, échantillonnées sur une journée. L'année 2020 sera analysée plus en détail.

Nous avons évoqué dans la phase 1 les données météorologiques accessibles.

Nous avons vu que les données météorologiques des années 2017, 2018, 2019 et 2020 ont pu être générées à l'aide du logiciel BioSIM.

Nous résumons dans le Tableau 2 les champs disponibles.

Tableau 2 : Champs disponibles générés par BioSIM.

Variable	Notation	Unité
Température de l'air (moyennée sur une journée)	T_{air}	[°C]
Température minimale (journalière)	T_{min}	[°C]
Température maximale (journalière)	T_{max}	[°C]
Température du point de rosée	T_{dew}	[°C]
Précipitations	Prcp	[mm]
Humidité relative	RelH	[%]
Irradiance solaire	SRad	[$W.m^{-2}$]

Nous pouvons définir **des paramètres intégrateurs** qui permettent d'incorporer les différentes variables météorologiques disponibles. Un paramètre est pertinent dans le cadre de notre étude, au regard des variables météorologiques accessibles. C'est le déficit de pression de vapeur (*vapor pressure deficit*, en anglais, ou VPD). Il se définit comme la différence entre la quantité d'eau que l'atmosphère pourrait contenir (100% d'humidité) à une température donnée et la quantité réellement présente dans l'atmosphère. Il correspond schématiquement à la place potentiellement disponible dans l'air pour

davantage de vapeur d'eau, à une température donnée. Nous avons détaillé son calcul dans le mémoire associé à ce rapport [7].

Une fonction permettant de calculer le déficit de pression de vapeur en [kPa], pour les données journalières, a été implémentée.

Nous pouvons représenter l'évolution temporelle du déficit de vapeur de pression de 2017 à 2020 (Figure 33 à Figure 36).

- Potentiel hydrique journalier pour chacun des trois peuplements et évolution du VPD (2017)

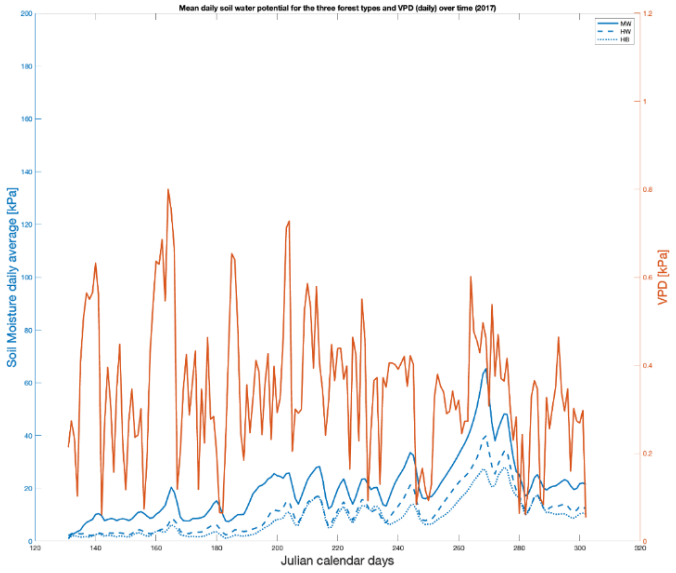


Figure 33 : Évolution du VPD et du potentiel hydrique du sol pour 2017.

[7] Courcot, B. (2023). *Suivi et modélisation du potentiel hydrique du sol dans un contexte de stress hydrique : le cas d'une érablière à bouleau jaune à la marge nordique de sa distribution.* [Mémoire de maîtrise, Université TÉLUQ].

- Potentiel hydrique journalier pour chacun des trois peuplements en bleu et évolution du VPD en orange (pour l'année 2018). Le peuplement MW est représenté en trait plein, le peuplement HW est représenté à l'aide de tirets et le peuplement HB en pointillés.

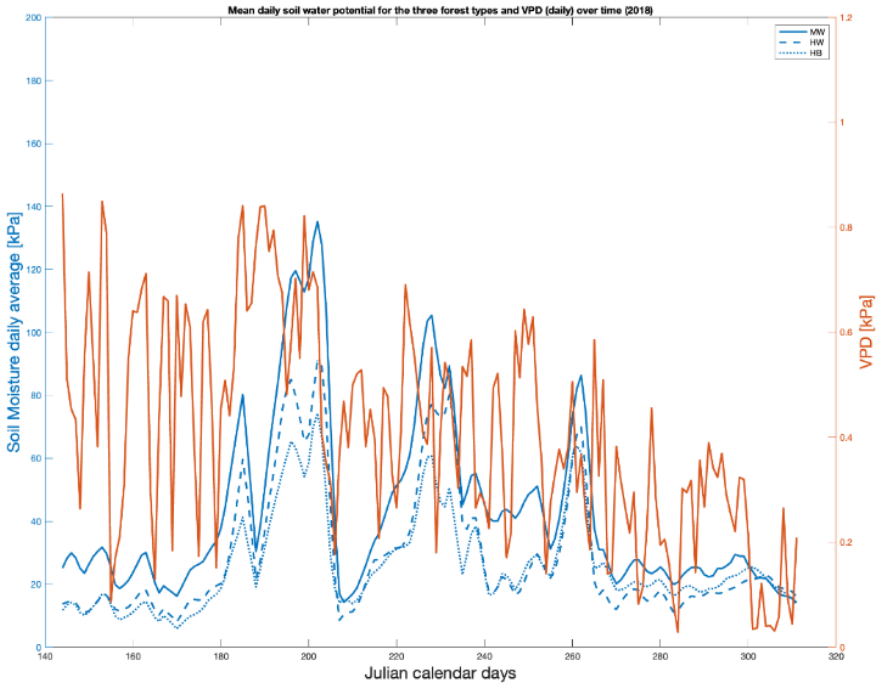


Figure 34 : Évolution du VPD et du potentiel hydrique du sol pour 2018.

- Potentiel hydrique journalier pour chacun des trois peuplements et évolution du VPD (2019)

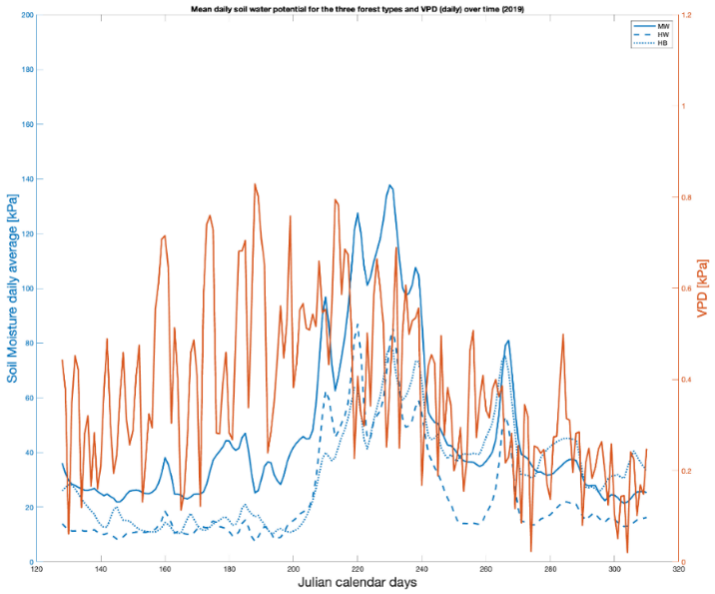


Figure 35 : Évolution du VPD et du potentiel hydrique du sol pour 2019.

- Potentiel hydrique journalier pour chacun des trois peuplements et évolution du VPD (2020)

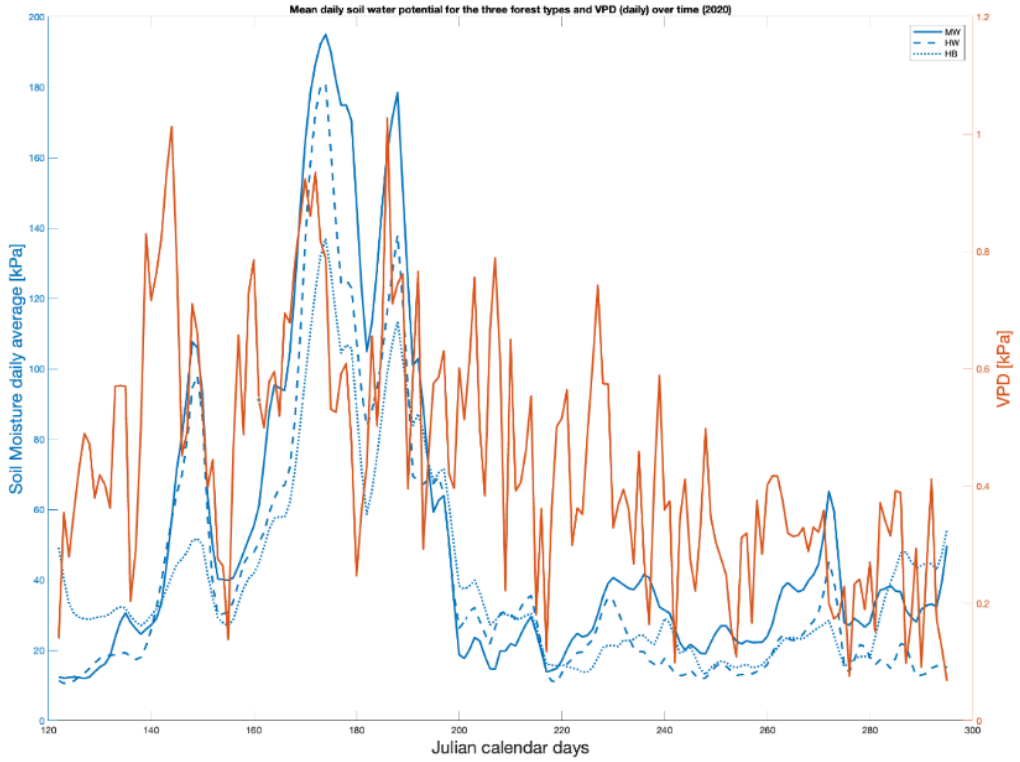


Figure 36 : Évolution du VPD et du potentiel hydrique du sol pour 2020.

Un déficit de pression de vapeur élevé implique des conditions sèches dans l'air et une augmentation des forces de transpiration sur la plante. Si celui-ci est trop élevé, il peut alors entraîner un flétrissement lorsque la prise d'eau des racines ne peut pas compenser la perte évaporative des feuilles.

Associé à un potentiel hydrique élevé, il peut être difficile pour les racines de suivre les pertes d'évaporation. À partir de 2018, nous constatons l'existence de périodes pour lesquelles le VPD et le potentiel sont simultanément élevés.

Nous pouvons souligner que les valeurs du potentiel hydrique observées en 2017 sont plus faibles que celles observées en 2018, 2019 et 2020. L'année 2017 correspond au démarrage de l'expérimentation, nous pouvons émettre l'hypothèse que les sondes de mesure n'étaient peut-être pas équilibrées et qu'il faut attendre 2018 pour obtenir des données plus fiables. Cet aspect est à considérer pour la suite des analyses.

Nous pouvons analyser les corrélations qui existent entre les variables d'intérêt.

- Représentation graphique des corrélations obtenues pour le peuplement MW (Figure 37)

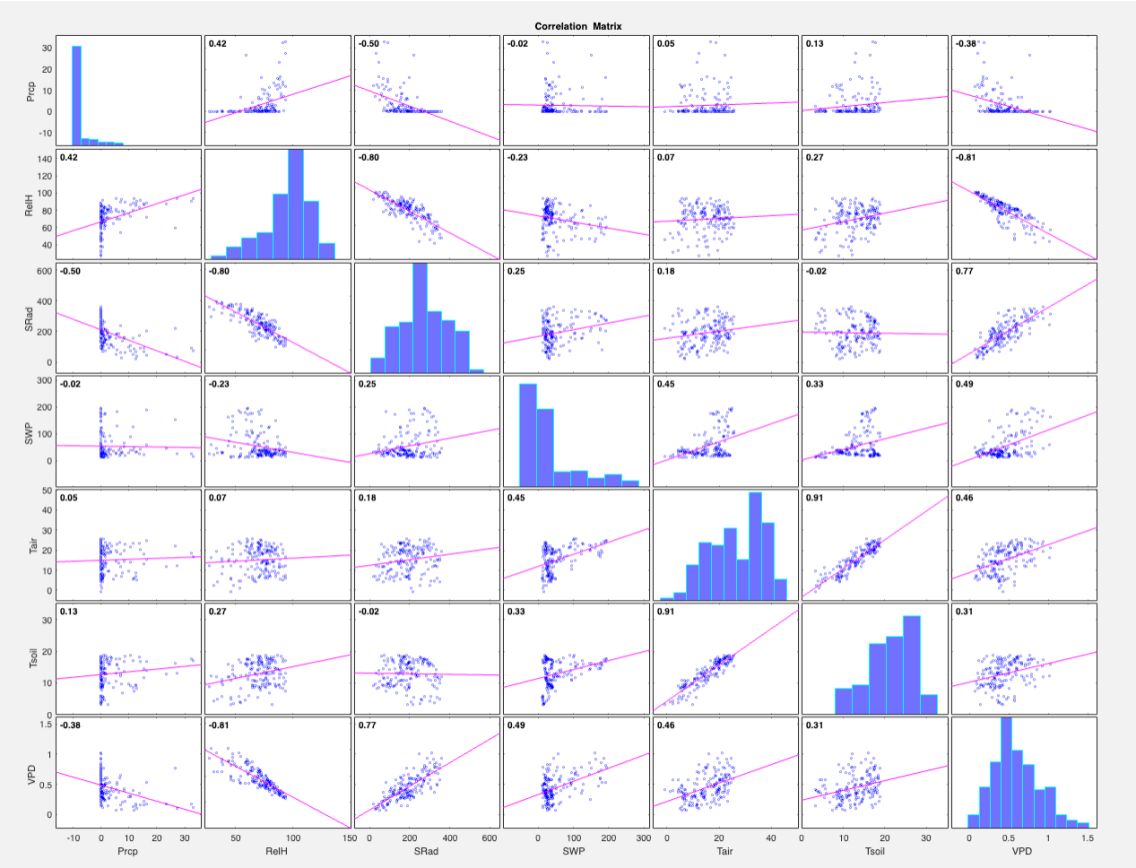


Figure 37 : Corrélation entre les variables d'intérêt pour le peuplement MW.

Nous pouvons détailler les coefficients de corrélation qui ont été obtenus (Tableau 3).

Tableau 3 : Coefficients de corrélation obtenus pour les variables d'intérêt.

	Prcp	RelH	SRad	Tair	VPD	Tsoil
Ψ_{MW}	-0,04	-0,28	0,32	0,53	**0,57**	0,42
Ψ_{HW}	-0,02	-0,23	0,25	0,45	**0,49**	0,33
Ψ_{HB}	0,00	-0,27	0,25	0,39	**0,49**	0,25

Pour les trois peuplements, nous observons la plus grande corrélation entre le potentiel hydrique du sol et le VPD, même si celle-ci ne demeure pas très élevée.

Pour les variables météo entre-elles, nous obtenons les corrélations représentées sur la Figure 38.

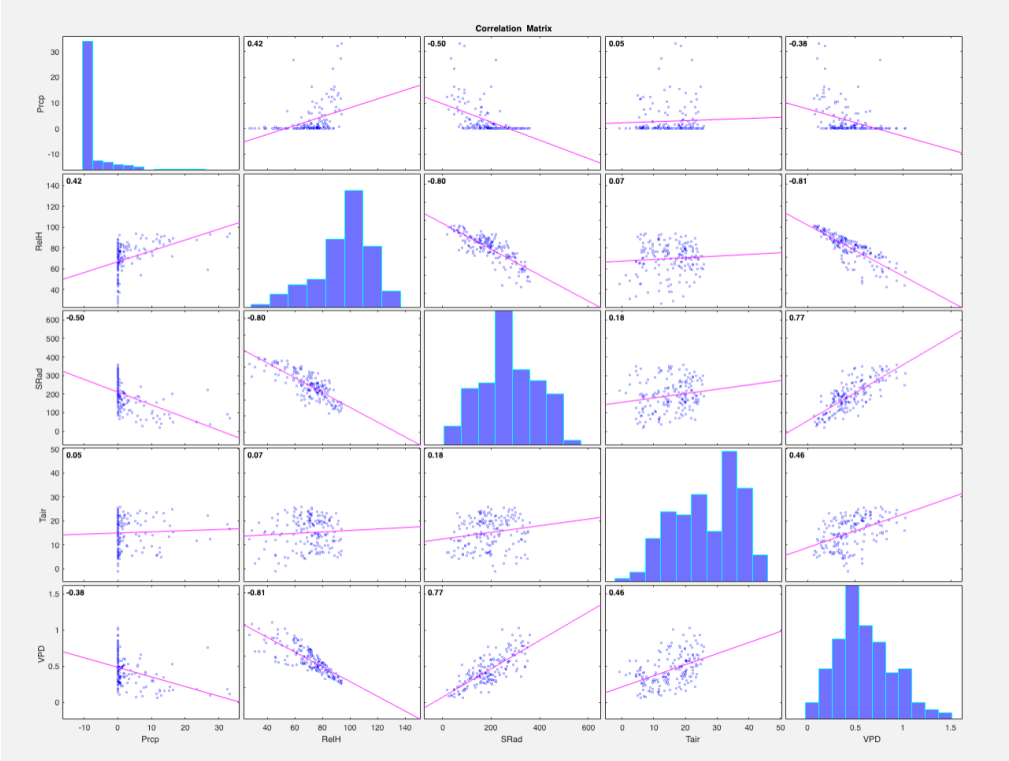


Figure 38 : Corrélations obtenues entre les variables météorologiques.

Les coefficients de corrélations obtenus pour les variables météorologiques sont résumés dans le Tableau 4.

Tableau 4 : Coefficients de corrélation obtenus pour les variables météorologiques.

	Prcp	RelH	SRad	Tair	VPD
Prcp					
RelH	0,42				
SRad	-0,50	**-0,80**			
Tair	0,05	0,07	0,18		
VPD	-0,38	**-0,81**	**0,77**	0,46	

Ce qu'il faut retenir :

Corrélation positive élevée entre les variables VPD et SRad

Corrélation négative élevée entre les variables VPD et RelH

Corrélation négative élevée entre SRad et RelH.

L'analyse des corrélations peut nous aider à sélectionner les variables prédictives, en privilégiant le déficit de pression de vapeur par rapport à l'irradiance solaire et l'humidité relative. Cette seconde phase de l'analyse a pour objectif de répondre à la première question de recherche qui a été formulée ainsi :

(QR1) *Quelles sont les variables climatiques considérées dans notre étude qui ont des conséquences importantes sur l'évolution temporelle du potentiel hydrique du sol mesuré à la Station de biologie des Laurentides entre 2017 et 2020 ?*

Il s'agit de se demander quels sont les paramètres qui peuvent potentiellement être des prédicteurs du potentiel hydrique du sol ?

Avant de répondre à cette question, nous allons créer une nouvelle variable catégorielle pour caractériser l'état du potentiel hydrique du sol : State_SM (*State of the soil moisture*).

De façon générale, cinq états sont définis en agriculture :

- Saturation,
- Capacité au champ (*field capacity*) : capacité de rétention maximale en eau du sol[8]
- Trigger d'irrigation (*irrigation trigger*)
- Point de flétrissement temporaire (*temporary wilting point*) : les racines ne peuvent plus vaincre les forces de rétention de l'eau
- Point de flétrissement permanent (*permanent wilting point*) : état irréversible, mort du végétal

Bien que les valeurs du potentiel hydrique du sol définissant chacun des états soient caractérisées pour des cultures spécifiques, celles-ci ne sont pas explicitement référencées pour les essences d'arbre rencontrées dans la forêt de feuillus ou la forêt boréale.

Nous allons donc définir quatre niveaux correspondant à quatre intervalles délimitant différentes valeurs du potentiel, de faible à très élevée (de S1 à S4). Ces valeurs notées SM1, SM2, SM3 et SM4 sont prises en nous reposant sur les valeurs proposées par Schock (2013)[9]. Ces valeurs représentent davantage des

[8] Elle correspond plus précisément à la quantité d'eau retenue, après 48 heures d'égouttement de l'eau libre vers la nappe phréatique, par un sol préalablement gorgé d'eau (par des pluies ou un arrosage intensif).

[9] Schock, C.C., Wang, F.X., Flock, R., Feibert, E., Schock, C.A., Pereira, A. (2013), Irrigation monitoring using soil water tension, Sustainable agriculture techniques EM 8900 – revised March 2013, Extension Service, Oregon State University

ordres de grandeur, que des valeurs seuils spécifiques aux peuplements d'arbres considérés dans notre étude.

Faible tension : [SM1; SM2]

Tension moyenne :] SM2 ; SM3]

Tension élevée :] SM3 ; SM4]

Tension très élevée : > SM4

En considérant chaque peuplement, nous aurions ainsi les intervalles définis dans le Tableau 5.

Tableau 5 : Intervalles définissant l'état du potentiel hydrique du sol.

State [kPa]	MixedWood (MW)	HardWood (HW)	HardBeech (HB)
S1	[$SM1_{mw}$; $SM2_{mw}$]	[$SM1_{hw}$; $SM2_{hw}$]	[$SM1_{hb}$; $SM2_{hb}$]
S2	]$SM2_{mw}$; $SM3_{mw}$]	]$SM2_{hw}$; $SM3_{hw}$]	]SM_{hb} ; $SM3_{hb}$]
S3	]$SM3_{mw}$; $SM4_{mw}$]	]$SM3_{hw}$; $SM4_{hw}$]	]$SM3_{hb}$; $SM4_{hb}$]
S4	> $SM4_{mw}$	> $SM4_{hw}$	> $SM4_{hb}$

Dans une première approche, nous allons considérer les mêmes valeurs pour les trois peuplements (Tableau 6).

Il est important de souligner qu'à cette étape, *les valeurs de potentiel proposées ne correspondent pas à un stress hydrique bien identifié, mais permettent uniquement de classifier les valeurs selon leur intensité.* L'utilisation d'une nouvelle variable, catégorielle, permettra le développement de méthodes de classification, telles que celles s'appuyant sur des arbres de décision.

Tableau 6 : Seuils du potentiel hydrique du sol définis dans une première approche.

State [kPa]	MW – HW - HB
S1 (Low)	[0; 40]
S2 (Medium)	]40 ; 80]
S3 (High)	]80; 120]
S4 (Very high)	> 120

Pour appuyer notre choix, nous avons évalué la densité du potentiel hydrique, mesuré en 2020, pour les trois peuplements (Figure 39).

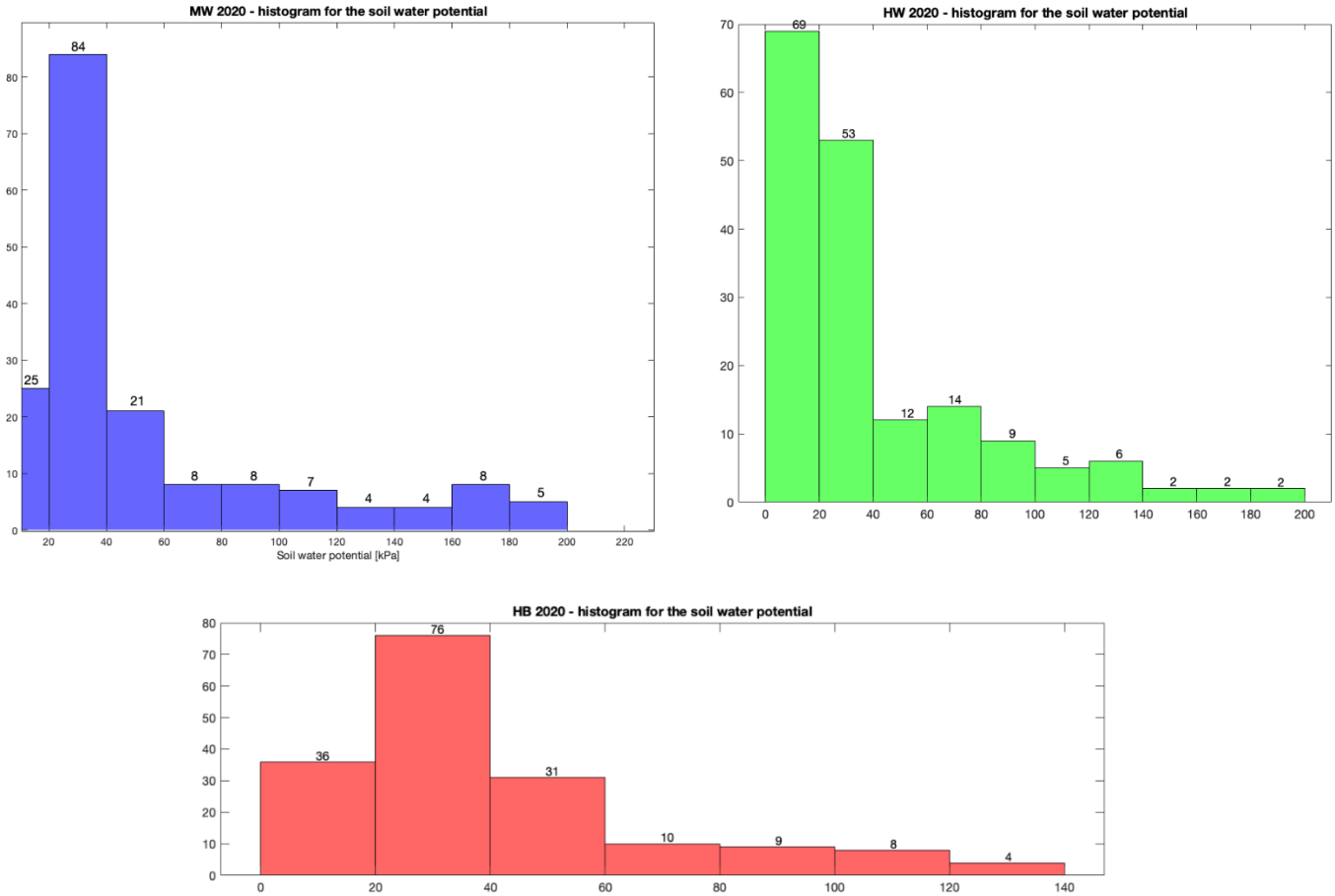
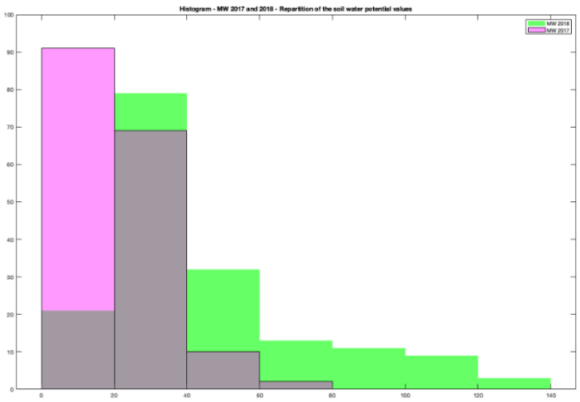


Figure 39 : Densité du potentiel hydrique du sol pour les peuplements MW (en bleu), HW (en vert) et HB (en rose) pour l'année 2020.

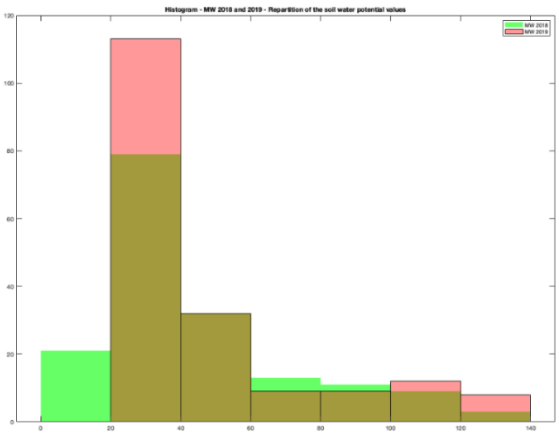
En 2020, le peuplement MW avait davantage de valeurs du potentiel supérieures à 140 kPa que les deux autres peuplements. Une grande majorité des valeurs se situent entre 0 et 40 kPa.

Nous pouvons également comparer l'évolution de la densité au cours des quatre années étudiées, pour le peuplement MW.



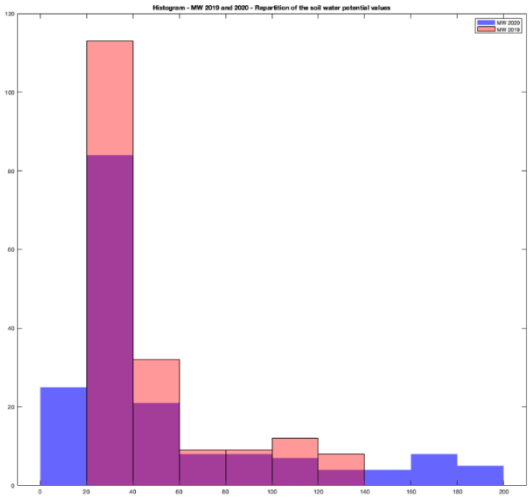
En ROSE : 2017

En VERT : 2018



En VERT : 2018

En ROSE : 2019



En ROSE CLAIR : 2019

En BLEU : 2020

Nous remarquons qu'entre 2017 et 2020 la répartition des valeurs de potentiel s'est décalée vers les valeurs plus élevées. En effet, la densité du potentiel a augmenté vers les valeurs supérieures à 80kPa, entre 2017 et 2020.

Classification à l'aide d'arbres de décision

Objectif : définir les variables qui permettent une prédiction plus précise de l'état du potentiel hydrique.

De façon générale[10], l'objectif principal des arbres de décision est de prédire la valeur d'une variable – la variable cible – connaissant différentes variables décrites par un ensemble de données. Cette technique de forage de données repose sur un apprentissage dit supervisé. En effet, il faut connaître la catégorie en sortie de la variable cible, qui est dans notre cas, l'état du potentiel hydrique SM_State.

Après avoir construit l'arbre, en suivant une certaine approche algorithmique, un ensemble de règles de décision est ainsi obtenu. Nous le verrons, il est possible de représenter graphiquement l'arbre final, afin de matérialiser ces règles qui aideront à la décision en hiérarchisant l'information, de manière récursive. La représentation adoptée sera de type DNF (*disjunctive normal form*), i.e., utilisant des formes disjonctives de conjonctions ((A et B) ou (C et D)).

L'arbre possède ainsi une racine (en haut), des branches, des nœuds internes – correspondant aux attributs – des nœuds terminaux (les feuilles) qui correspondent à une classe.

Pour départager les différents choix, différents critères de division peuvent être utilisés. Dans le cadre de notre projet, il s'agira de l'entropie, qui permet de mesurer l'incertitude d'une variable, associée à une distribution de probabilité $P=(p_1, \ldots, p_n)$. Elle se définit ainsi : $E(P) = -\sum_i p_i \ log_2(p_i)$

Il est intéressant de souligner que l'approche choisie est une approche descendante (*top-down*), dans laquelle on ne regarde qu'une hypothèse à la fois, sans retour en arrière. On va choisir à chaque itération la meilleure alternative, en privilégiant l'hypothèse la plus simple (d'après le principe du rasoir d'Occam). L'algorithme utilisé par défaut dans Matlab est l'algorithme CART (Classification And Regression Trees). Nous le testerons, ainsi que deux autres variantes, qui intègrent de façon plus spécifique les valeurs prédictives d'entrée. En effet, l'approche *curvature* et *interaction-curvature* sont également accessibles.

---

[10] Référence pour les arbres de decision : Cours de la Pre Mezghani, TÉLUQ, module 5 INF6409 (2015) et Cours du Pr Lounis, UQAM, DIC9380 (2020).

Dans le cas de la première, le prédicteur qui permet la meilleure scission (embranchement) est celui qui minimise la valeur de p du test statistique de courbure entre chaque prédicteur considéré X_i et la réponse y. L'hypothèse nulle testée est la suivante : les 2 variables testées (X_i et y) ne sont pas associées.

L'autre approche s'appuie sur le test d'interaction qui vérifie l'hypothèse nulle suivante : il n'y a pas d'interaction entre une paire de prédicteurs (X_1 , X_2) et la réponse y.

Cette approche permet de mettre en évidence des interactions importantes entre les prédicteurs.

Les deux approches, ainsi que l'approche par défaut, qui ne sélectionne pas les prédicteurs au préalable, seront testées.

Enfin, il est important dans ce type d'approche de se questionner sur la profondeur de l'arbre pour éviter les problèmes d'*overfitting*. En effet, si l'arbre est trop profond, il risque de capturer des aberrations et donc sera moins bon à prédire de nouveaux cas. Certaines méthodes (implémentées dans Matlab et utilisées), comme le post-élagage (*pruning*), peuvent être appliquées pour tester différentes tailles d'arbres et ainsi, éviter ce type de problèmes. Nous opterons pour un élagage visant à obtenir un arbre de décision moins profond.

Six variables ont été sélectionnées pour tenter de classer les valeurs du potentiel (Tableau 7).

Quatre variables proviennent des données météorologiques générées grâce au logiciel BioSIM, une variable a été calculée (VPD) et une variable provient des mesures expérimentales (mean_T).

Tableau 7 : Variables sélectionnées pour la classification.

Variable	Notation	Unité
Température moyenne du sol (mesurée)	mean_T	[°C]
Température de l'air (moyennée sur une journée)	Tair	[°C]
Précipitations	Prcp	[mm]
Humidité relative	RelH	[%]
Irradiance solaire	SRad	[W.m-2]
Déficit de pression de vapeur	VPD	[kPa]

Approche mise en place sur les données de 2020 pour MW, HW, HB

Objectif 1 : prédire les données de 2019 à partir du modèle entraîné avec les données de 2020.

Nous testerons également la prédiction des données de 2020 à partir de modèles entrainés avec les données de 2019.

Démarche

(1) Modèle CART par défaut[11]

(2) Modèle avec choix des prédicteurs de type *curvature*

   a. *Sans élagage*

   b. *Avec élagage*

   c. *Avec un nombre de branches maximal (fixé à 10)*

(3) Modèle avec choix des prédicteurs de type *interaction-curvature*

   a. *Sans élagage*

   b. *Avec élagage*

   c. *Avec un nombre de branches maximal (fixé à 10)*

Le Tableau 8 décrit les différents modèles entrainés avec les données propres à chaque peuplement, en 2020.

Tableau 8 : Description des différents modèles testés.

Nom donné au modèle	Description
M1_0	Algorithme CART (défaut)
M1_1	Algorithme CART avec MaxSplit = 10
M2_0	Sélection des prédicteurs utilisant le test de courbure, sans élagage
M2_1	Sélection des prédicteurs utilisant le test de courbure, avec élagage
M2_2	Sélection des prédicteurs utilisant le test de courbure, avec MaxSplit = 10
M3_0	Sélection des prédicteurs utilisant le test d'interaction, sans élagage
M3_1	Sélection des prédicteurs utilisant le test d'interaction, avec élagage
M3_2	Sélection des prédicteurs utilisant le test d'interaction, avec MaxSplit = 10

Le Tableau 9 synthétise les pourcentages d'exactitude dans la prédiction des données de 2019, pour chaque peuplement.

[11] Se référer à la fonction fitctree de Matlab : https://www.mathworks.com/help/stats/fitctree.html

Les arbres ont été entrainés avec les données de 2020, et les données de 2019 ont été prédites.

Tableau 9 : Pourcentages d'exactitude [%] obtenus pour la prédiction des données de 2019, pour les trois peuplements.

Modèle \ Peuplement	MW	HW	HB
M1_0	**57**	67	43
M1_1	55	67	39
M2_0	50	66	46
M2_1	50	66	46
M2_2	52	67	50
M3_0	51	72	52
M3_1	51	72	52
M3_2	52	**73**	**60**

M2_0 et M2_1 s'avèrent être des modèles identiques (l'option avec ou sans élagage n'a pas eu d'impact sur la prédiction). Il en est de même pour M3_0 et M3_1.

Pour les peuplements HW et HB, le modèle M3_2 (sélection des prédicteurs par interaction et 10 embranchements au maximum) permet d'obtenir une meilleure prédiction des données de 2019.

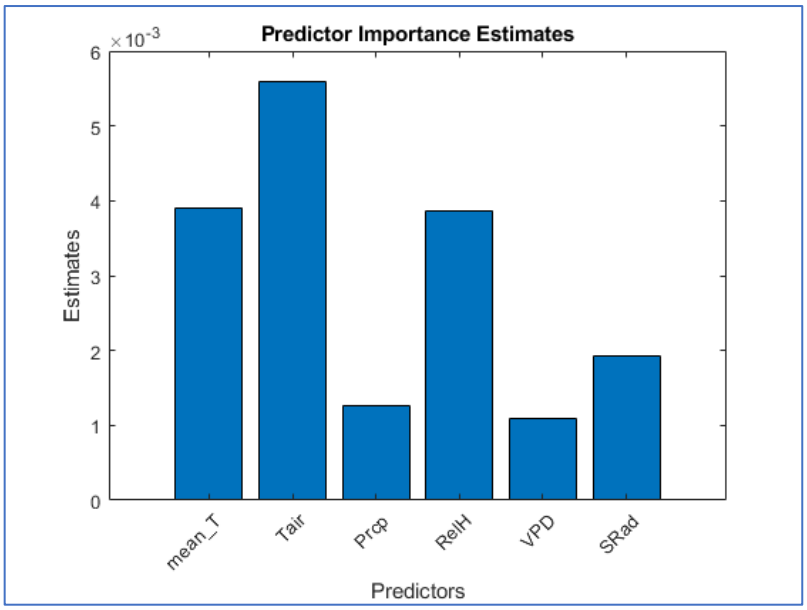
Pour le peuplement MW, le modèle CART par défaut permet d'obtenir le pourcentage de prédictions exactes (*accuracy*) le plus élevé.

Néanmoins, les pourcentages obtenus restent faibles si nous considérons qu'une prédiction aléatoire donnerait un pourcentage d'exactitude de 25% (pour les 4 classes définies : ¼ x 100).

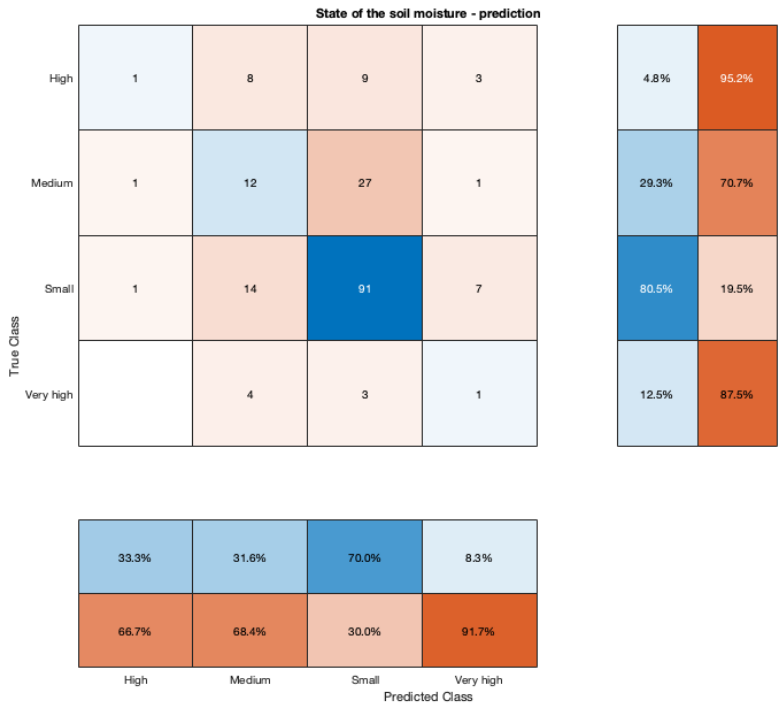
Pour les modèles M1_0 et M3_2, nous avons tracé l'importance des variables prédictives (voir l'Annexe V pour plus de précisions).

Résultats des tests effectués, pour les modèles donnant le pourcentage d'exactitude le plus élevé

Modèle M1_0 pour MW (entraînement 2019, prédiction 2020)

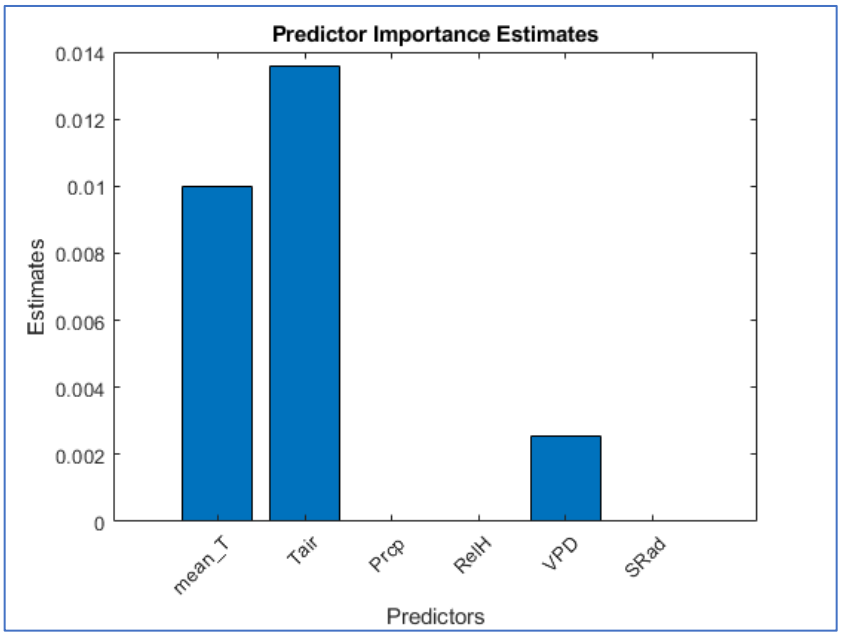


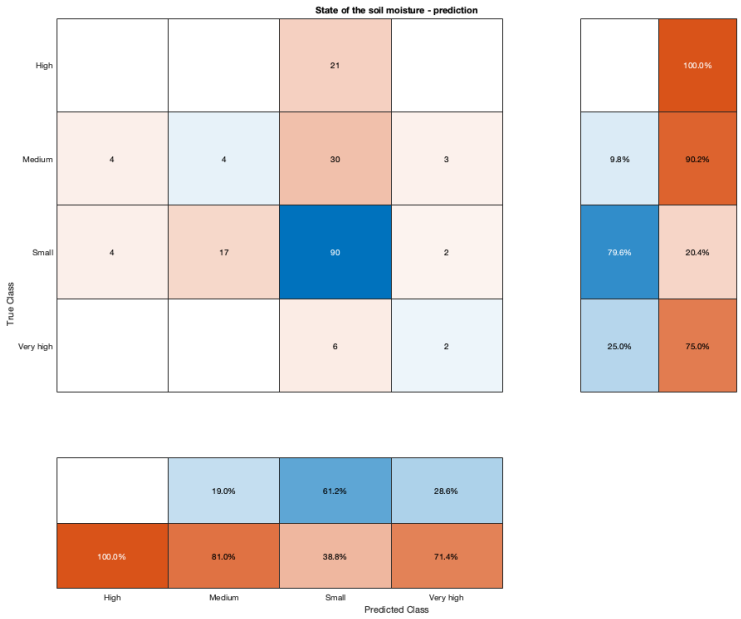
Pour le peuplement MW, avec l'algorithme CART simple, la température de l'air ressort comme variable au plus fort pouvoir prédicteur du potentiel hydrique. La température moyenne du sol et l'humidité relative semblent également participer à la prédiction de ce dernier. L'irradiance solaire, les précipitations et le VPD ressortent également, mais plus faiblement.



La matrice de confusion qui résulte de ce premier modèle nous indique que 91 états *small* du potentiel sur 113 (80,5%) ont été correctement prédits comme small. 12 états *medium* sur 41 (29,3%) ont été correctement prédits comme *medium.* Pour les états high 4,8% ont été prédits correctement.

Modèle M3_2 pour MW - interaction par paires de prédicteurs

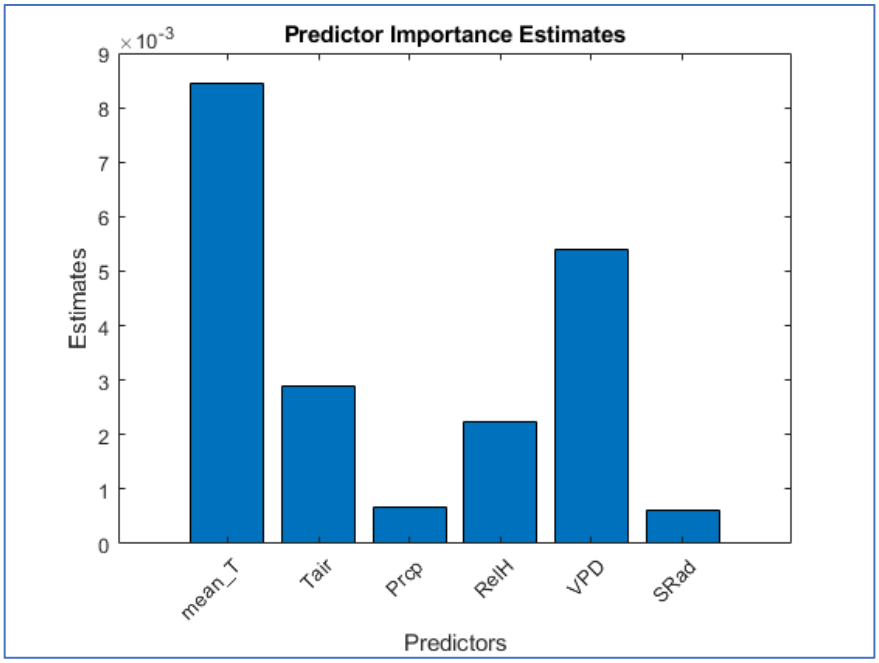




Avec la sélection au préalable des prédicteurs, les résultats sont différents : La température de l'air et la température moyenne du sol ressortent également, plus fortement, et seul le VPD est conservé.

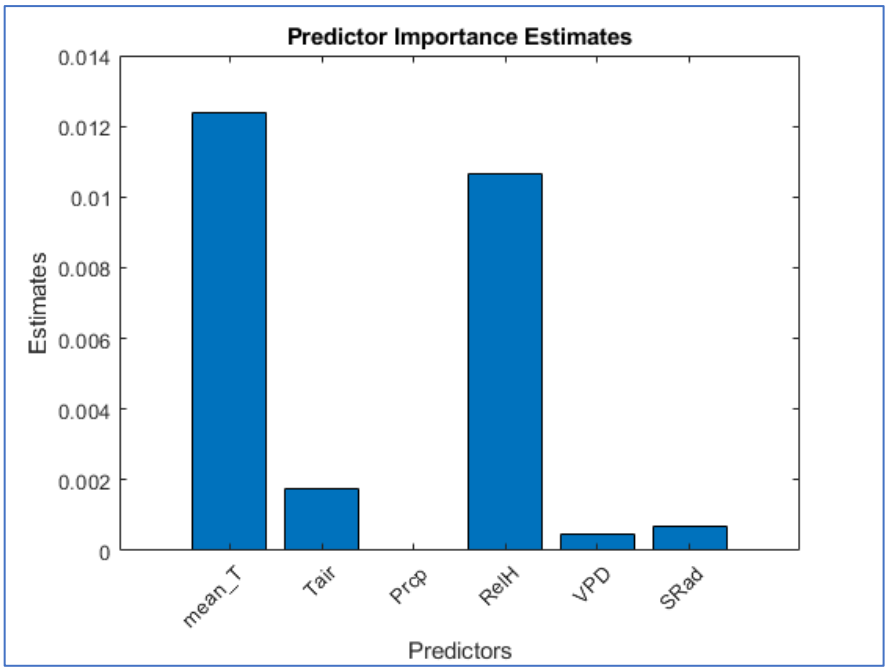
Une explication pourrait être la forte corrélation entre le VPD et SRad et entre le VPD et RelH (voir résultats obtenus pour l'étude des corrélations). En effet, lors de la sélection par pairs, seul le VPD est alors conservé.

Modèle M1_0 pour HW



Pour le peuplement HW, la température du sol semble être plus prédictive du potentiel que la température de l'air, contrairement à ce qui a été mis en évidence pour MW. Le VPD surpassant la température de l'air.

Modèle M3_2 pour HW - interaction par paire de prédicteurs

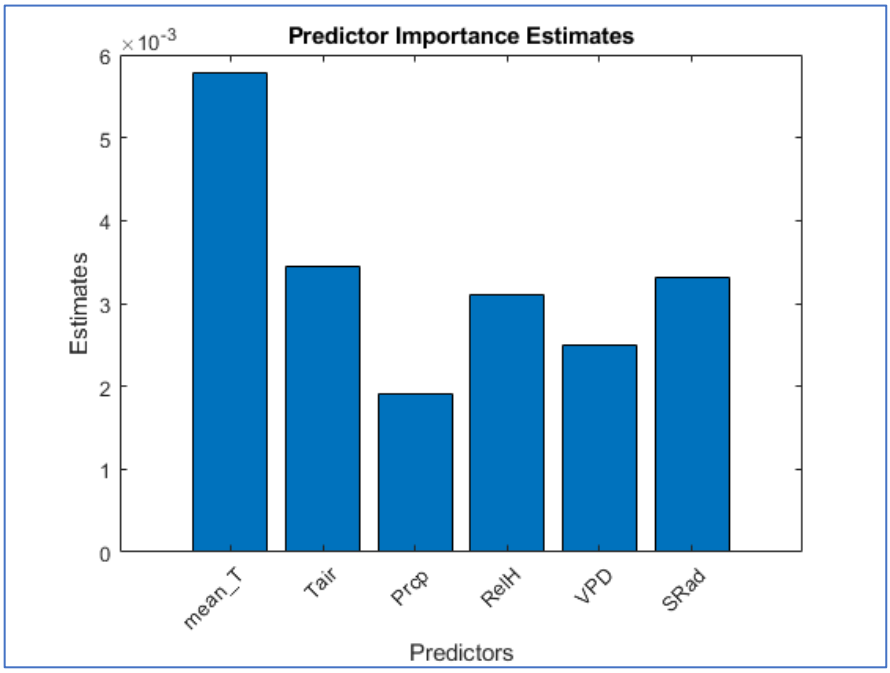


Avec le modèle M3_2, c'est l'humidité relative qui semble ressortir, pour HW.

La température moyenne du sol ayant toujours un fort pouvoir de prédiction.

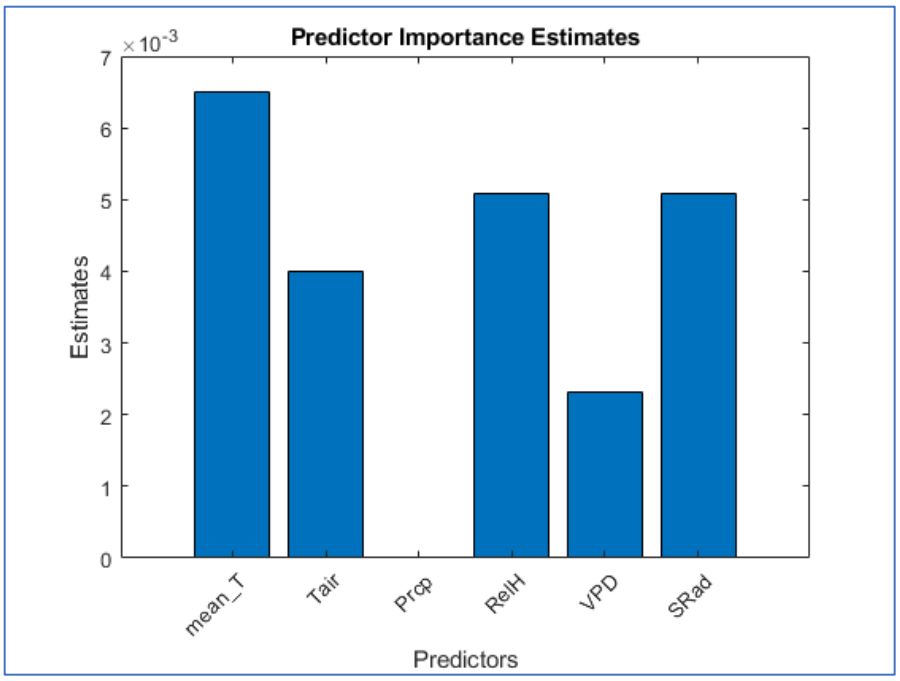
Les précipitations ne ressortent plus.

Modèle M1_0 pour HB



Pour le peuplement HB, la température du sol a un pouvoir de prédiction supérieur, comme pour HW. Les autres variables ressortent également de manière comparable.

Modèle M3_2 pour HB – interaction par paire de prédicteurs



Avec la sélection au préalable des variables prédictives, la température moyenne du sol reste la variable au plus fort pouvoir de prédiction du potentiel hydrique. RelH et SRad semblent surpasser le VPD, et Tair.

Les précipitations ne semblent pas participer à la prédiction du potentiel dans ce cas de figure.

Est-ce que les variables prédictives du potentiel hydrique changent durant les périodes de sécheresse-flash ?

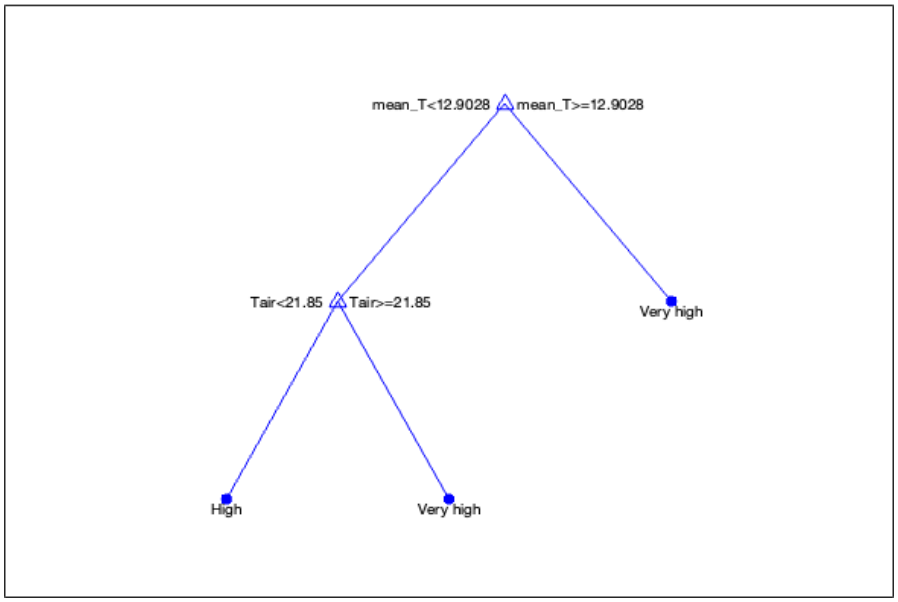
Pour répondre à cette question, nous avons considéré dans un second temps les données correspondantes aux trois périodes de sécheresse-flash de 2020. Ces périodes de sécheresse-flash ont été définies dans le mémoire (chapitre 3). Elles correspondent à des périodes de 8 jours au moins de déficit hydrique.

Le modèle choisi pour la classification est l'algorithme CART par défaut, sans sélection au préalable des prédicteurs.

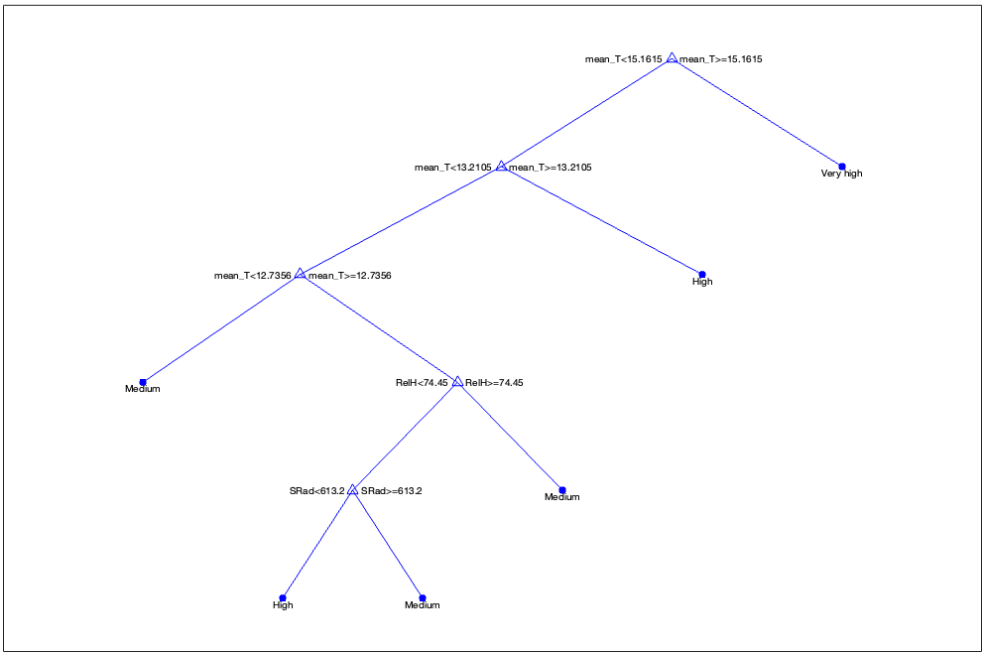
Dans les trois cas, la température moyenne du sol ressort comme variable prédictive la plus importante.

Analysons plus en détail **la sécheresse-flash du mois de juin 2020** et comparons les trois arbres de décision obtenus, pour chaque peuplement considéré.

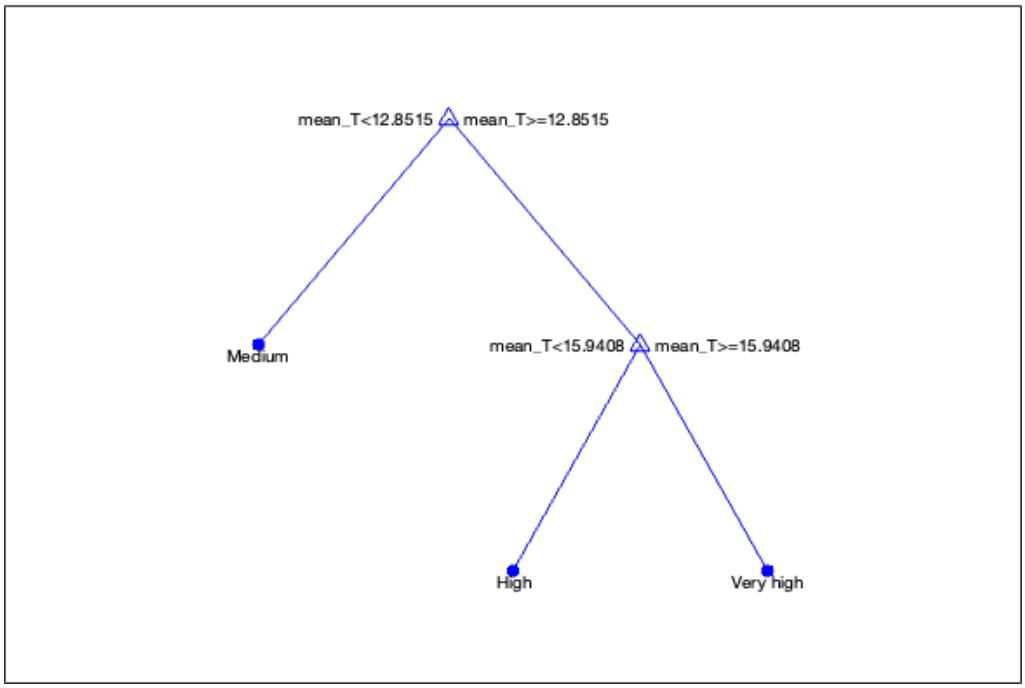
- Arbre de décision obtenu pour *le peuplement MW*



- Arbre de décision obtenu pour *le peuplement HW*



- Arbre de décision obtenu pour *le peuplement HB*



Ce qui ressort de la comparaison des trois arbres, spécifiques à chaque peuplement, est que *le critère d'embranchement sur la variable température moyenne du sol pour basculer vers une valeur très élevée du potentiel hydrique (classe Very high) diffère selon les peuplements.*

Pour le peuplement HB, pour lequel le hêtre est majoritaire, le potentiel hydrique bascule vers une valeur considérée arbitrairement comme *très élevée* pour une température moyenne du sol supérieure à 15.9°C.

Cette température est de 15.1°C pour le peuplement HW et de 12.9°C pour le peuplement MW.

Ainsi, le trigger de basculement vers un potentiel très élevé, dans un contexte de sécheresse-flash, repose pour le peuplement HB sur une température du sol plus élevée. Pour le peuplement MW, la température du sol est plus basse, suggérant alors une moins bonne résistance face à l'élévation de température du sol.

Trois degrés Celsius séparent les triggers de basculement. Dans un contexte de changement climatique, la meilleure résistance des hêtres aux élévations de température du sol peut être un critère d'adaptation positif pour ce peuplement.

Une limite cependant à cette analyse : nous avons considéré des valeurs seuils de potentiel identiques pour les trois peuplements.

Est-ce qu'une valeur du potentiel hydrique du sol égale à 120 kPa veut dire la même chose pour chaque peuplement ?

L'avantage des méthodes de classification utilisant les arbres de décision est l'obtention de règle de type *Si…alors…,* qui permettent d'expliquer la dépendance entre les variables. L'approche a un pouvoir explicatif plus grand que des méthodes d'apprentissage machine exploitant les réseaux de neurones.

Il peut donc s'avérer plus simple de se reposer sur des arbres de décision qui fournissent des résultats plus clairs permettant ainsi de faire plus facilement des liens avec les phénomènes qui ont lieu à l'échelle des arbres.

Néanmoins, nous avons mis en évidence le manque d'exactitude des arbres de décision dans la prédiction des valeurs du potentiel hydrique, et en particulier quand celles-ci sont élevées.

Nous pouvons faire les constats suivants :

(1) Le pourcentage d'exactitude pour la prédiction du potentiel hydrique du sol mesuré en 2019, se situe entre 39% (HB) et 73% (HW), pour tous les tests effectués.

Le modèle M3 (sélection des variables prédictives avec l'approche d'interaction par pair) avec un nombre de branches fixé à 10, permet d'obtenir les meilleures prédictions pour les peuplements HW et HB.

Pour le peuplement MW, nous obtenons 57% d'exactitude avec l'approche CART; c'est-à-dire que sur toutes les valeurs de potentiel hydrique considérées, pour le peuplement MW, plus de la moitié est correctement prédite. Mais ce pourcentage reste faible en comparaison avec les 25% d'exactitude que nous obtiendrions avec un classement purement aléatoire des quatre classes considérées.

Ce que nous remarquons, c'est **l'incapacité des modèles - et cela est vrai pour tous les modèles testés - à prédire correctement les valeurs élevées du potentiel (catégories *High* et *Very High*) se situant au-dessus de 80kPa.**

Nous sommes devant une **situation débalancée** (*imbalanced conditions*), pour laquelle les classes *High* et *Very High* sont sous-représentées par rapport aux deux autres classes *Small* et *Medium*. Cette situation engendre des limites à l'utilisation de méthodes de classification telles que celles reposant sur l'algorithme CART.

Une approche de type SMOTE (*synthetic minority over-sampling technique*)[12] ou une de ses améliorations[13] pourraient être testées pour essayer d'améliorer l'exactitude du modèle M3.

(2) Pour écarter toute corrélation entre la température du sol et le potentiel qui pourrait avoir été introduite dans la première phase d'analyse, lors de l'ajustant de la valeur mesurée du potentiel à la température du sol, nous avons également déterminé et testé un modèle utilisant comme données d'entrée les mesures non-ajustées du potentiel hydrique. Les résultats obtenus ne différaient pas.

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over– sampling technique. J. *Artif. Intell. Res*. *16*, 321–357 (2002)

[13] Fernández Alberto, García Salvador, Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets. Springer.

Il a été montré, dans d'autres contextes expérimentaux, que le potentiel hydrique du sol était très sensible à la température de ce dernier (Seyfried, 2001)[14], et cette sensibilité pouvait être douze fois plus grande lorsque les sols étaient saturés, en comparaison avec des sols plus secs (Evett, 2006)[15].

Les modèles testés favorisant les classes majoritaires Small et Medium qui représentent de plus faibles tensions du potentiel - caractéristiques de sols humides – nous pouvons nous demander si la sur-sensibilité du potentiel vis-à-vis de la température du sol ne se matérialise pas par une sur-importance de la capacité prédictive de la variable mean_T (température du sol) ?

(3) Nous pourrions ainsi tenter d'améliorer les approches de classification utilisées dans le but de mettre en évidence les variables qui permettraient la meilleure prédiction du potentiel hydrique du sol. Cependant, n'y a-t-il pas une limite intrinsèquement liée au choix de la méthode ?

Les arbres de décision proposent une approche statique, qui ne prend pas en compte les liens temporels entre les variables.

En effet, l'impact d'une élévation de température à un temps t_1 se fait ressentir sur le potentiel à un temps différent t_2. Il conviendra alors de définir le délai entre l'augmentation ou la diminution de chaque variable exogène considérée et la variation observée du potentiel hydrique du sol.

En d'autres termes, il s'agit tout d'abord de considérer les données comme des séries chronologiques (temporelles) et *de déterminer, non pas la corrélation entre les variables, mais quelle série cause une variation de la série temporelle du potentiel hydrique, et combien de points temporels sont nécessaires pour permettre cette prédiction.*

Cette causalité a été décrite par Granger (1969) qui lui a donné son nom[16]. Elle peut s'exprimer ainsi : Une série temporelle $X_{1,t}$ causerait – au sens de Granger – la série $X_{2,t}$ lorsque la connaissance du passé de $X_{1,t}$ entraine une prévision de $X_{2,t}$ distincte (et meilleure) de celle fondée uniquement sur le passé de $X_{2,t}$. Pour

[14] Seyfried, M. and Murdock, M. (2001), Response of a New Soil Water Sensor to Variable Soil, Water Content, and Temperature. *Soil Sci. Soc. Am. J.*, *65*: 28-34.

[15] Evett, S.R., Tolk, J.A. and Howell, T.A. (2006), Soil Profile Water Content Determination: Sensor Accuracy, Axial Response, Calibration, Temperature Dependence, and Precision. *Vadose Zone Journal*, *5*: 894-907

[16] Granger, C. W. J. (1969) Investigating Causal Relations by Econometric Models and Cross-Spectral Methods, *Econometrica*, 37, pp. 424–459.

s'assurer de cela, l'erreur quadratique moyenne de prédiction de $X_{2,t}$ en considérant $X_{1,t}$ est comparée à celle obtenue sans celle-ci. Si cette erreur est significativement inférieure quand $X_{1,t}$ est considérée, alors $X_{1,t}$ cause $X_{2,t}$ au sens de Granger.

En pratique, un test statistique est réalisé sur ces erreurs, avec pour hypothèse nulle : X_1 ne cause pas X_2, au sens de Granger. Si l'hypothèse nulle est rejetée, alors il est possible de conclure à une causalité de Granger entre X_1 et X_2.

Analyse de la causalité de Granger entre les variables prédictives et le potentiel hydrique du sol

Le seuil de significativité est fixé à $p<0.05$.

Le postulat fait pour effectuer le test de la causalité de Granger est le suivant :

On suppose qu'il existe p, un entier naturel non nul, tel que X_1 et X_2 suivent un modèle de type VAR(p) - Vector Auto Regressive, d'ordre p.

Test1 : *Est-ce que la température moyenne du sol mean_T cause le potentiel hydrique Ψ, au sens de Granger ?*

Hypothèse nulle H_0: La température ne cause pas le potentiel, au sens de Granger.

$H_1 = 0$ (l'hypothèse nulle n'est pas rejetée.)

La dernière valeur de la température du sol considérée au temps t n'aide donc pas à la prédiction du potentiel, à ce même temps t.

Nous testons la causalité de Granger, **en introduisant alors un délai dans la variable exogène.**

Ainsi, pour un délai de 4 jours (lag=4), l'hypothèse nulle peut être rejetée, ce qui nous permet de conclure que les 4 dernières valeurs de la température du sol aident à prédire le potentiel hydrique du sol.

Le test est réalisé pour les trois peuplements MW, HW et HB.

Test2 : *Est-ce que la température de l'air cause le potentiel hydrique, au sens de Granger ?*

$H_1 = 0$

Il n'y a pas de causalité de Granger entre Tair et le potentiel.

Le test est réalisé avec un délai compris entre 2 et 10 jours. L'hypothèse nulle n'est jamais rejetée.

Les tests réalisés pour le peuplement MW (données de 2020) sont résumés dans le Tableau 10.

Tableau 10 : Résultats obtenus pour l'analyse de la causalité de Granger.

Variable testée (X_1)	Lag (jour)	H0 R pour Rejetée NR pour Non Rejetée	Conclusion X_1 cause-t-il le potentiel au sens de Granger ?
mean_T	1	NR	Pas de causalité de Granger entre la température du sol et le potentiel Ψ
mean_T	4	R	La causalité apparait à un délai de 4 jours.
Tair	1	NR	Pas de causalité de Granger entre Tair et Ψ
Tair	2 à 10	NR	La non-causalité est maintenue sur 10 jours
RelH	1	R	Causalité de Granger entre RelH et Ψ
RelH	2	NR	La causalité disparaît au lag 2
Prcp	1	R	Causalité de Granger entre Prcp et Ψ
Prcp	2	NR	La causalité disparaît au lag 2
VPD	1	R	Causalité de Granger entre VPD et Ψ
VPD	2	NR	La causalité disparaît au lag 2
SRad	1	R	Causalité entre SRad et Ψ
SRad	2	R	La causalité se maintient au lag 2
SRad	3	NR	La causalité disparaît au lag 3

Pour le peuplement *HB, quel que soit le délai choisi, la température du sol ne cause pas au sens de Granger le potentiel hydrique.*

Pour les peuplements HW et HB, la causalité entre le VDP et le potentiel disparaît au lag 3.

Pour le peuplement HB, la causalité entre SRad et le potentiel disparaît comme pour le peuplement MW, au délai 3. En revanche, pour le peuplement HW, elle disparaît au délai 5.

Nous avons également réalisé les tests sur des séries différenciées, nous assurant alors de leur stationnarité. Cela ne semble influer que sur le délai à considérer.

Que faut-il retenir de cette analyse de la causalité de Granger et que pouvons-nous conclure ?

(1) La température de l'air, considérée seule, ne semble pas causer le potentiel hydrique du sol, au sens de Granger. Cette variable ne permet pas une meilleure prédiction du potentiel.

Ce résultat semble contredire le résultat obtenu lors de l'analyse avec les arbres de décision qui accordait une grande importance à la température de l'air comme variable prédictive.

(2) Il faut considérer les quatre dernières valeurs temporelles de la température du sol pour qu'une causalité avec le potentiel apparaisse. En d'autres termes, les quatre dernières valeurs journalières de la température du sol permettent une meilleure prédiction du potentiel.

Pour le peuplement HB, la causalité entre la température du sol et le potentiel n'est pas concluante.

(3) Pour les autres variables, il semblerait que l'interaction avec le potentiel se fasse sans délai. En effet, la causalité est mise en évidence sans décalage, et disparaît très rapidement, suggérant une causalité immédiate entre l'humidité relative, le VPD, la radiation solaire et le potentiel hydrique (Ψ) du sol.

Nous pouvons nous demander si les relations de causalité changent lorsque nous considérons les périodes de sécheresses-flash (dry-down).

Le peuplement MW a été considéré dans une première approche.

Si nous isolons les trois périodes de mai, juin et septembre 2020, les relations de causalité changent.

En particulier, il apparait une causalité au sens de Granger entre la température de l'air et le potentiel, en mai et en juin, mais pas en septembre, période pour laquelle les températures sont plus basses.

La température du sol cause également le potentiel, pour les 3 périodes, mais si nous considérons davantage de points temporels (8 jours avant et après la période de sécheresse), ces deux causalités disparaissent, comme remarqué quand toute la série chronologique était étudiée.

À ce stade, nous ne pouvons pas conclure sur le rôle des périodes de sécheresse.

En effet, ces changements de causalité observés peuvent être dus à l'approche utilisée, et en particulier à l'hypothèse d'existence d'un modèle de type VAR(p) pour modéliser les séries temporelles.

Nous avons réalisé plusieurs tests afin d'évaluer la fiabilité d'une telle hypothèse.

Pour le peuplement HB, l'ajustement de type VAR(1) avec la température du sol ne semble pas optimal, car le critère d'information de Akaike élevé, supérieur à 1000, signifiant que la perte d'information lors de la modélisation est importante. Ce qui pourrait expliquer la non-causalité de Granger entre la température du sol et le potentiel hydrique.

Revenons à notre question de recherche (QR1) :

(QR1) *Quelles sont les variables climatiques considérées dans notre étude qui ont des conséquences importantes sur l'évolution temporelle du potentiel hydrique du sol mesuré à la Station de biologie des Laurentides entre 2017 et 2020 ?*

Après avoir montré l'intérêt de modèles reposant sur la temporalité des données, au regard d'approches de classification comme les arbres de décision, nous nous sommes intéressés aux relations de causalité entre les variables prédictives et le potentiel hydrique du sol.

Nous avons mis en évidence des relations de causalité, au sens de Granger, entre les variables climatiques considérées dans notre étude et le potentiel hydrique du sol.

Cependant, nous avons également montré les limites de ces relations, principalement liées au modèle de Granger lui-même.

Cette analyse de la causalité, au sens de Granger, nous a permis de mettre en évidence le lien de causalité entre la température moyenne du sol et le potentiel hydrique lorsque les quatre dernières données temporelles de la température étaient considérées.

Nous allons revenir aux méthodes de classification utilisant les arbres de décision, et nous allons introduire de nouvelles *features*, qui vont nous permettre d'introduire une forme de temporalité dans l'analyse.

Nous l'avons souligné, les précipitations journalières ne semblent pas être une variable prédictive du potentiel. Au lieu d'introduire un délai à (t-2 jours), (t-3 jours) et (t-4 jours), nous allons explorer une variable cumulative des précipitations. En effet, les précipitations correspondent à un effet ponctuel, lié à un événement. Cet aspect ponctuel représente peut-être un événement marginal au regard des autres variables météorologiques considérées, et l'importance prédictive des précipitations peut alors ressortir comme ayant un faible poids dans l'arbre de décision.

Ainsi, nous allons introduire une variable qui représente le cumul des précipitations sur 2 jours.

Retour aux arbres de décision

Nous allons créer les nouvelles *features* suivantes :

- Température du sol considérée avec un délai de 2 jours : mean_T_lag_2days
- Température du sol considérée avec un délai de 3 jours : mean_T_lag_3days
- Température du sol considérée avec un délai de 4 jours : mean_T_lag_4days
- Précipitations cumulées sur 2 jours : prcp_cumul_2days

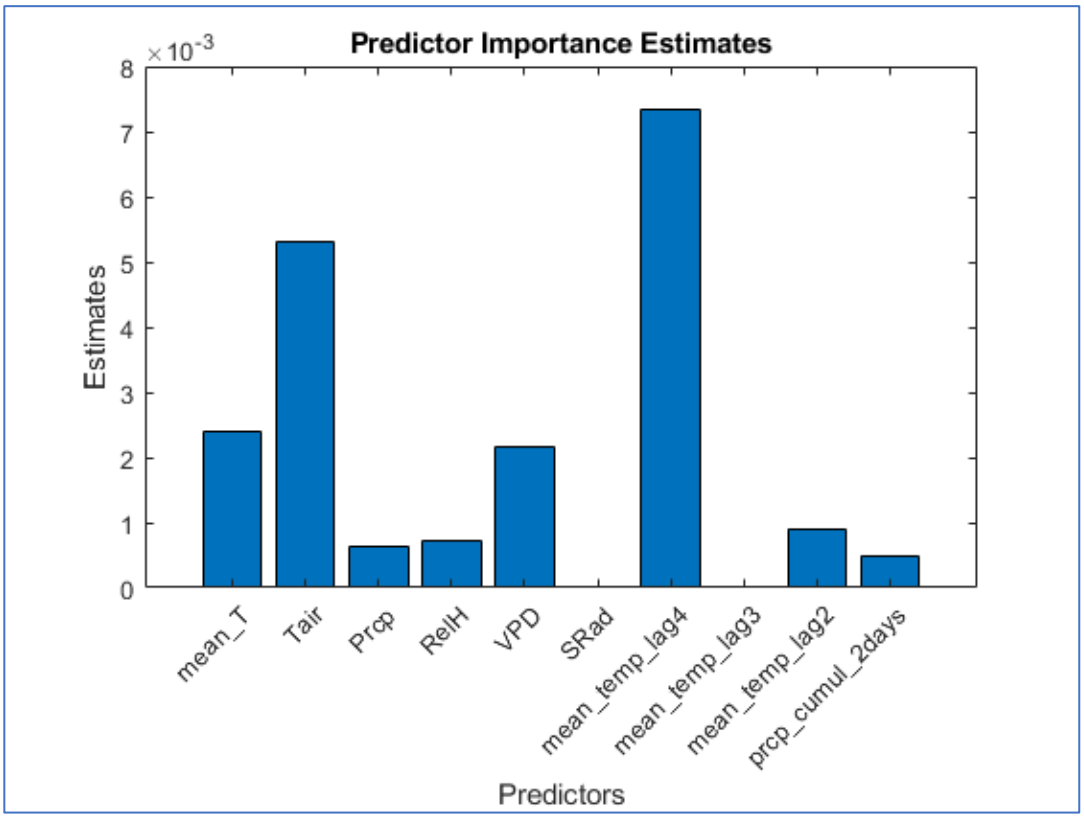
Les trois modèles M1, M2 et M3 sont de nouveau testés, avec l'introduction de ces nouvelles variables prédictives.

Résultats obtenus à la suite de l'introduction de ces quatre variables

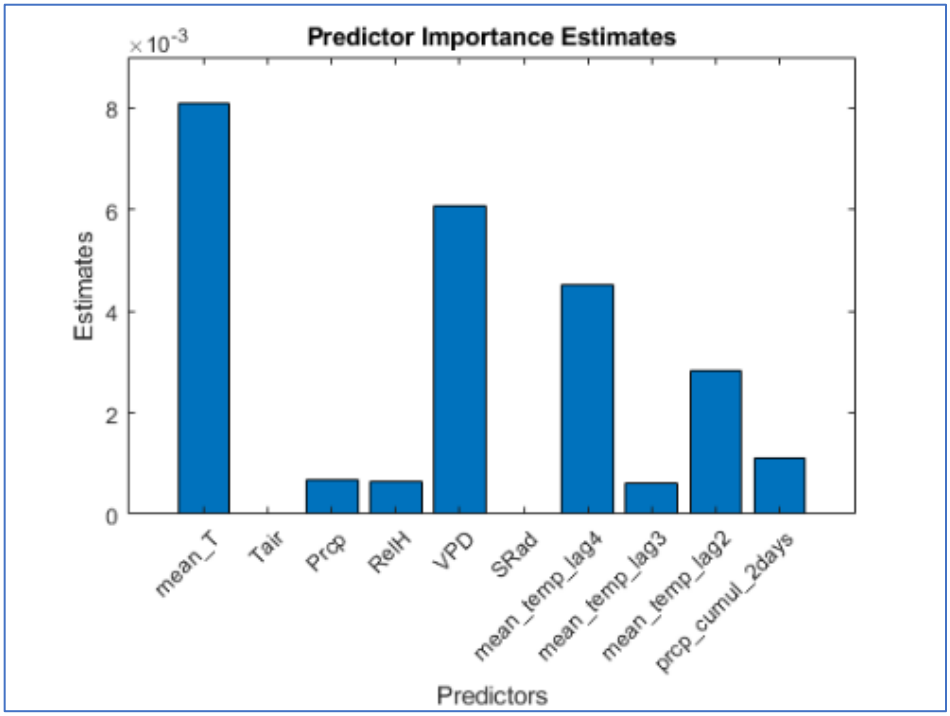
Pour les trois modèles testés, les pourcentages de prédictions exactes sont très proches. Nous présentons ici les résultats obtenus avec le modèle CART par défaut.

Pour MW : 53%
Pour HW : 64%
Pour HB : 46%

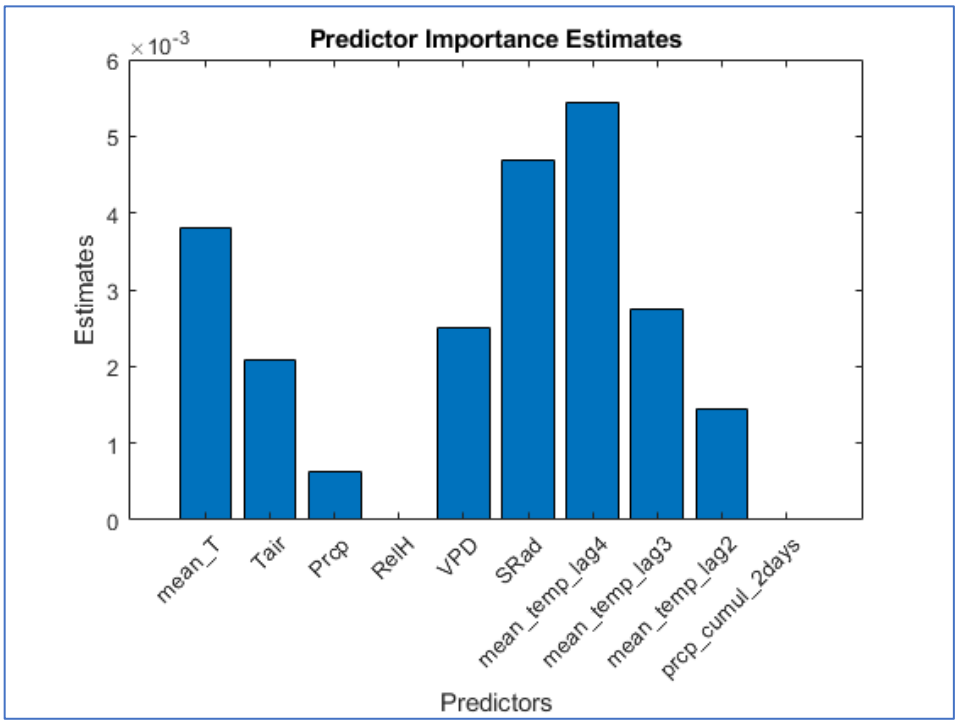
Modèle M1_0 (avec nouvelles features) pour MW



Modèle M1_0 (avec nouvelles features) pour HW



Modèle M1_0 (avec nouvelles features) pour HB



Nous avons précisé précédemment les avantages d'une méthode de classification s'appuyant sur les arbres de décision. Néanmoins, nous l'avons également souligné, la prédiction des valeurs de 2019 n'est pas très exacte, et en particulier pour les valeurs élevées du potentiel hydrique du sol.

Nous allons explorer une variété d'approches dans l'objectif d'augmenter le pourcentage d'exactitude du modèle, mais aussi afin d'obtenir un modèle qui permette de mieux prédire les valeurs élevées du potentiel hydrique. Dans notre contexte de stress climatiques, nous visons à mieux prédire ces valeurs extrêmes.

Matlab permet d'accéder à une boîte à outils qui facilite l'exploration de plusieurs modèles. Nous l'avons utilisé pour tester différentes approches de classification. L'idée centrale étant de générer un modèle capable de classifier le potentiel hydrique du sol selon sa valeur.

Nous avons conservé deux nouvelles *features* qui ont été créées : la température du sol considérée avec un délai de 4 jours et le cumul des précipitations sur 2 jours.

Les modèles ont été entrainés sur les données de 2020 et testés sur les données de 2019.

Une première session de tests a été réalisée avec 4 classes de potentiel, définies avec les mêmes critères que ceux utilisés pour les arbres de décision précédemment analysés : [Small, Medium, High, Very high]. Les intervalles considérés sont, pour rappel, les suivants : [0;40[, [40;80[, [80;120[et supérieur ou égal à 120 kPa.

La distinction entre la classe Medium et High est difficile à faire, suggérant peut-être une similitude trop grande entre les deux classes. Cela nous a donc amenés à réduire le nombre de classes, pour que les bornes des intervalles choisis soient plus discrimants et permettent une meilleure classification.

Ainsi, une seconde session de tests a été faite avec **trois classes** de potentiel : [Small, Medium, High]

Les intervalles considérés sont alors les suivants : [0;40[, [40;100[et supérieur ou égal à 100 kPa.

Entraînement du modèle sur toutes les données de 2020 et test sur les données de 2019

- Résultat obtenu avec une approche de classification de type SVM (machine à support de vecteurs), avec l'utilisation d'un noyau quadratique (Boser, 1996)

Au regard de la matrice de confusion obtenue, 85,7% des états *High* sont correctement prédits avec ce modèle et 83,5% des états *Small* sont également correctement prédits.

Model 2: SVM
Status: Trained

Training Results

Accuracy (Validation)	78.2%
Total cost (Validation)	38
Prediction speed	~5700 obs/sec
Training time	1.3664 sec

Test Results

Accuracy (Test)	59.6%
Total cost (Test)	74

Model Hyperparameters

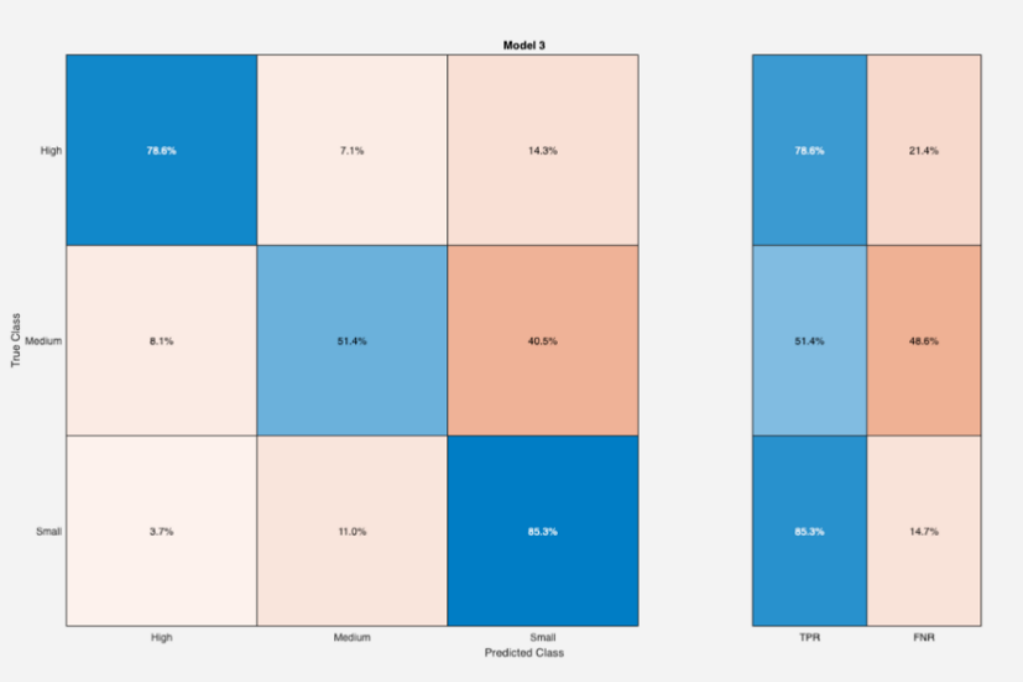
Preset: Quadratic SVM
Kernel function: Quadratic
Kernel scale: Automatic
Box constraint level: 1
Multiclass method: One-vs-One
Standardize data: true

Cette approche permet d'obtenir une prédiction plus élevée des valeurs hautes du potentiel (classe High), ce qui est préférable dans notre cas. En effet, nous visons la prédiction des valeurs élevées du potentiel hydrique du sol, afin d'anticiper les périodes de stress hydriques.

Ainsi, une méthode de classification capable de classer avec une plus grande exactitude les valeurs hautes sera privilégiée.

Une approche reposant sur des réseaux de neurones a également été testée, mais elle prédit moins bien les valeurs élevées du potentiel (86 % pour SVM contre 79%).

- Comparaison avec une approche reposant sur des réseaux de neurones (3 couches)



78,6% des états appartenant à la classe *High* sont correctement prédits comme *High*.

Phase 3 : Analyse des changements d'état du potentiel

À partir des trois états du potentiel définis à la phase 2, nous allons créer un modèle de chaînes de Markov à temps discret.

L'objectif est d'obtenir les probabilités de transitions d'un état à un autre, mais aussi le temps (en jours) mis pour atteindre un état, en partant d'un état différent.

L'idée centrale est de réaliser cette analyse pour chaque année (de 2017 à 2020) et ce, pour chaque peuplement, afin de mettre en évidence les différences dans les évolutions des changements d'état.

L'objectif est de répondre à la question de recherche (QR2) :

(QR2) Pouvons-nous définir différents états de ce potentiel hydrique du sol et déterminer les fréquences et les probabilités associés aux changements d'état ainsi mis en évidence ?

Nous considérons 3 classes pour décrire l'état du potentiel hydrique du sol en testant les intervalles de potentiel suivants :

(A) Small, Medium et High : [0;40[, [40;100[et *supérieur ou égal à 100 kPa*

Nous allons définir des chaînes de Markov à temps discret pour pouvoir évaluer les probabilités de passer d'un état du potentiel à un autre état, en nous appuyant sur les données expérimentales.

Description des chaînes de Markov à temps discret n

C'est un processus de Markov, i.e. un processus stochastique (aléatoire) qui possède la propriété de Markov.

La propriété de Markov se définit ainsi : l'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et n'est pas dépendante des états précédents.

La notion de chaîne permet de modéliser des phénomènes dynamiques aléatoires dans lesquels le passé n'intervient que dans le dernier instant de la chaîne (mémoire courte).

Le processus est une chaîne de Markov homogène (i.e., indépendant de n), si :

La probabilité de l'état X au temps (t_{n+1}) connaissant la probabilité de X aux temps (t_n), (t_{n-1})….(t_0) est égale à la probabilité de l'état X au temps (t_{n+1}) connaissant uniquement celle de X au temps (t_n), notée $P_{i,j}$

Les probabilités $P_{i,j}$ sont appelées probabilités de transition de la chaîne.

Elles sont supérieures ou égales à 0, et la somme des $P_{i,j}$ est égale à 1 ($\forall\ i, j\ \in\ Z$).

De façon simple, nous pouvons illustrer une chaîne de Markov à temps discret en considérant une pièce de monnaie non-truquée, qui possède deux états : Pile et Face (Figure 40).

Il est alors possible de passer de l'état Pile à l'état Face, avec une probabilité égale à 0,5. Mais il est également possible de passer de l'état Pile à l'état Pile, avec la même probabilité. Ainsi, nous pouvons définir toutes les transitions possibles, et leur probabilité associée. Ces probabilités constituent la matrice des probabilités de transition.

Figure 40 : Illustration d'une chaîne de Markov avec une pièce de monnaie non-truquée.

Le potentiel hydrique du sol peut être caractérisé, nous l'avons vu, au moyen d'une variable catégorielle qui représente son *état*. Cet état est défini en fonction d'intervalles de valeurs, que nous avons fixés.

Démarche

(1) Identifier quels jours ont eu lieu chaque état distinct du potentiel hydrique du sol.

(2) Créer la matrice de transition P (3x3), telle que :

	Low	Medium	High
Low	P(1,1)	P(1,2)	P(1,3)
Medium	P(2,1)	P(2,2)	P(2,3)
High	P(3,1)	P(3,2)	P(3,3)

Par exemple, le coefficient P(2,1) caractérise le nombre de transitions de l'état *Low* à l'état *Medium,* pendant la période d'échantillonnage choisie.

(3) Création d'un modèle de chaînes de Markov (mc) à temps-discret (*discrete-time Markov chain model*) caractérisé par la matrice de transition P.

La fonction de Matlab *dtmc()* est utilisée, telle que :

mc = dtmc(P, 'StateNames', State_SM)

La probabilité empirique de transitionner à (t+1) vers un état de point de potentiel élevé (High), étant donné un état de potentiel moyen (Medium) à t, est égale à :

P_{medium_high} = mc.P(3,2)

(4) Représentations graphiques

Deux types de représentations graphiques ont été faites pour permettre de visualiser les probabilités de transition, d'un état du potentiel à un autre état, et le nombre de jours nécessaire pour que chaque transition se fasse.

L'utilisation d'un tel processus présuppose une mémoire du système à court terme. L'état futur du système peut être déduit du passé, mais avec un historique assez faible. Cette mémoire à court terme est une des limites des chaînes de Markov à temps discret.

Cependant, cette approche permet de considérer le système en termes de transition d'état et de visualiser ces transitions.

Phase 4 : analyse des séries temporelles (TS)

Objectif : répondre à la question de recherche QR3

(QR3) Est-ce qu'un modèle prédictif de type ARIMAX, développé avec des données provenant de la SBL, permet de prédire l'évolution temporelle du potentiel hydrique du sol de ce même site, en considérant l'influence de variables climatiques exogènes ?

Cette phase se concentre sur *la temporalité des données en les considérant comme des séries temporelles.*

Dans une première approche, nous utiliserons les données relatives aux trois peuplements étudiés, échantillonnées sur une journée. Les années 2017 à 2020, incluses, seront analysées. (Cf. phase 2)

Les données météorologiques obtenues grâce au logiciel BioSIM ont été concaténées aux données de la SBL qui comprennent donc la température du sol journalière moyenne et le potentiel hydrique du sol journalier moyen, chaque variable ayant été préparée pour permettre l'analyse des séries temporelles (voir la phase 1 d'analyse).

Pour les données journalières, nous avons sélectionné les mesures moyennées par le biais de la fonction *groupsummary()* dans Matlab qui permet de considérer les mesures réalisées sur une journée et d'en déterminer une valeur moyenne. Contrairement à la fonction *sampling_frequency()* que nous avons implémentée, nous n'obtenons pas dans ce cas l'incertitude associée.

Nous avons ensuite déterminé la mesure moyenne par rapport aux deux sondes de température, mais aussi pour celles du potentiel hydrique.

Cette opération est possible, car les deux sondes indépendantes ont le même nombre de mesures. Il n'est pas donc nécessaire de calculer une moyenne pondérée, chaque moyenne journalière considérée ayant la même pondération égale à ½.

Nous avons donc :

$$mean_{temperature_{sampled}} = \frac{1}{2}\,(mean_{TMPA} + mean_{TMPB})$$

$$mean_{soil\ moisture_{sampled}} = \frac{1}{2}\,(mean_{SMSC} + mean_{SMSD})$$

Tout d'abord nous allons évaluer les paramètres de forme suivants : le coefficient d'aplatissement (*kurtosis*) et le coefficient d'asymétrie (*skewness)* pour le potentiel hydrique qui nous permettront de caractériser la distribution des données.

Nous nous attendons à une distribution leptocurtique (kurtosis normalisé $\gamma_2 > 0$) signe de valeurs anormales plus fréquentes.

Résultats obtenus en 2020

		MW	**HW**	**HB**
Kurtosis	Ψ	4,6	5,5	4,7
Skewness	Ψ	1,6	1,7	1,5

Résultats obtenus en 2017

		MW	**HW**	**HB**
Kurtosis	Ψ	5,4	4,3	3,4
Skewness	Ψ	1,3	1,2	0,9

Lorsque la valeur du kurtosis (non normalisé) est supérieure à 3, la distribution des données met en évidence davantage de données extrêmes. Ce que nous mettons de nouveau en évidence ici pour le potentiel hydrique du sol.

Pour l'asymétrie des données (*skewness*), des valeurs supérieures à 0, comme nous l'observons, indiquent que les données du potentiel hydrique se répartissent davantage à droite de la moyenne, exhibant alors une répartition asymétrique, s'éloignant d'une distribution normale.

Description de la démarche de Box-Jenkins

Nous allons appliquer dans cette approche, la démarche de Box-Jenkins pour l'analyse des séries temporelles (TS).

La méthodologie de Box-Jenkins17 comporte cinq étapes dont les objectifs sont d'identifier, sélectionner et valider des modèles autorégressifs et de moyenne mobile, pour des séries temporelles discrètes et univariées.

Les cinq étapes se définissent ainsi :

(1) Établir la stationnarité des séries temporelles.

Si celles-ci ne sont pas stationnaires, il faut alors différencier successivement les séries (ordre 1, ordre 2, etc.) jusqu'à atteindre la stationnarité.

Pour vérifier la stationnarité d'une série temporelle, plusieurs tests statistiques sont possibles, nous utiliserons dans notre cas le test de Philips-Perron.

En représentant la fonction d'autocorrélation de la série (ACF) et sa fonction d'autocorrélation partielle (PACF) il est également possible, visuellement, d'établir si la série est stationnaire. En effet, si les deux fonctions diminuent de façon exponentielle, ou tendent vers 0 après quelques délais temporels (appelés lags), cela permet de conclure à la stationnarité.

Qu'est-ce qu'une série stationnaire et pourquoi est-ce important de s'assurer de la stationnarité du signal d'intérêt ?

Une série temporelle stationnaire présente une structure du processus qu'elle représente, qui n'évolue pas dans le temps. On distinguera deux types de stationnarité : une forte et une faible (ou de second ordre)18.

Au sens fort, que nous regardions les propriétés statistiques du processus au point t ou au point (t+k), la série aura le même comportement. Pour conclure à ce type de stationnarité, il faut s'intéresser à la distribution conjointe de probabilité des t premières valeurs et la comparer à la distribution des (t+k) valeurs suivantes. La série sera alors stationnaire au sens fort, si les deux distributions sont égales.

---

[17] Box, G. E. P., Jenkins, G. M. , Reinsel, G.C. *Time Series Analysis: Forecasting and Control*. 4rd ed. Englewood Cliffs, NJ: Prentice Hall, 2008

[18] Cours Pr Jérôme Antoni, *Traitement du signal,* INSA Lyon (France)

La stationnarité faible sera souvent privilégiée, car elle est plus simple à montrer. En effet, il faudra montrer que l'espérance de la série est constante au cours du temps ; que sa variance est également constante au cours du temps, et que l'autocorrélation entre un point au temps t et un autre point au temps (t+k) dépend seulement du décalage temporel k.

Ainsi, pour résumer, les statistiques d'un signal stationnaire ne dépendent pas du temps.

Cette stationnarité est importante à vérifier lorsqu'on s'intéresse aux séries temporelles, car de nombreuses approches d'analyse reposent sur cette hypothèse, en particulier les régressions linéaires. Une régression linéaire effectuée sur une série non-stationnaire peut produire des régressions dites fallacieuses, car l'autocorrélation entre les valeurs d'une même série n'étant pas prise en compte, l'indice de corrélation R^2 peut être faussement élevé, impliquant une corrélation entre les variables qui est fausse.

Notons enfin qu'il devrait être aussi nécessaire de s'assurer de l'ergodicité[19] de la série, i.e. que la moyenne temporelle du signal soit égale à la moyenne d'un ensemble de mesures. L'hypothèse est généralement admise, car elle est difficile à montrer.

(2) Déterminer le type de processus suivi par le signal

Processus de type autorégressif (AR) ? : la fonction d'autocorrélation diminue graduellement, mais la fonction d'autocorrélation partielle tend vers zéro après quelques lags.

Processus de type moyenne mobile (MA) ? : l'ACF tend vers zéro après quelques lags, mais la fonction d'autocorrélation partielle (PACF) diminue graduellement.

Processus combinant les deux types (AR et MA) : à la fois l'ACF et la PACF diminuent graduellement.

Cette approche repose donc sur la visualisation des deux fonctions.

[19] Théorème majeur en physique statistique (fondement de la théorie cinétique des gaz de Boltzman), l'ergodicité est également importante en traitement du signal. Elle permet de relier la théorie (la valeur moyenne d'une grandeur calculée de manière statistique) à l'expérience (la moyenne d'un très grand nombre de mesures réalisées au cours du temps).

Les processus dits autorégressifs et ceux dits de type moyenne mobile seront explicités lorsque les modèles ARIMAX seront abordés.

(3) Spécifier le type de modèle choisi

S'il a été montré que la série temporelle n'est pas stationnaire et qu'une différentiation d'ordre 1 est nécessaire, alors il est possible d'analyser directement le signal non stationnaire, en combinant l'approche ARMA avec un certain degré de différenciation. Il s'agit alors d'un modèle appelé ARIMA[20]. Ainsi, il est possible de manipuler directement les séries d'origine, sans manipulation intermédiaire.
Cette approche est intéressante, car lors de l'utilisation du modèle à des fins prédictives, il n'est pas obligatoire de revenir à l'échelle d'origine en intégrant le résultat.

Le modèle ARIMA est défini en précisant les différents degrés du processus autorégressif (noté p), de la différenciation (noté d) et du processus de moyenne mobile (noté q). Ainsi, on parlera d'un modèle ARIMA(p,d,q).

L'enjeu est donc de déterminer à cette étape les valeurs de p, d et q qui permettent d'obtenir un ajustement (*fit*) de qualité. Des données dites d'entraînement seront choisies pour cette étape.

À cette étape, une évaluation du modèle peut se faire en traçant le graphique d'autocorrélation des résidus, pour s'assurer qu'ils ne sont pas autocorrélés : il faut en effet que les résidus correspondent bien à des aléas de mesure.

Le test de Ljung-Box peut être appliqué aux résidus pour tester s'ils sont autocorrélés (voir l'Annexe VI).

Hypothèse nulle du test : les données testées sont indépendamment distribuées (elles ne sont pas corrélées).

[20] Référence pour le modèle ARIMA : *Non-seasonal ARIMA*, Hyndman & Athanasopoulos (en particulier pour l'analyse des graphiques ACF et PACF)

- Expressions mathématiques et définition du corrélogramme[21,22,23]

La série $\{X_t\}$ est un processus autorégressif d'ordre p, AR(p) si, pour ce processus stationnaire on peut expliquer sa valeur à l'instant t en utilisant ses p termes précédents, tels que :

$$\forall\, t \in Z, x_t = \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \cdots + \Phi_p x_{t-p} + \varepsilon_t$$

$$\varepsilon_t \text{ un bruit blanc et } \Phi_i \text{ les paramètres du modèle avec } \Phi_p \neq 0 \text{ pour un ordre } p$$

On peut alors l'exprimer en un polynôme d'ordre p :

$$L_p(B)x_t = (1 - \Phi_1 B - \Phi_2 B^2 - \cdots - \Phi_p B^p).$$

B représente l'opérateur de rétropropagation (*backshift operator*), tel que : $By_t = y_{t-1}$

Il faut préciser que la notation utilisée dans Matlab remplace l'opérateur B par la lettre L.

C'est l'existence de la racine de ce polynôme L_p qui est testée dans le test de stationnarité de Philips-Perron. En effet, si ce polynôme admet une racine égale à 1, la série temporelle n'est pas stationnaire.

Le modèle est une *auto*-régression car on effectue une régression à partir de termes passés de la série elle-même.

Un processus $\{X_t\}$ est considéré MA (*moving average*) d'ordre q, si on peut exprimer sa valeur à l'instant t, comme une combinaison linéaire d'erreurs aléatoires ε (un bruit blanc), tel que :

$$\forall\, t \in Z, x_t = \varepsilon_t + \sum_{i=1}^{q} -\theta_i \varepsilon_{t-i}$$

Par convention, le signe négatif est adopté.

En combinant les deux processus, nous obtenons un processus ARMA d'ordre p et q.

[21] Cours de Vincent Lefieux, ENSAE-ENSAI formation continue (OpenClassrooms)
[22] Krispin, R. (2019) Hands-on Time series analysis with R, Ed. Packt Publishing Ltd (Birmingham, UK)
[23] Nielsen, A. (2019). Practical time series analysis, Ed. O'Reilly Media Inc.

En traçant l'autocorrélation pour différentes périodes on obtient un *corrélogramme* qui donne des indications sur les valeurs possibles de p et q.

Le corrélogramme ACF (*autocorrelation function)* permet d'évaluer la dépendance du potentiel à ses données passées. Il faut ainsi définir le nombre de périodes pour lequel la variable montre une forte corrélation avec ses données passées. L'ACF permet d'estimer le paramètre q de l'approche MA(q). En effet, la fonction d'autocorrélation d'un processus MA(q) tend vers zéro à la période (q+1).

Le corrélogramme PACF (*partial autocorrelation function)* permet quant à lui d'évaluer la corrélation partielle qui existe entre les *lags* de la série temporelle. En traçant les coefficients de la PACF, cela permet d'estimer le paramètre p de l'approche AR(p). L'autocorrélation partielle d'un processus AR(p) tend en effet vers zéro à la période (p+1). Il faut donc repérer à quel *lag* cela se produit pour avoir une estimation de p.

(4) Évaluation de la qualité de l'ajustement

Pour s'assurer que le modèle décrive suffisamment bien les données temporelles, il faudra déterminer la qualité de l'ajustement (*goodness of fit*). Le critère d'information de Akaike (AIC) ou le critère bayésien d'information (BIC) permettent de discriminer les modèles explorés.

Définition mathématique du critère d'information de Akaike, d'après (Krispin, 2019) :

$$AIC = 2k - 2\ln(\widehat{L})$$

Avec k = p+q

Et L la valeur maximale de la fonction de vraisemblance *(likelihood*)

Définition mathématique du critère bayésien d'information (Krispin, 2019) :

$$BIC = \ln(n)k - 2\ln(\hat{L})$$

Avec n, le nombre d'observations en entrée.

Nous voyons que si ln(n) > 2, alors le BIC est plus pénalisant que le coefficient AIC.

Il faut noter qu'une valeur basse de AIC ou BIC ne garantit pas que le modèle sélectionné respecte les hypothèses du modèle ARIMA. Il est donc recommandé d'analyser également la distribution des résidus.

(5) Validation du modèle sur des données test

Des données différentes des données d'entraînement sont ici utilisées pour valider le modèle.

(6) Prédiction sur un horizon futur

Quand le modèle est validé sur des données test, il peut être utilisé pour prédire des données sur un horizon futur.

Mise en place de la méthode de Box-Jenkins

(1) Stationnarité des séries temporelles décrivant l'évolution du potentiel hydrique du sol

Nous allons utiliser le test de Philips-Perron de la racine unitaire (*Philips-Perron test for unit root*).

Le seuil de significativité (p) est fixé à 0,05.

Aucun délai n'est considéré.

T-test standard.

Hypothèse nulle du test H_0 : présence d'une racine égale à 1

Si la série temporelle (TS) n'est pas stationnaire, c'est-à-dire qu'une des racines du polynôme exprimant la TS est égale à 1, alors il faudra la différencier, puis refaire le test de Philips-Perron pour s'assurer de la stationnarité de la série temporelle différenciée une fois. Différencier à nouveau, si cela n'est pas encore le cas.

Les données de 2020 sont utilisées.

Les résultats obtenus sont résumés dans le Tableau 11.

Tableau 11 : Résultats des tests de Philips-Perron.

Potentiel \ Peuplement	MW	HW	HB
Ψ	Non rejet de H_0 p = 0,42 Série temporelle non stationnaire (présence d'une racine égale à 1)	Non rejet de H_0 p = 0,33 Série temporelle non stationnaire	Non rejet de H_0 p = 0,43 Série temporelle non stationnaire
$\Delta\Psi = (\Psi_t - \Psi_{t-1})$ (différence d'ordre 1)	Rejet de H_0 p = $1x10^{-3}$ La différence d'ordre 1 de Ψ par rapport au temps est stationnaire.	Rejet de H_0 p = $1x10^{-3}$ La différence d'ordre 1 de Ψ par rapport au temps est stationnaire.	Rejet de H_0 p = $1x10^{-3}$ La différence d'ordre 1 de Ψ par rapport au temps est stationnaire.

Ainsi, le test de Philips-Perron montre que les trois séries temporelles du potentiel hydrique du sol ne sont pas stationnaires, ce qui implique la nécessité de les différencier; leur différence d'ordre 1 étant alors stationnaire.

Nous montrons de la même façon que pour les trois peuplements considérés, les séries temporelles de la température moyenne du sol ne sont pas stationnaires ($p > 0.05$), mais que leurs différences premières le sont.

Pour les variables météorologiques de notre étude, seule la série temporelle représentant la température de l'air n'est pas stationnaire ($p = 0,21$).

Conclusion *: Il sera nécessaire de différencier les séries temporelles du potentiel hydrique du sol, pour s'assurer de leur stationnaire, postulat permettant d'appliquer les approches de type autorégressives avec moyenne mobile (ARMA).*

La non-stationnarité des trois potentiels hydriques du sol, correspondant aux trois peuplements étudiés, peut également être mise en évidence en traçant l'autocorrélation des trois séries temporelles et en observant que celle-ci varie au cours du temps, en décroissant lentement vers 0 (Figure 41).

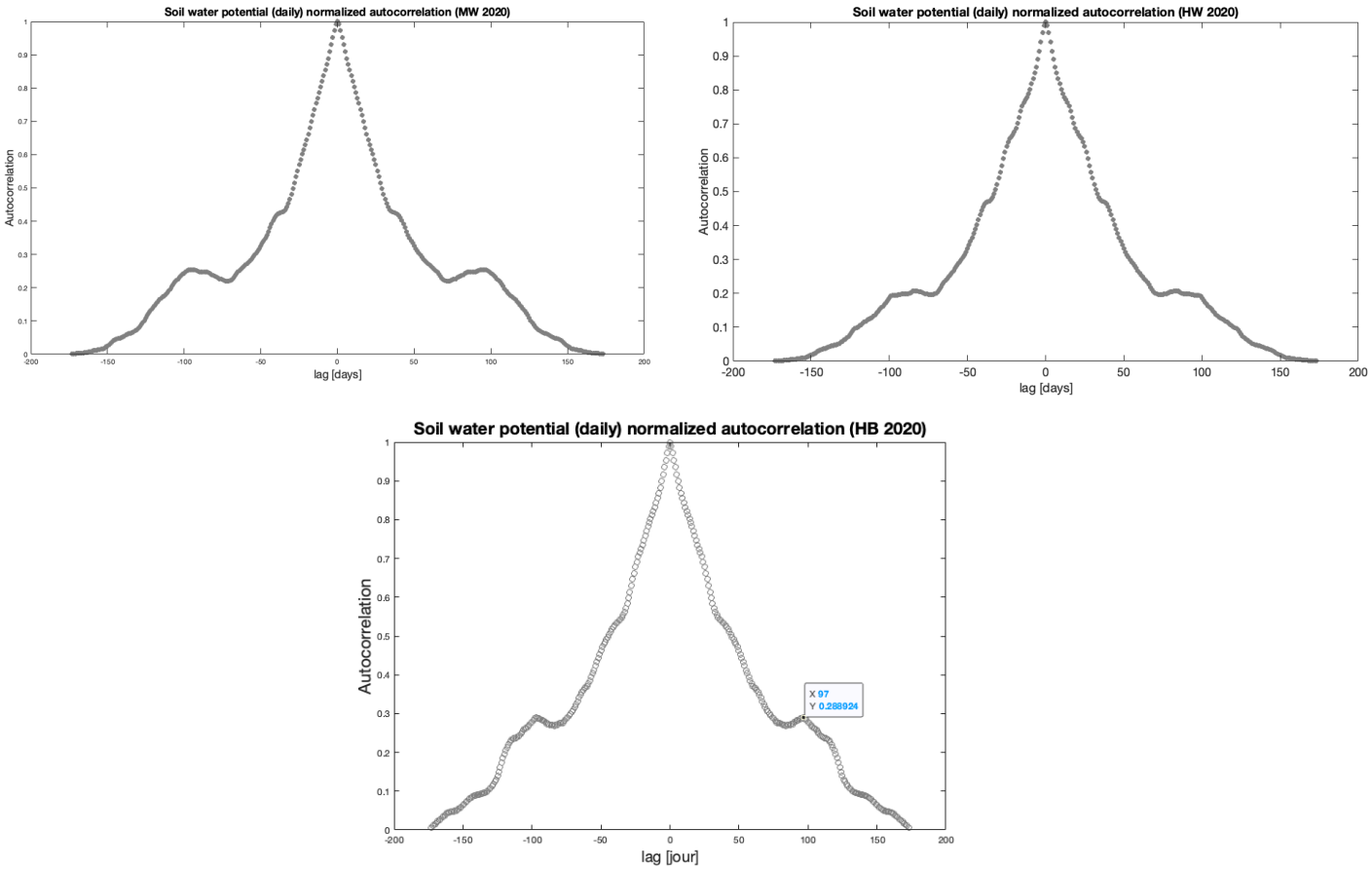
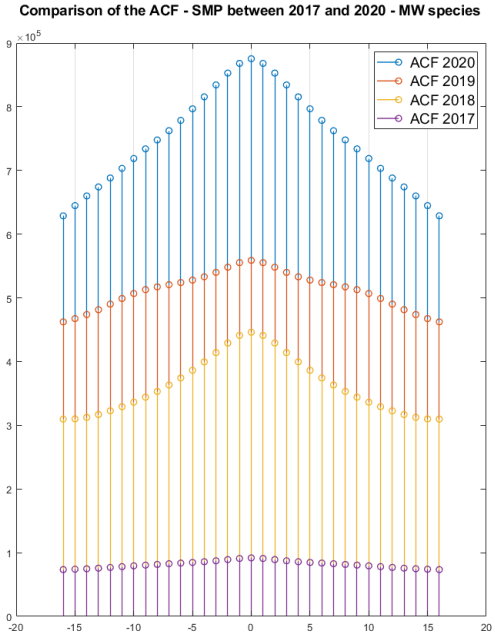
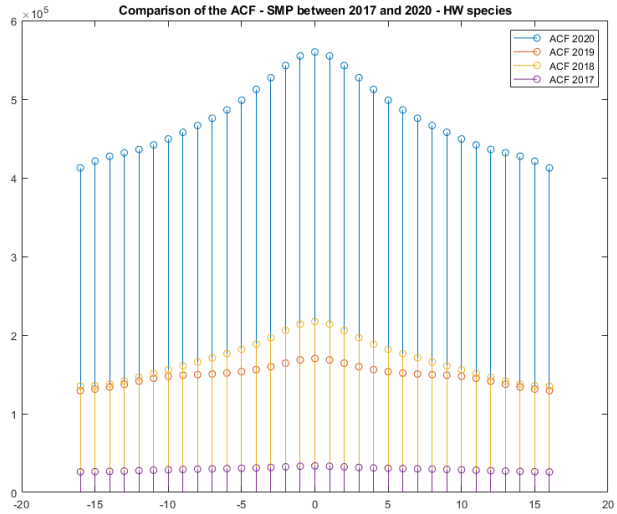
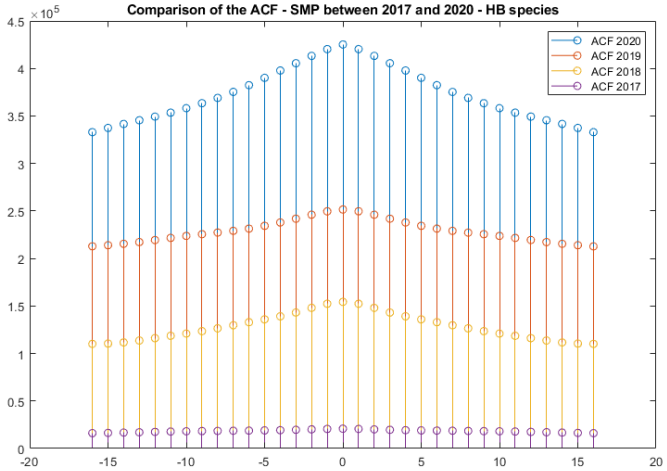


Figure 41 : Autocorrélation pour chaque peuplement (2020).

Analyse de l'évolution de l'autocorrélation entre 2017 et 2020



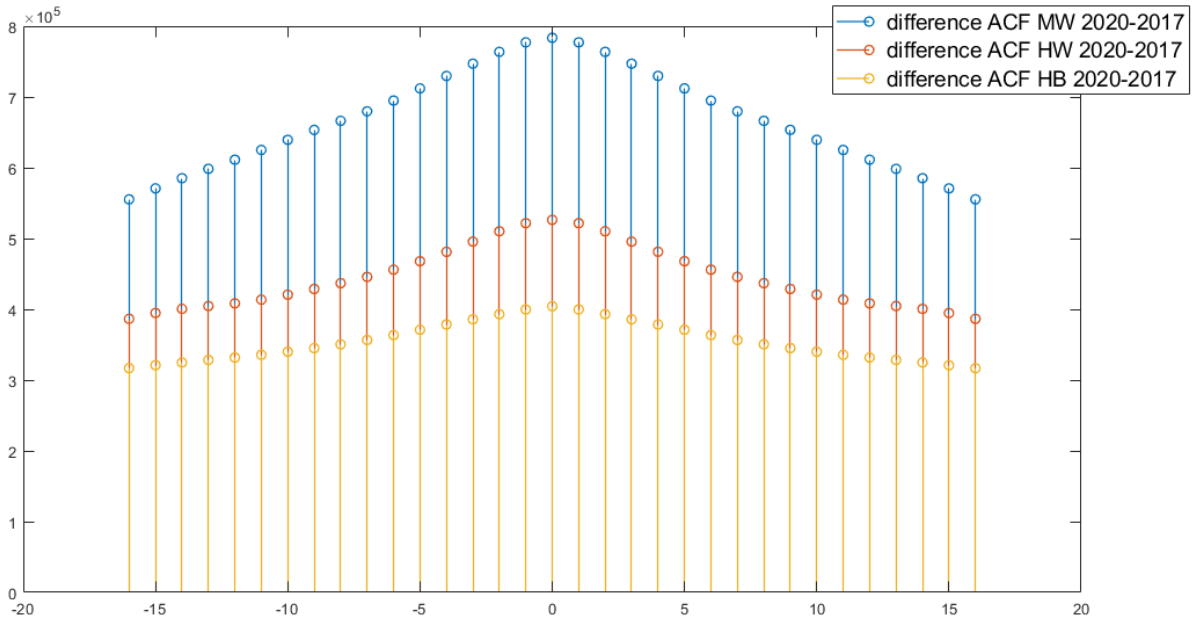




Pour les trois peuplements, la valeur des coefficients de l'autocorrélation du potentiel hydrique augmente entre 2017 et 2020.

Si nous traçons la différence des coefficients obtenus en 2020 et en 2017, pour les trois peuplements, nous mettons en évidence l'augmentation de l'autocorrélation.

La plus forte augmentation entre 2017 et 2020 est mise en évidence pour le peuplement MW, puis HW et enfin HB.



La fonction d'autocorrélation représente la dépendance statistique entre les valeurs de la série chronologique x[n] au cours du temps. La valeur x[1] correspondant à la valeur de la série observée au temps t_1. L'autocorrélation peut se traduite par le produit moyen de la séquence x[n] avec une version décalée de m d'elle-même, i.e., x[n+m].

Elle s'exprime ainsi : $\phi_{xx}[m] = \lim_{N\to+\infty} \frac{1}{2N+1} \sum_{n=-N}^{n=N} x[n]x[n+m]$

Il a été montré que l'augmentation de l'autocorrélation au cours du temps était un signe de rapprochement d'un point de bifurcation (voir le chapitre 1 du mémoire).

Le peuplement HB a connu la moins grande augmentation entre 2017 et 2020.

Nous pouvons aussi, de façon similaire, tracer le potentiel au temps (t+1) en fonction de la valeur mesurée au temps (t). Le coefficient de corrélation de Pearson est ainsi évalué pour le peuplement MW, entre 2017 et 2020, selon la démarche décrite par (Laitinen, 2021)[24]. La Figure 42 et la Figure 43 représentent respectivement l'année 2018 et l'année 2020.

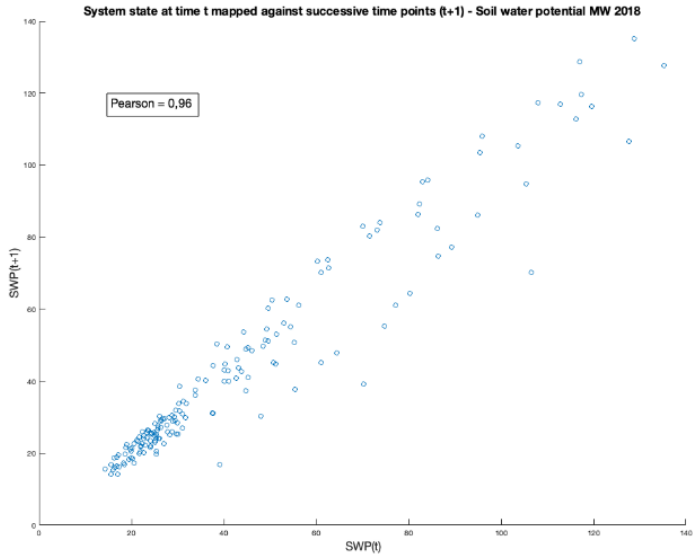


Figure 42 : Représentation du potentiel hydrique à (t+1) en fonction des valeurs mesurées au temps t (MW, 2018).

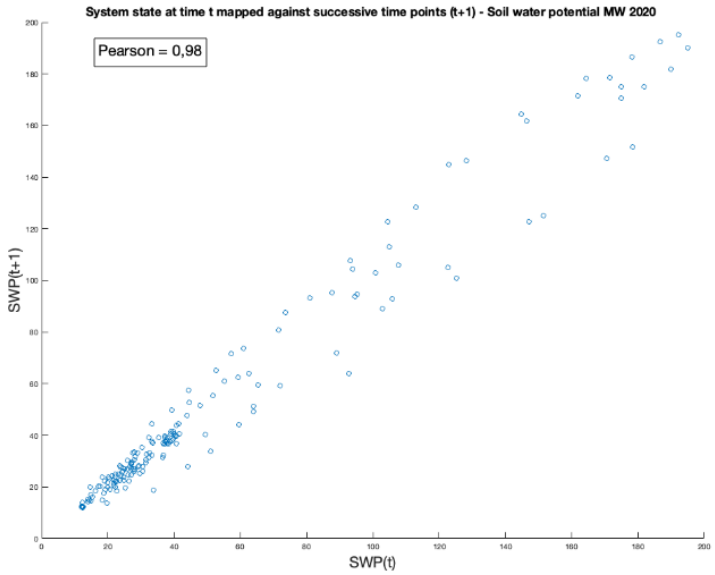


Figure 43 : Représentation du potentiel hydrique à (t+1) en fonction des valeurs mesurées au temps t (MW, 2020).

[24] Laitinen, V., Dakos, V., Lahti, L. (2021) Probabilistic early warning signals, *Ecology and Evolution*

L'approche est inspirée de la Figure 44 issue de (Laitinen, 2021) qui montre l'augmentation de l'autocorrélation au lag-1 à l'approche d'un point de basculement (f) en comparaison avec un signal éloigné de ce point (e).

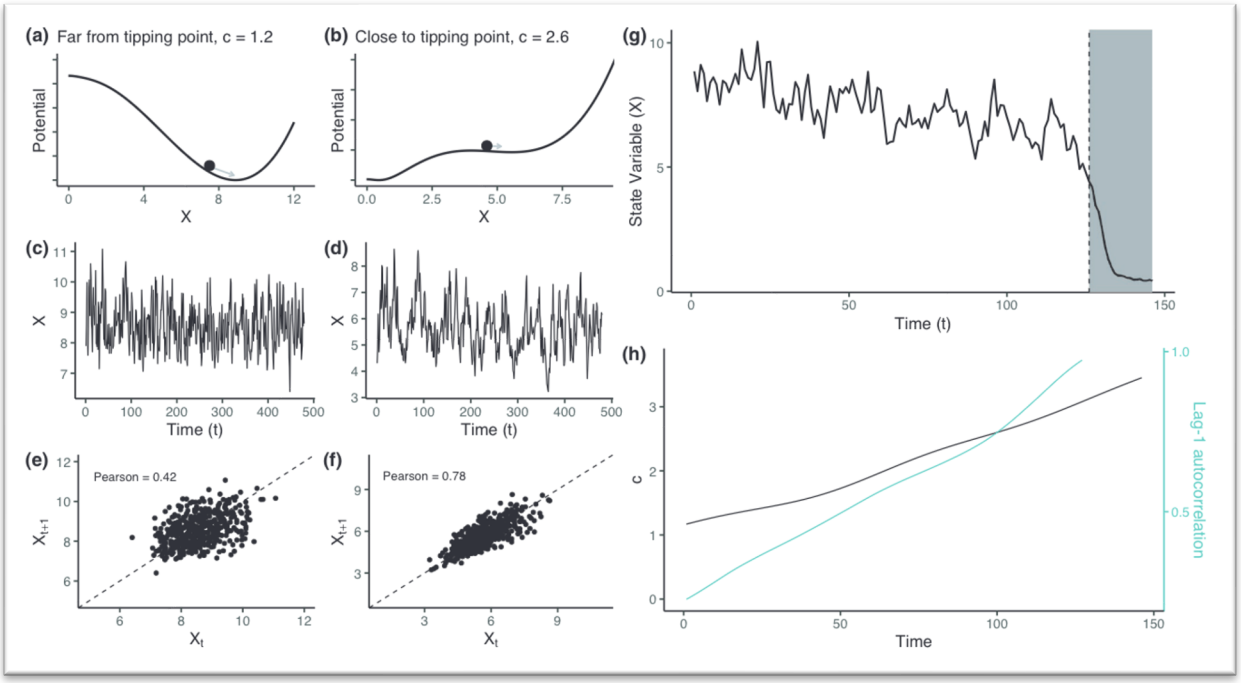


Figure 44 : Figure extraite de Laitinen (2021). Comparaison entre (a) un système loin d'un point de basculement et (b) proche d'un point de basculement. L'évolution temporelle de chacun des systèmes est représentée en (c) et en (d).

Avant de poursuivre la méthode de Box-Jenkins, nous allons étudier les corrélations-croisées entre le potentiel hydrique et les variables météorologiques.

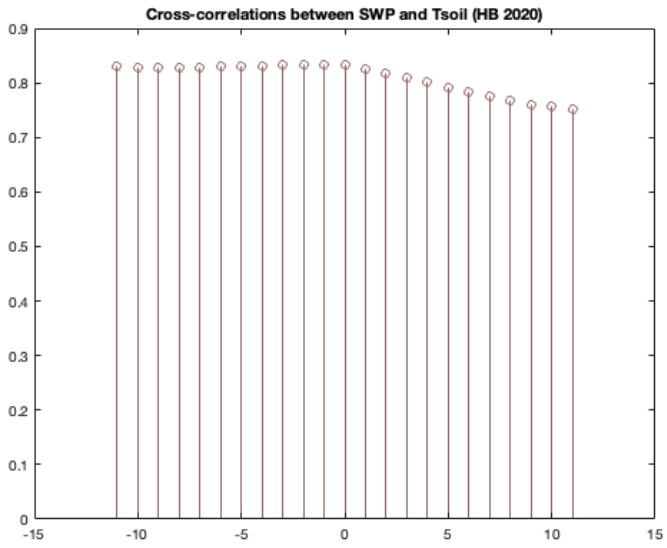
Analyse des corrélation-croisées

Cette analyse permettra de mettre en évidence les dépendances temporelles entre les variables. Cet aspect a été abordé, dans une certaine mesure, lorsque nous avons étudié la causalité au sens de Granger. Or, dans cette étude, nous regardions si des valeurs antérieures à celle du potentiel mesuré au temps t permettaient une meilleure prédiction de cette même valeur. Nous n'avons pas exploré la dépendance postérieure.

En traitement du signal, la corrélation croisée est la mesure de la similarité entre deux signaux. Lorsqu'elle est maximale, les signaux sont fortement associés. À la différence de l'autocorrélation, deux séries chronologiques distinctes sont considérées, soient x[n] et y[n]. Ce qui se traduit alors mathématiquement par le produit de convolution suivant : $(\phi_x * \phi_y)(n) = \lim_{N\to+\infty} \frac{1}{2N+1} \sum_{n=-N}^{n=N} x[n]y[n+m]$

Nous allons calculer les corrélations-croisées pour un lag de 11 jours. Les coefficients ont été normalisés, car nous comparons des mesures qui ont des ordres de grandeur différents. L'ajout d'un lag correspond à un décalage de la seconde série temporelle, et la corrélation est alors calculée en considérant la série temporelle décalée.

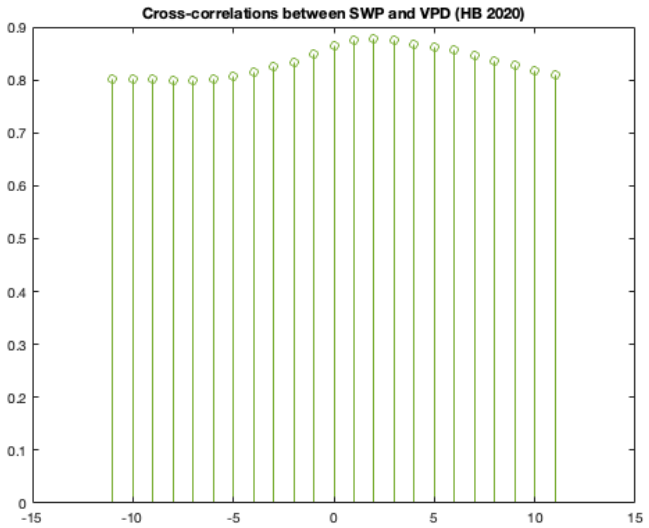
- **Pour le peuplement HB**



La corrélation croisée entre le potentiel et la température du sol reste forte de t à (t - 4 jours).

Cela suggère que la température du sol mesurée jusqu'au jour (t – 4j) est associée au potentiel mesuré au jour t.

Nous retrouvons les résultats obtenus avec la causalité de Granger.



Pour le VPD, cette corrélation croisée reste forte de t à (t + 4 jours).

Le VPD mesuré jusqu'au jour (t+4j) est associé au potentiel mesuré au jour t.

L'influence de l'humidité relative sur le potentiel se fait ressentir avant celle du VPD (Figure 45). En effet, les corrélations-croisées sont fortes jusqu'à 6 jours avant le maximum du potentiel hydrique, avec une valeur maximale 2 jours avant (t-2j). L'irradiance solaire exerce une influence sur le potentiel hydrique qui s'étale sur plusieurs jours (t + 6 jours).

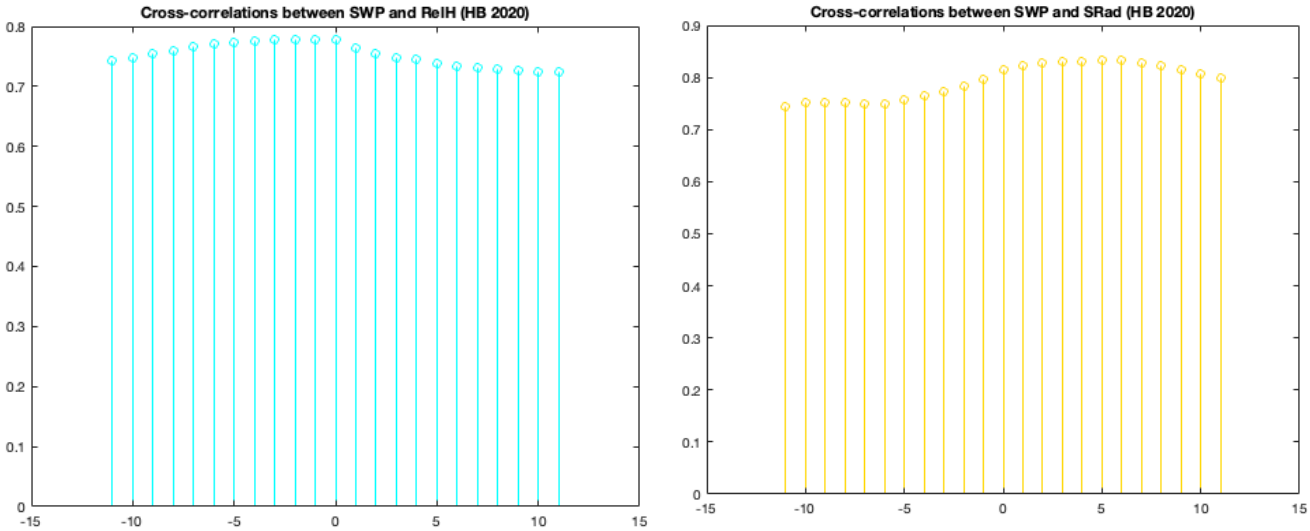
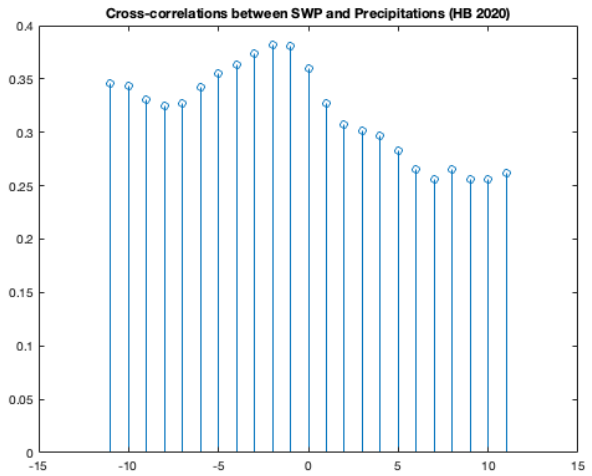
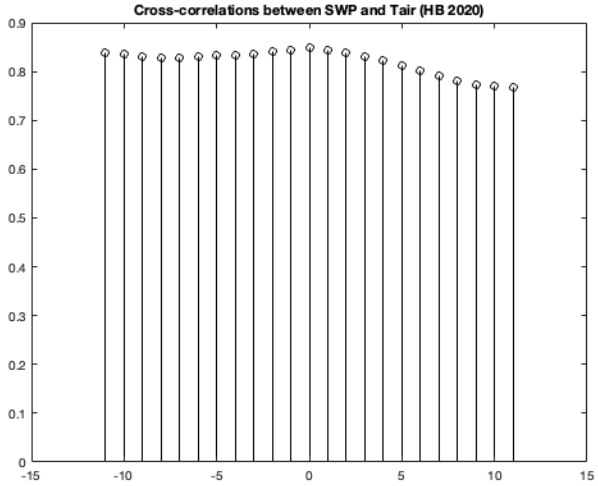


Figure 45 : Coefficients de corrélation-croisée entre le potentiel hydrique du sol et l'humidité relative (en bleu) et l'irradiance solaire (en jaune)



Concernant les précipitations, rappelons-nous que nos données étant journalières, leur impact se mesure un jour après sur le potentiel.

Nous observons que les précipitations sont maximales un jour avant que ne le soit le potentiel hydrique. Nous retrouvons donc le décalage entre les précipitations et leur effet sur le potentiel. Notons la diminution importante au temps t de la corrélation-croisée. L'influence entre les deux variables ne dure pas dans le temps : il est ponctuel.



Pour la température de l'air (en noir), elle est fortement associée au potentiel hydrique au temps t (lag nul). Elles ont une simultanéité dans leur influence réciproque.

Les coefficients calculés pour le peuplement HB sont présentés dans le Tableau 12.

Tableau 12 : Coefficients de corrélation-croisée obtenus pour le peuplement HB (2020).

	Tsoil	VPD	RelH	SRad	Prcp	Tair
t – 6j	0.8294	0.8013	0.7700	0.7503	0.3417	0.8307
t – 5j	0.8304	0.8070	0.7730	0.7569	0.3546	0.8326
t – 4j	0.8313	0.8154	0.7752	0.7653	0.3630	0.8340
t – 3j	0.8323	0.8246	0.7771	0.7722	0.3732	0.8367
t – 2j	**0.8329**	0.8342	**0.7785**	0.7824	**0.3816**	0.8416
t – 1j	**0.8329**	0.8492	0.7776	0.7977	0.3808	0.8447
t = 0	0.8319	0.8657	0.7777	0.8151	0.3594	**0.8479**
t + 1j	0.8247	0.8762	0.7639	0.8234	0.3274	0.8438
t + 2j	0.8171	**0.8787**	0.7542	0.8282	0.3074	0.8388
t + 3j	0.8090	0.8739	0.7482	0.8307	0.3017	0.8316
t + 4j	0.8005	0.8661	0.7443	0.8307	0.2963	0.8228
t + 5j	0.7916	0.8607	0.7393	0.8323	0.2827	0.8129
t + 6j	0.7826	0.8561	0.7337	**0.8334**	0.2647	0.8023

Nous notons également les coefficients de corrélation-croisée obtenus pour le peuplement MW (Tableau 13).

Tableau 13 : Coefficients de corrélation-croisée obtenus pour le peuplement MW (2020).

	Tsoil	VPD	RelH	SRad	Prcp	Tair
t – 6j	0.7780	0.7414	0.7176	0.6880	0.3156	0.7866
t – 5j	0.7790	0.7468	**0.7182**	0.6919	0.3295	0.7884
t – 4j	0.7805	0.7572	0.7169	0.7001	0.3359	0.7916
t – 3j	0.7820	0.7689	0.7150	0.7106	0.3423	0.7961
t – 2j	0.7831	0.7822	0.7122	0.7221	**0.3465**	0.8011
t – 1j	**0.7834**	0.7981	0.7085	0.7368	0.3440	0.8050
t = 0	0.7825	0.8172	0.7042	0.7557	0.3146	**0.8078**
t + 1j	0.7772	0.8313	0.6943	0.7714	0.2784	0.8051
t + 2j	0.7701	**0.8380**	0.6867	0.7828	0.2534	0.7991
t + 3j	0.7621	0.8372	0.6817	0.7893	0.2483	0.7901
t + 4j	0.7542	0.8334	0.6784	0.7921	0.2432	0.7812
t + 5j	0.7462	0.8302	0.6749	0.7959	0.2313	0.7727
t + 6j	0.7376	0.8259	0.6716	**0.7986**	0.2180	0.7631

Enfin, ce qui ressort pour le peuplement HW est résumé dans le Tableau 14.

Tableau 14 : Coefficients de corrélation-croisée obtenus pour le peuplement HW (2020).

	Tsoil	VPD	RelH	SRad	Prcp	Tair
t – 6j	0.7829	0.7389	**0.7108**	0.6902	0.3054	0.7928
t – 5j	0.7837	0.7453	0.7095	0.6929	0.3165	0.7941
t – 4j	0.7850	0.7572	0.7061	0.7001	0.3262	0.7970
t – 3j	0.7868	0.7720	0.7017	0.7112	0.3359	0.8025
t – 2j	0.7885	0.7881	0.6968	0.7247	**0.3416**	0.8099
t – 1j	**0.7889**	0.8076	0.6899	0.7423	0.3330	0.8157
t = 0	0.7870	0.8291	0.6815	0.7630	0.2934	**0.8182**
t + 1j	0.7812	0.8441	0.6709	0.7806	0.2518	0.8150
t + 2j	0.7731	**0.8492**	0.6635	0.7920	0.2240	0.8076
t + 3j	0.7636	0.8454	0.6590	0.7973	0.2243	0.7966
t + 4j	0.7534	0.8402	0.6549	0.8005	0.2263	0.7844
t + 5j	0.7430	0.8345	0.6510	0.8033	0.2218	0.7719
t + 6j	0.7327	0.8264	0.6482	**0.8035**	0.2131	0.7595

Nous pouvons résumer les trois résultats obtenus sous la forme d'un schéma (Figure 46).

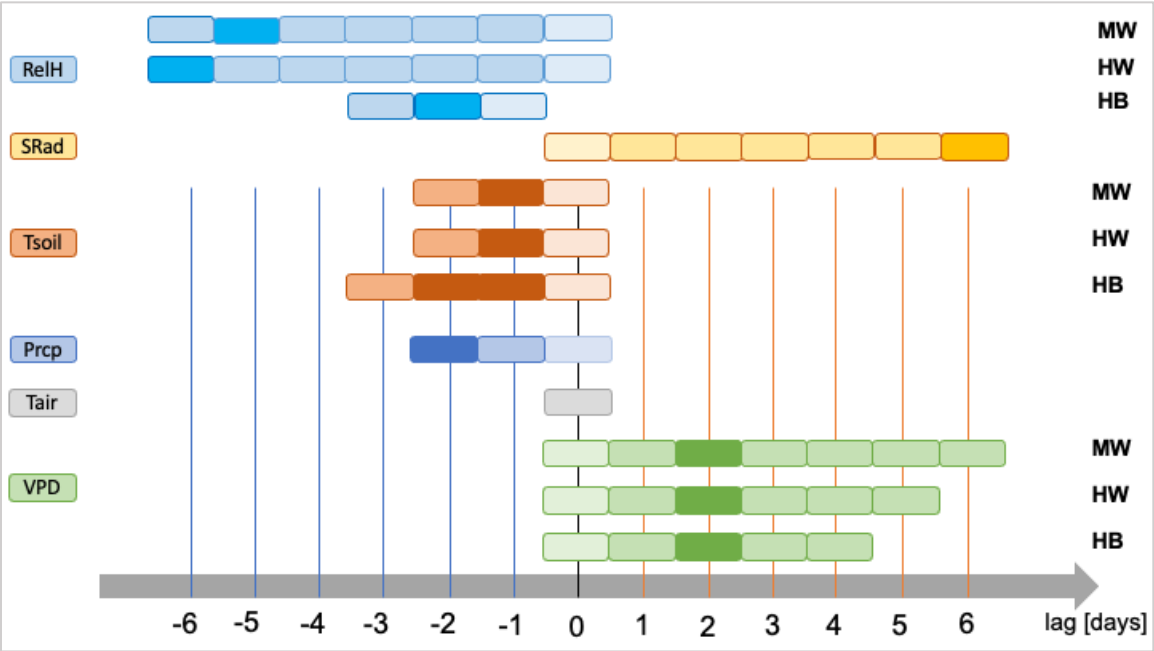


Figure 46 : Résumé de l'étude des corrélations-croisées obtenues en 2020 pour les trois peuplements.

L'analyse des corrélations-croisées nous permet de montrer que chaque peuplement va présenter des corrélations différentes au cours du temps. Même si pour les variables, à l'exception de l'humidité relative, les maxima se situent à des délais souvent identiques*, les corrélations s'estompent à des jours différents. Cela est le cas pour le VPD, dont l'influence se fait sentir pour MW jusqu'à 6 jours après la valeur*

la plus élevée du potentiel. Pour HB, après 4 jours, cette influence a diminué comparativement au jour t (lag = 0).

Si nous regardons la température moyenne du sol, elle influence le potentiel hydrique jusqu'à 3 jours avant. Nous avions mis en évidence une temporalité de 4 jours avec la causalité de Granger, ce qui reste comparable lorsque nous regardons les coefficients de corrélation-croisée, qui restent élevés même 5 jours après. À (t-4jours) le coefficient devient légèrement inférieur à celui obtenu à (t), mais il reste très proche.

L'influence de l'humidité relative est maximale 2 jours avant pour le peuplement majoritairement composé de hêtres, mais celle-ci est plus longue pour HW et MW (6 et 5 jours avant, respectivement).

Le potentiel hydrique et la température de l'air mesurés au jour (t) ont leur corrélation-croisée maximale, ne montrant donc aucune influence antérieure ou postérieure de la température de l'air sur le potentiel hydrique du sol.

Corrélations-croisées durant les sécheresses-flash

Nous allons déterminer les corrélations-croisées sur les données relatives aux trois périodes de sécheresse-flash (dry-down) afin de mettre en évidence d'éventuels changements.

(1) Corrélations-croisées obtenues pour HB (sécheresse-flash de mai dd1, juin dd2 et septembre dd3) et représentations graphiques des coefficients

dd1	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.6877	0.6077	0.6144	0.5768	0.6444
t – 7j	0.7368	0.6362	0.6826	0.6118	0.6898
t – 6j	0.7877	0.6632	0.7590	0.6443	0.7371
t – 5j	0.8372	0.7137	0.8107	0.6948	0.7854
t – 4j	0.8835	0.7655	0.8567	0.7537	0.8260
t – 3j	0.9204	0.8142	0.8880	0.8024	0.8595
t – 2j	0.9485	0.8515	0.9195	0.8504	0.8948
t – 1j	0.9668	0.8909	0.9354	0.8907	0.9201
t = 0	0.9786	0.9325	0.9443	0.9300	0.9297
t + 1j	0.9458	0.9309	0.8879	0.9234	0.9148
t + 2j	0.9072	0.9229	0.8343	0.9114	0.8909
t + 3j	0.8608	0.9026	0.7873	0.8982	0.8602
t + 4j	0.8074	0.8860	0.7327	0.8803	0.8099
t + 5j	0.7500	0.8623	0.6807	0.8621	0.7441
t + 6j	0.6923	0.8269	0.6371	0.8379	0.6807
t + 7j	0.6376	0.7812	0.6057	0.8030	0.6132
t + 8j	0.5860	0.7415	0.5738	0.7710	0.5508

dd2	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.6253	0.5265	0.6206	0.5006	0.5995
t – 7j	0.6829	0.5778	0.6788	0.5443	0.6566
t – 6j	0.7345	0.6267	0.7301	0.5890	0.7065
t – 5j	0.7807	0.6769	0.7724	0.6331	0.7525
t – 4j	0.8283	0.7342	0.8117	0.6852	0.8001
t – 3j	0.8761	0.7906	0.8514	0.7371	0.8482
t – 2j	0.9151	0.8363	0.8858	0.7812	0.8932
t – 1j	0.9435	0.8815	0.9062	0.8255	0.9299
t = 0	0.9620	0.9226	0.9187	0.8710	0.9562
t + 1j	0.9390	0.9337	0.8864	0.8878	0.9404
t + 2j	0.9127	0.9394	0.8529	0.8987	0.9189
t + 3j	0.8839	0.9405	0.8185	0.9005	0.8935
t + 4j	0.8536	0.9283	0.7901	0.9078	0.8677
t + 5j	0.8223	0.9073	0.7663	0.9019	0.8380
t + 6j	0.7888	0.8816	0.7432	0.8869	0.8069
t + 7j	0.7528	0.8535	0.7185	0.8776	0.7709
t + 8j	0.7153	0.8225	0.6925	0.8510	0.7278

dd3	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.6287	0.5555	0.6379	0.5381	0.6163
t – 7j	0.6793	0.6194	0.6861	0.5809	0.6665
t – 6j	0.7319	0.6820	0.7369	0.6323	0.7238
t – 5j	0.7871	0.7340	0.7938	0.6943	0.7671
t – 4j	0.8359	0.7881	0.8398	0.7369	0.8146
t – 3j	0.8765	0.8211	0.8799	0.7730	0.8628
t – 2j	0.9074	0.8654	0.9059	0.8277	0.8968
t – 1j	0.9377	0.9132	0.9332	0.8891	0.9258
t = 0	0.9698	0.9419	0.9690	0.9251	0.9540
t + 1j	0.9364	0.9316	0.9332	0.9180	0.9300
t + 2j	0.9004	0.9344	0.8918	0.9154	0.8992
t + 3j	0.8628	0.9122	0.8565	0.9039	0.8532
t + 4j	0.8256	0.8924	0.8226	0.8985	0.8169
t + 5j	0.7871	0.8751	0.7879	0.8928	0.7738
t + 6j	0.7470	0.8544	0.7523	0.8713	0.7219
t + 7j	0.7082	0.8145	0.7185	0.8360	0.6580
t + 8j	0.6732	0.7751	0.6829	0.7962	0.6106

(2) Corrélations-croisées obtenues pour HW (sécheresse-flash de mai dd1, juin dd2 et septembre dd3)

dd1	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.4631	0.3693	0.4701	0.3623	0.4166
t – 7j	0.5337	0.4134	0.5516	0.4087	0.4835
t – 6j	0.6156	0.4647	0.6497	0.4630	0.5602
t – 5j	0.7014	0.5323	0.7369	0.5293	0.6408
t – 4j	0.7824	0.6027	0.8057	0.5951	0.7170
t – 3j	0.8478	0.6680	0.8400	0.6457	0.7893
t – 2j	0.8997	0.7265	0.8562	0.6906	0.8632
t – 1j	0.9368	0.7847	0.8569	0.7356	0.9257
t = 0	0.9593	0.8409	0.8532	0.7872	0.9622
t + 1j	0.9420	0.8644	0.8063	0.8099	0.9557
t + 2j	0.9131	0.8761	0.7618	0.8235	0.9267
t + 3j	0.8759	0.8808	0.7197	0.8327	0.8882
t + 4j	0.8342	0.8941	0.6695	0.8376	0.8391
t + 5j	0.7893	0.8963	0.6259	0.8421	0.7838
t + 6j	0.7412	0.8724	0.5991	0.8386	0.7312
t + 7j	0.6912	0.8249	0.5891	0.8222	0.6692
t + 8j	0.6416	0.7801	0.5784	0.8027	0.6042

dd2	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.6075	0.5005	0.6213	0.4836	0.5893
t – 7j	0.6650	0.5528	0.6786	0.5266	0.6459
t – 6j	0.7163	0.6021	0.7270	0.5690	0.6951
t – 5j	0.7621	0.6529	0.7653	0.6109	0.7409
t – 4j	0.8089	0.7109	0.7995	0.6620	0.7884
t – 3j	0.8553	0.7687	0.8332	0.7142	0.8360
t – 2j	0.8952	0.8184	0.8638	0.7607	0.8815
t – 1j	0.9275	0.8688	0.8839	0.8093	0.9201
t = 0	0.9515	0.9145	0.8992	0.8596	0.9495
t + 1j	0.9308	0.9276	0.8672	0.8791	0.9328
t + 2j	0.9061	0.9344	0.8342	0.8920	0.9094
t + 3j	0.8780	0.9359	0.8000	0.8949	0.8804
t + 4j	0.8480	0.9225	0.7733	0.9034	0.8509
t + 5j	0.8181	0.8991	0.7532	0.8959	0.8194
t + 6j	0.7877	0.8720	0.7347	0.8778	0.7903
t + 7j	0.7562	0.8445	0.7154	0.8683	0.7599
t + 8j	0.7228	0.8161	0.6938	0.8423	0.7226

dd3	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.5436	0.4721	0.5486	0.4482	0.5310
t – 7j	0.5939	0.5299	0.5992	0.4908	0.5784
t – 6j	0.6543	0.5924	0.6607	0.5451	0.6395
t – 5j	0.7291	0.6554	0.7411	0.6153	0.7009
t – 4j	0.7987	0.7172	0.8114	0.6663	0.7664
t – 3j	0.8536	0.7612	0.8625	0.7058	0.8294
t – 2j	0.8878	0.8040	0.8844	0.7508	0.8712
t – 1j	0.9166	0.8518	0.9012	0.8057	0.9080
t = 0	0.9462	0.8888	0.9287	0.8498	0.9448
t + 1j	0.9183	0.8982	0.8974	0.8645	0.9338
t + 2j	0.8824	0.9159	0.8613	0.8851	0.9003
t + 3j	0.8446	0.9034	0.8332	0.8904	0.8446
t + 4j	0.8085	0.8820	0.8073	0.8919	0.8051
t + 5j	0.7731	0.8667	0.7788	0.8951	0.7682
t + 6j	0.7344	0.8519	0.7485	0.8870	0.7219
t + 7j	0.6950	0.8186	0.7199	0.8621	0.6548
t + 8j	0.6572	0.7827	0.6863	0.8252	0.5970

(3) Corrélations-croisées obtenues pour MW (sécheresse-flash de mai dd1, juin dd2 et septembre dd3)

dd1	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.4684	0.3856	0.4553	0.3699	0.4121
t – 7j	0.5401	0.4331	0.5383	0.4189	0.4707
t – 6j	0.6252	0.4882	0.6356	0.4744	0.5441
t – 5j	0.7131	0.5548	0.7213	0.5364	0.6257
t – 4j	0.7959	0.6195	0.7930	0.5957	0.7079
t – 3j	0.8612	0.6741	0.8366	0.6412	0.7864
t – 2j	0.9125	0.7249	0.8646	0.6847	0.8619
t – 1j	0.9464	0.7778	0.8747	0.7317	0.9210
t = 0	0.9654	0.8340	0.8754	0.7862	0.9547
t + 1j	0.9436	0.8596	0.8289	0.8114	0.9477
t + 2j	0.9124	0.8779	0.7829	0.8290	0.9246
t + 3j	0.8728	0.8882	0.7384	0.8431	0.8934
t + 4j	0.8262	0.9041	0.6820	0.8504	0.8482
t + 5j	0.7745	0.9025	0.6323	0.8553	0.7916
t + 6j	0.7196	0.8724	0.5999	0.8468	0.7362
t + 7j	0.6651	0.8201	0.5874	0.8227	0.6722
t + 8j	0.6123	0.7732	0.5743	0.7944	0.6085

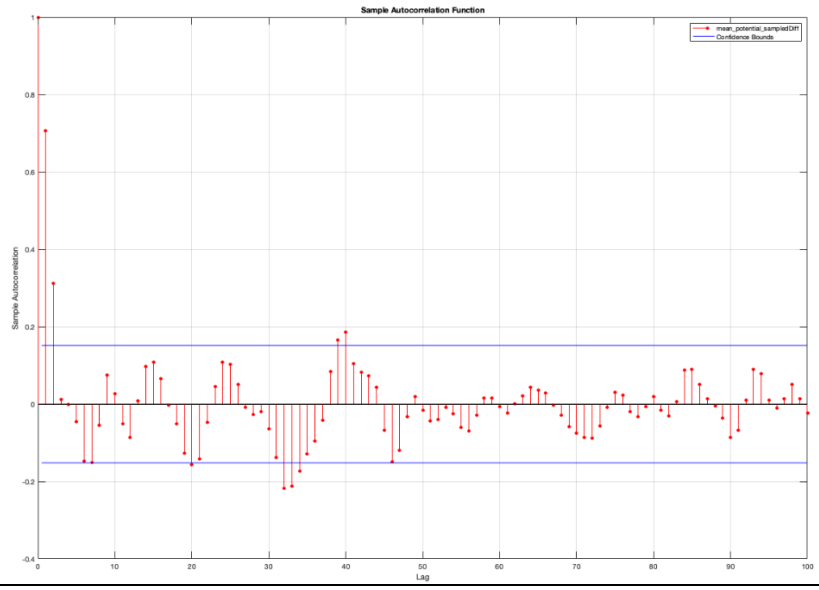
dd2	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.6031	0.5062	0.6084	0.4839	0.5897
t – 7j	0.6578	0.5574	0.6616	0.5276	0.6457
t – 6j	0.7102	0.6074	0.7121	0.5749	0.6974
t – 5j	0.7601	0.6593	0.7580	0.6241	0.7474
t – 4j	0.8100	0.7172	0.7996	0.6797	0.7966
t – 3j	0.8587	0.7743	0.8403	0.7340	0.8439
t – 2j	0.9003	0.8231	0.8759	0.7797	0.8873
t – 1j	0.9335	0.8712	0.8998	0.8238	0.9238
t = 0	0.9584	0.9136	0.9178	0.8682	0.9518
t + 1j	0.9405	0.9256	0.8882	0.8833	0.9360
t + 2j	0.9202	0.9348	0.8577	0.8965	0.9160
t + 3j	0.8980	0.9426	0.8255	0.9014	0.8929
t + 4j	0.8743	0.9362	0.8007	0.9109	0.8716
t + 5j	0.8496	0.9201	0.7808	0.9049	0.8482
t + 6j	0.8220	0.8995	0.7606	0.8915	0.8239
t + 7j	0.7891	0.8762	0.7357	0.8873	0.7922
t + 8j	0.7508	0.8478	0.7068	0.8646	0.7487

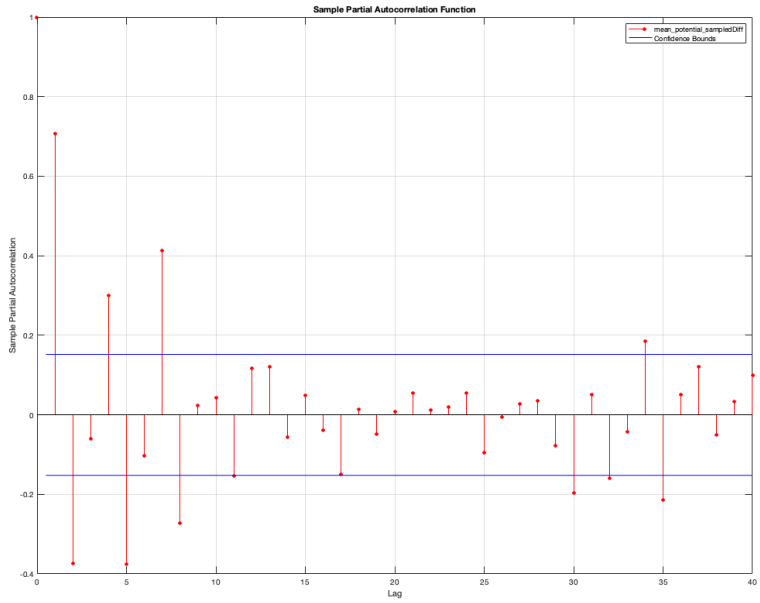
dd3	Tsoil	VPD	RelH	SRad	Tair
t – 8j	0.5574	0.4843	0.5619	0.4625	0.5454
t – 7j	0.6053	0.5436	0.6083	0.5036	0.5919
t – 6j	0.6619	0.6061	0.6654	0.5573	0.6495
t – 5j	0.7324	0.6658	0.7415	0.6258	0.7046
t – 4j	0.7997	0.7262	0.8098	0.6751	0.7687
t – 3j	0.8536	0.7657	0.8622	0.7143	0.8307
t – 2j	0.8905	0.8105	0.8895	0.7642	0.8729
t – 1j	0.9244	0.8583	0.9146	0.8216	0.9106
t = 0	0.9571	0.8927	0.9462	0.8632	0.9488
t + 1j	0.9283	0.9004	0.9114	0.8729	0.9368
t + 2j	0.8946	0.9185	0.8733	0.8871	0.9070
t + 3j	0.8578	0.9055	0.8423	0.8876	0.8551
t + 4j	0.8216	0.8875	0.8133	0.8897	0.8174
t + 5j	0.7851	0.8730	0.7835	0.8932	0.7797
t + 6j	0.7466	0.8579	0.7535	0.8846	0.7313
t + 7j	0.7091	0.8245	0.7260	0.8625	0.6647
t + 8j	0.6746	0.7911	0.6949	0.8315	0.6117

(2) Confirmation du type de processus

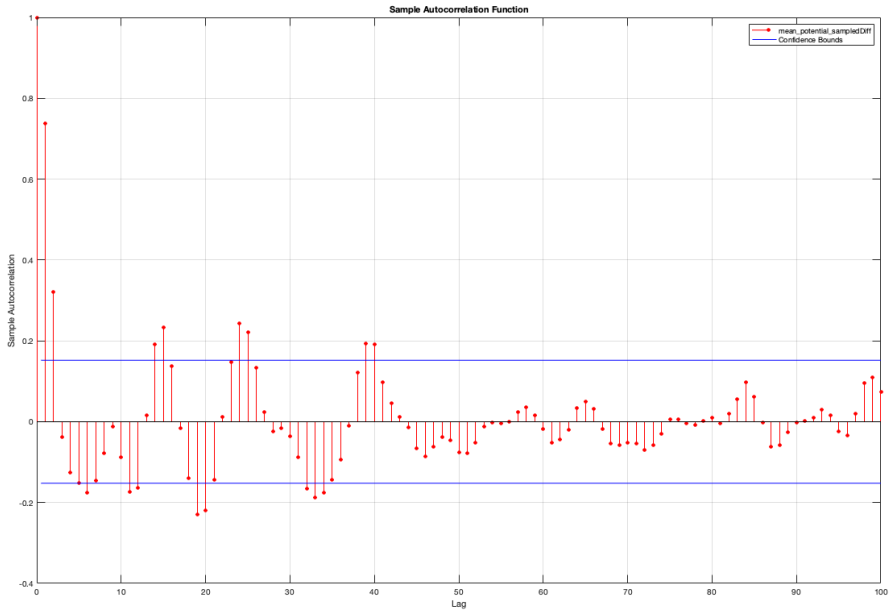
Pour ce faire, nous allons tracer les corrélogrammes ACF et PACF de chaque peuplement étudié, pour l'année 2020.

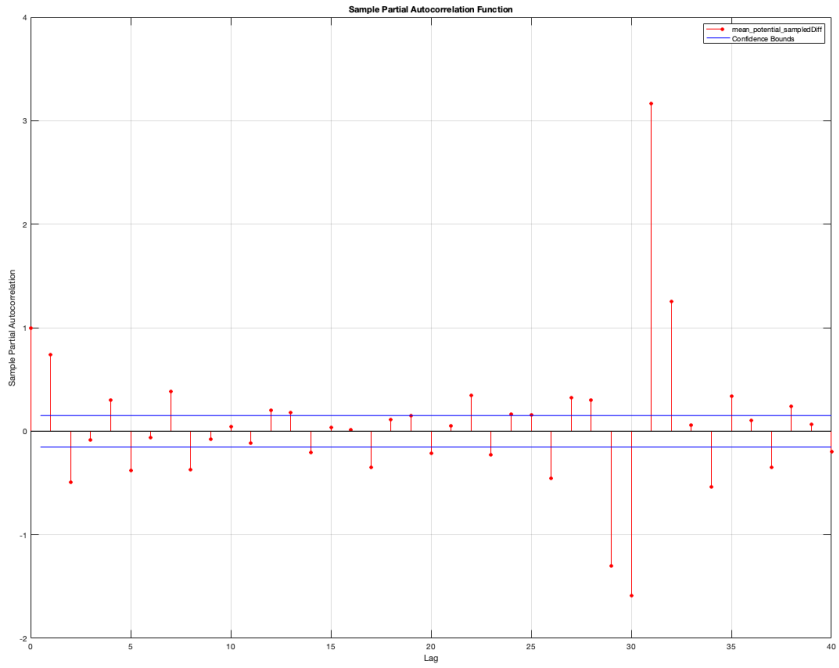
(1) ACF et PACF pour le potentiel hydrique du sol du peuplement MW, différencié 1 fois



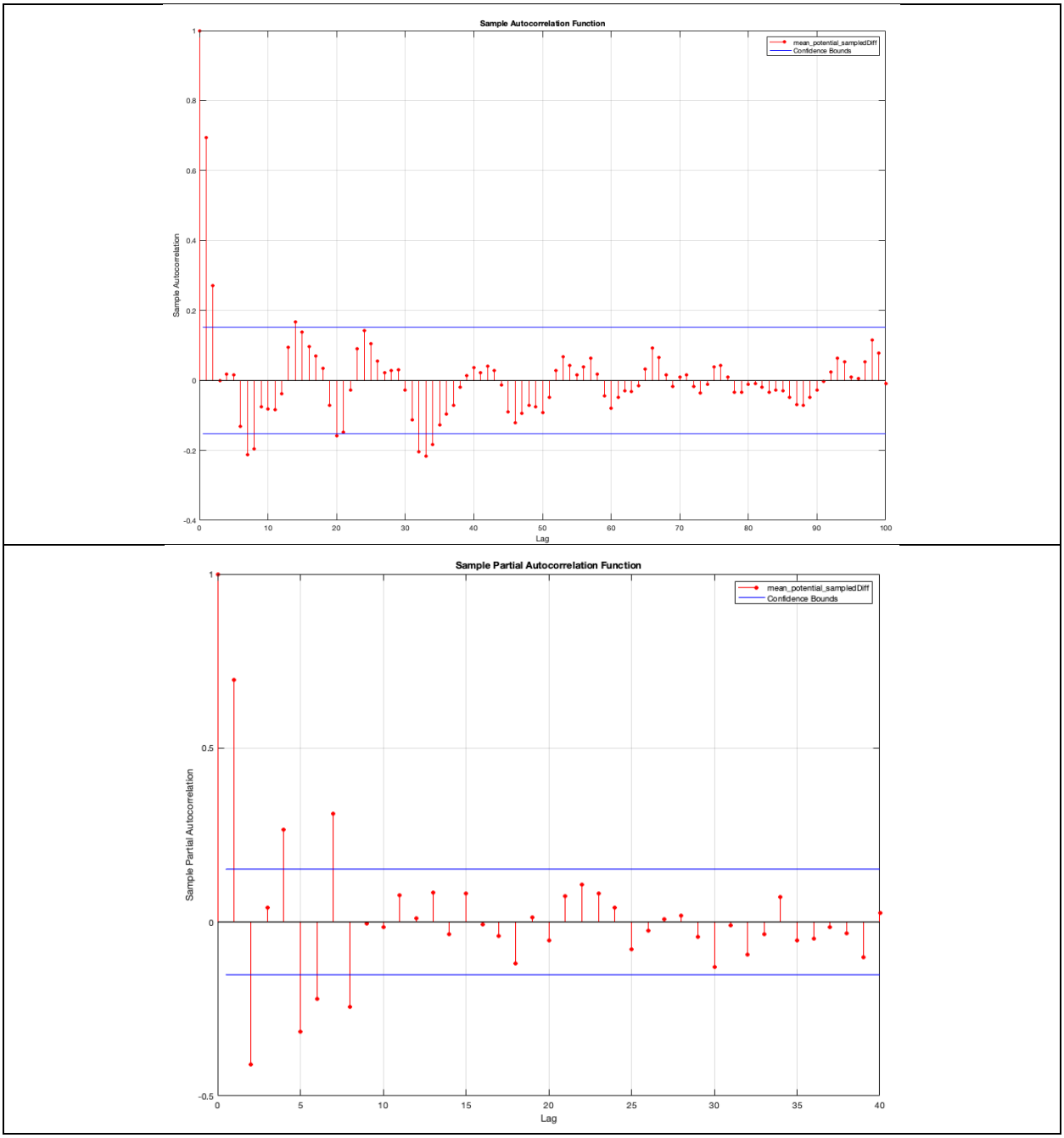


(2) ACF et PACF pour le potentiel hydrique du sol du peuplement HW, différencié 1 fois





(3) ACF et PACF pour le potentiel hydrique du sol du peuplement HB, différencié 1 fois

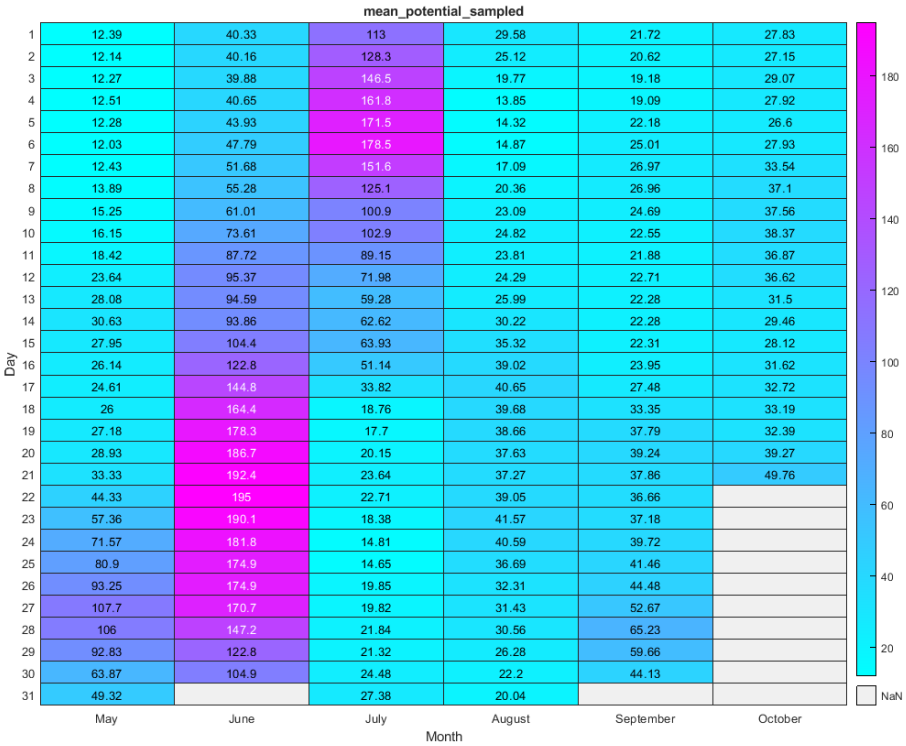


L'analyse des corrélogrammes donne des pistes pour des valeurs initiales de p et q pour le modèle ARMA(p,q). Soulignons ici que le fait que le graphique ACF et le graphique PACF ne montrent pas de *cutoff* clair, nous conforte dans le choix d'un modèle de type ARMA. En effet, l'autocorrélation se maintient au cours du temps, justifiant la pertinence d'utiliser ce type d'approche.

Nous avons différencié les trois séries temporelles pour les rendre stationnaires.

Les données ont été collectées entre mai et octobre 2020. Une saisonnalité pourrait être présente dans les données, avec un potentiel hydrique en mai qui s'apparente à celui mesuré en septembre, laissant suggérer une saisonnalité à prendre en compte. Néanmoins, cette saisonnalité nous semble trop longue pour que les modèles potentiels puissent l'exprimer.

Pour nous assurer de l'absence d'une trop forte saisonnalité dans les données du potentiel hydrique sur la période d'étude, nous pouvons tracer ses valeurs sous forme de cartes thermiques (*heatmap)* (Figure 47). Si un patron est présent, il sera visible dans la représentation colorée (gradient de couleur allant du bleu clair au rose foncé, des valeurs basses du potentiel aux valeurs élevées).



Day	May	June	July	August	September	October
1	12.39	40.33	113	29.58	21.72	27.83
2	12.14	40.16	128.3	25.12	20.62	27.15
3	12.27	39.88	146.5	19.77	19.18	29.07
4	12.51	40.65	161.8	13.85	19.09	27.92
5	12.28	43.93	171.5	14.32	22.18	26.6
6	12.03	47.79	178.5	14.87	25.01	27.93
7	12.43	51.68	151.6	17.09	26.97	33.54
8	13.89	55.28	125.1	20.36	26.96	37.1
9	15.25	61.01	100.9	23.09	24.69	37.56
10	16.15	73.61	102.9	24.82	22.55	38.37
11	18.42	87.72	89.15	23.81	21.88	36.87
12	23.64	95.37	71.98	24.29	22.71	36.62
13	28.08	94.59	59.28	25.99	22.28	31.5
14	30.63	93.86	62.62	30.22	22.28	29.46
15	27.95	104.4	63.93	35.32	22.31	28.12
16	26.14	122.8	51.14	39.02	23.95	31.62
17	24.61	144.8	33.82	40.65	27.48	32.72
18	26	164.4	18.76	39.68	33.35	33.19
19	27.18	178.3	17.7	38.66	37.79	32.39
20	28.93	186.7	20.15	37.63	39.24	39.27
21	33.33	192.4	23.64	37.27	37.86	49.76
22	44.33	195	22.71	39.05	36.66	
23	57.36	190.1	18.38	41.57	37.18	
24	71.57	181.8	14.81	40.59	39.72	
25	80.9	174.9	14.65	36.69	41.46	
26	93.25	174.9	19.85	32.31	44.48	
27	107.7	170.7	19.82	31.43	52.67	
28	106	147.2	21.84	30.56	65.23	
29	92.83	122.8	21.32	26.28	59.66	
30	63.87	104.9	24.48	22.2	44.13	
31	49.32		27.38	20.04		

Figure 47 : Représentation du potentiel hydrique du sol sous forme de carte thermique pour le peuplement MW.

Nous pouvons observer qu'il n'y a pas de patron marqué de saisonnalité dans les données du potentiel hydrique durant la période étudiée (de mai à octobre).

En considérant les graphiques ACF, nous opterons pour une saisonnalité courte, de 10 jours maximum, qui pourra être testée avec une moyenne mobile de fenêtre maximale de cet ordre de grandeur.

Cependant, pour les trois peuplements, nous constatons une forte diminution de la fonction d'autocorrélation au lag 3. Ainsi (q+1) = 3, ce qui implique q est égal à 2.

Pour les valeurs de p, les trois graphiques de la fonction d'autocorrélation partielle (PACF) nous suggèrent également une forte diminution de l'autocorrélation partielle au lag 3. Ainsi (p+1) = 3, p est donc égal à 2.

Une nouvelle diminution est observée autour du lag 10, laissant aussi suggérer la nécessité d'explorer des valeurs de p plus élevées.

Les trois graphiques ACF et PACF nous suggèrent donc de tester les valeurs de p et q suivantes :

	MW	HW	HB
(PACF) p	2 7	2 9	2 8
(ACF) q	2 7	2 7	2 11

À la suite de cette étape, nous allons essayer de déterminer un modèle de type ARIMAX(p,d,q) (*AutoRegressive Integrated Moving Average with Explanatory variables*), un modèle autorégressif un peu plus complexe, qui prend en compte la différenciation de la série (d=1) et l'influence de facteurs exogènes.

Phase 5 : développement des modèles de type ARIMAX(p,d,q)

(3) Détermination du modèle ARIMAX(p,d,q) *et* (4) évaluation de la qualité de l'ajustement

Pourquoi privilégier l'approche ARIMAX ?

Elle permet une analyse prédictive lorsque les séries temporelles sont non stationnaires et multivariées, exhibant également n'importe quelle *structure* (*pattern*) dans leurs données (tendance, saisonnalité, cyclicité).

La méthode ARIMA, que nous avons définie précédemment, ne s'applique qu'aux séries stationnaires univariées. Dans l'approche ARIMAX *il est possible d'intégrer des variables exogènes au modèle.*

Si nous souhaitons intégrer, dans une première approche, les variables comme la température de l'air, le déficit de pression de vapeur (VPD), les précipitations et la température moyenne du sol, qui, nous l'avons montré, exercent une influence sur l'évolution temporelle du potentiel hydrique, un modèle ARIMAX semble plus adapté.

Soulignons une des limites du modèle ARIMAX : il ne considère qu'une relation symétrique entre les variables, c'est-à-dire que la variable à prédire est considérée comme étant influencée par les variables exogènes. Or l'inverse peut s'avérer également. Pour considérer toutes les variables comme endogènes, il faudrait se tourner vers des modèles de type VAR (*Vector AutoRegression model*).[25]

Nous pouvons néanmoins considérer que ce sont les variables météorologiques qui influent sur le potentiel hydrique du sol et que l'inverse n'aurait pas d'explications physiologiques au niveau des mécanismes qui ont lieu à l'échelle de l'arbre.

Soit le modèle ARIMAX(p,d,q) avec :

p l'ordre de la régression AR(p),

d le degré de différenciation,

q l'ordre de la moyenne mobile MA(q).

[25] Hyndman, R.J. and Athanasopoulos, G. (2018). *Forecasting hierarchical and grouped time series* In Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on January 27th, 2023

Nous avons montré que le degré de différenciation devra être égal à 1 (d=1), pour s'assurer de la stationnarité des séries temporelles.

L'analyse des corrélogrammes des séries différenciées nous a donné des pistes pour des valeurs initiales de p et q.

Démarche effectuée dans Matlab

(1) Création d'un *template* correspondant au type de modèle choisi (fonction *arima* de Matlab)

Les valeurs de p, d et q sont choisies.

C'est la structure mathématique du modèle, dont les coefficients seront ensuite optimisés.

(2) Estimation du modèle choisi (fonction *estimate* de Matlab)

À cette étape, pour un modèle ARIMAX, il faut préciser les variables exogènes sélectionnées.

Il faut également définir un intervalle de données de pré-échantillonnage (Y0 pre-sample).

Cet intervalle doit contenir au minimum (p+d) valeurs. En effet, pour un modèle autorégressif, il faudra considérer au minimum les p premières valeurs de la série temporelle. Si une différenciation est faite, il faut alors ajouter un point par ordre de différenciation effectué.

Il faut ensuite définir l'intervalle de données qui permettra d'estimer le modèle.

Celui-ci peut commencer au point temporel (p+d+1).

(3) Prédictions à partir du modèle estimé (fonction *forecast* de Matlab)

Il faudra ici définir également un autre intervalle de données de pré-échantillonnage Y0 et un intervalle de données sur lequel la prédiction se fera (l'horizon de la prédiction).

Il faudra également préciser l'intervalle de prédiction des variables exogènes. Celui-ci est égal à l'horizon de la prédiction (il ne peut pas être plus grand, et par défaut, il est pris égal à l'horizon).

Plusieurs stratégies sont possibles pour déployer les trois étapes.

Nous avons implémenté une fonction pour combiner l'étape 1 et 2.

À l'issue de l'étape 2, il est recommandé, comme souligné précédemment, de tracer le graphique ACF des résidus et de vérifier la normalité de leur distribution. Les coefficients AIC et BIC sont également comparés, si d'autres modèles sont testés. Enfin, nous pouvons tracer le potentiel hydrique ajusté sur le modèle choisi, tel que : (résidu = valeur mesurée – valeur ajustée)

Pour mieux comprendre les différents intervalles de données qui doivent être définis pour chaque étape, un schéma est proposé[26] (Figure 48). Les trois périodes expliquées sont représentées graphiquement : pré-échantillonnage (*presample*), estimation, prédiction (*forecast*).

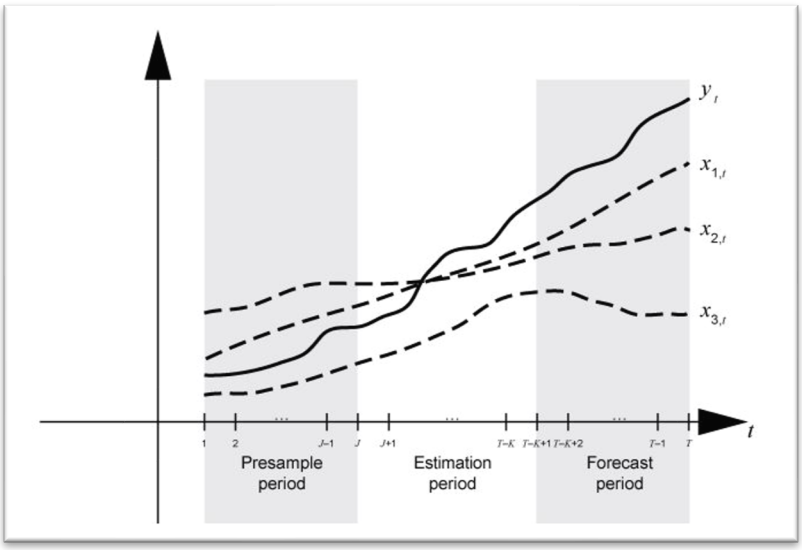


Figure 48 : Trois types d'intervalles sont considérés pour l'estimation du modèle : pré-échantillonnage, estimation et prédiction.

Pour l'étape de prédiction, nous l'avons souligné, il faut également préciser un intervalle de données de pré-échantillonnage, comme illustré sur la (Figure 49).

[26] https://www.mathworks.com/help/econ/time-base-partitions-for-estimation.html

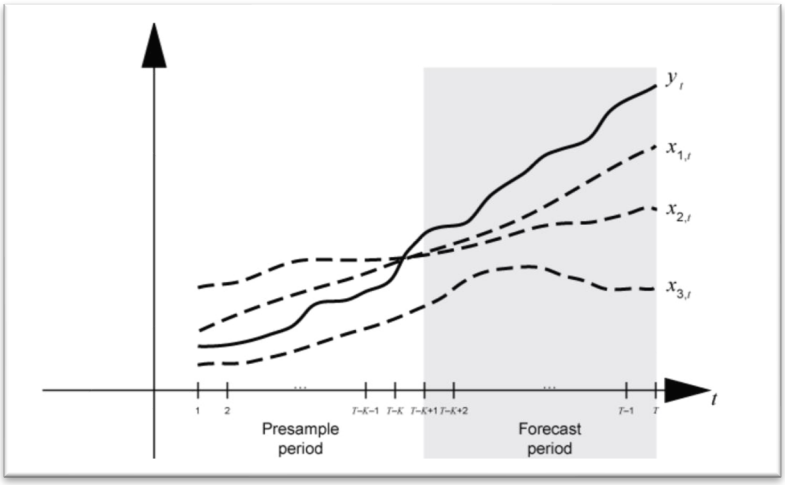


Figure 49 : Intervalle de pré-échantillonnage pour l'étape de prédiction.

Pour valider le modèle, une stratégie conseillée est de séparer au début les données d'intérêt en ces trois intervalles, et de valider le modèle en prédisant des données existantes, avec pour intervalle de pré-échantillonnage, les intervalles qui ont servi à l'estimation du modèle (intervalle de pré-échantillonnage de l'estimation et l'intervalle de l'estimation)[27]. Souvent, cela va correspondre aux données au temps [1 :p+d] pour le pré-échantillonnage de l'estimation puis [p+d+1 : T-K] pour l'estimation). L'horizon de prédiction sera alors [T-K+1:T].

En considérant les données sous forme de vecteurs, cela donne alors les données de pré-échantillonnage Y0, d'estimation Y et l'horizon de prédiction (Figure 50). Les variables exogènes X_i sont également définies sur l'intervalle d'estimation du modèle.

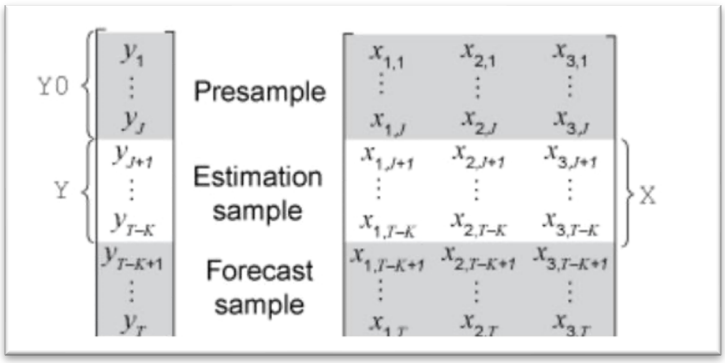


Figure 50 : Intervalles représentés sous forme de vecteurs.

[27] Se référer au manuel de Matlab, chapitre : *Time base partitions for estimation*

Obtention de trois modèles ARIMAX(p,d,q)

Les variables exogènes considérées sont :

- Température moyenne du sol
- Température de l'air
- Le déficit de pression de vapeur
- L'humidité relative
- L'irradiance solaire
- Les précipitations

Soulignons que nous avons mis en évidence que l'influence de ces variables sur le potentiel hydrique du sol s'étend parfois jusqu'à 6 jours. Il est donc impératif de tester des valeurs de q allant jusqu'à 6 jours.

Différentes valeurs des paramètres p et q sont alors testées, et la qualité des ajustements est obtenue et comparée afin de sélectionner les meilleurs paramètres (Tableau 15).

Une fonction a été implémentée pour tester de façon systématique différentes valeurs de p et q, en estimant un grand de nombres de modèles ARIMAX(p,1,q).

Les critères d'information *AIC* et *BIC* sont reportés dans le Tableau 15 pour les valeurs testées.

Pour chaque modèle, nous optons pour une distribution gaussienne des innovations et le terme constant dans la description du modèle est pris égal à zéro.

Remarque : quelle différence entre les termes *innovation* et *résidus* ?

En traitement du signal, et à l'origine en traitement de l'information, le terme *innovation* se définit comme la différence entre la valeur observée d'une variable au temps t et sa valeur qui a été prédite en s'appuyant sur l'information disponible *avant* ce temps t. Pour le *résidu*, la valeur au temps t est également considérée.

Tableau 15 : Tests réalisés en faisant varier les valeurs des paramètres p et q.

	MW **AIC** **BIC**	**HW** **AIC** **BIC**	**HB** **AIC** **BIC**	**Commentaires**
(1,0,0)	1126 1136			Un modèle AR(1) simple est testé et ne permet pas un bon ajustement du potentiel hydrique du sol, pour les 3 peuplements
(0,0,1)	1657 1666			Même constant avec un modèle purement MA, d'ordre 1
(1,1,1)	1042 1052			En permettant la différenciation d'ordre 1 de la série temporelle (d=1), nous avons un modèle ARIMA(1,1,1). Le coefficient AIC diminue alors.
(12,1,1)	941 993	889 941		
(12,1,2)	941 997	**887** 943		
(12,1,3)	940 998			
(12,1,4)	**932** 993	890 952	*773* 835	
(12,1,5)	938 1003			
(7,1,2)	960 1000			
(7,1,3)	962 1005			
(7,1,3)	961 1007			
(8,1,4)		907 957		
(9,1,4)		902 955	**779** 832	Pour HB, davantage de coefficients d'ajustement du modèle sont statistiquement significatifs, en comparaison avec (10,1,4) et (12,1,4).
(10,1,4)		894 950	*778* 834	
(11,1,4)		893 952		

Résultats des tests réalisés sur les trois peuplements

Pour les trois peuplements étudiés, trois modèles de type ARIMAX(p,d,q) ont été entrainés sur leurs données de 2020. Le terme constant dans le modèle a été pris égal à zéro. Une distribution gaussienne des résidus a été choisie. Une distribution de type t de Student, mieux adaptée à des données ayant un kurtosis élevé (données plus extrêmes) a été testée, car intuitivement plus adaptée à nos données, mais l'erreur associée à chaque point ajusté devenait très élevée, au cours du temps.

Trois variables exogènes ont été introduites initialement dans le modèle : T_{air}, VPD, T_{sol}

Nous avons également testé plusieurs modèles en ajoutant comme variable exogène, les précipitations, mais le coefficient d'ajustement associé au VPD ne devenait plus significatif dans ce cas-ci.

Étant donnés les résultats obtenus lors de l'analyse des variables prédictives du potentiel hydrique, nous avons opté pour les trois variables qui semblaient avoir le plus d'influence sur ce dernier, en ajoutant néanmoins les précipitations. Cette variable ne semble pas améliorer la prédiction du potentiel, étant données les imprécisions liées aux données générées par BioSIM, mais dans l'hypothèse de l'obtention de futures valeurs plus précises, elle a été intégrée dans les modèles.

Ainsi, l'irradiance solaire et l'humidité relative n'ont pas été considérées dans les modèles ARIMAX sélectionnés.

Les résultats obtenus sont les suivants :

Pour le peuplement MW : ARIMAX(12,1,4)

Pour le peuplement HW : ARIMAX(12,1,2)

Pour le peuplement HB : ARIMAX(9,1,4)

Après les premiers tests de prédictions, nous affinons les modèles pour MW et HW.

Ainsi, nous allons utiliser les trois modèles suivants pour les tests prédictifs (phase 5 de la méthodologie) :

Pour le peuplement MW : ARIMAX(14,1,4)

Pour le peuplement HW : ARIMAX(12,1,4)

Pour le peuplement HB : ARIMAX(9,1,4)

Avec quatre variables exogènes considérées : Tair, VPD, Tsol et Précipitations

Pour le peuplement (MW), le modèle ARIMAX(14,1,4) est défini par l'équation suivante :

$$(1 - \phi_1 L - \cdots - \phi_{14} L^{14})(1 - L) y_t = \beta_1 X_1 + \cdots + \beta_4 X_4 + (1 + \theta_1 L + \cdots + \theta_4 L^4)\varepsilon_t + c$$

Les coefficients ϕ_i sont les coefficients autorégressifs (AR) du polynôme L (défini par l'opérateur de rétro-propagation B), les coefficients β_j sont les coefficients des variables exogènes X_j et les coefficients θ_l sont les coefficients du polynôme associé à la moyenne mobile, et ε_t représente une série de variables aléatoires (de distribution gaussienne) de moyenne égale à 0 et de variance σ^2. Un terme constant c peut être ajouté, mais il est pris égal à zéro.

Remarque : une variable aléatoire gaussienne a une fonction de densité de probabilité (*Probability Density Function*) qui s'écrit sous la forme générale suivante :

$f_x = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-m)^2}{2\sigma^2}}$, avec m la moyenne et σ l'écart-type (*standard deviation*).

On a ainsi $\varepsilon_t = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{t^2}{2\sigma^2}}$

Pour rappel : La fonction *arima()* de Matlab permet d'obtenir l'équation associée au modèle : un objet de type arima est obtenu en sortie de la fonction. À partir de cet objet créé, la fonction *estimate()* ajuste alors tous les coefficients et la variance σ^2 sur les données appelées *données d'entraînement*. Pour initialiser le modèle, la fonction estimate() nécessite comme paramètre d'entrée l'intervalle de données de pré-échantillonnage, évoqué précédemment.

Après avoir appliqué la fonction estimate() au modèle ARIMAX(14,1,4) sur les données de potentiel hydrique du sol mesurées en 2020 pour le peuplement MW, nous obtenons les coefficients résumés dans le tableau ci-dessous.

Paramètres	Valeur obtenue
AR{1}	0.66424
AR{2}	-0.20053
AR{3}	-0.87331
AR{4}	0.83831
AR{5}	-8.2728e-05
AR{6}	-0.85164
AR{7}	0.75428
AR{8}	-0.15047
AR{9}	-0.37761
AR{10}	0.37946
AR{11}	-0.083551
AR{12}	-0.19184
AR{13}	0.14467
AR{14}	-0.052685
MA{1}	0.50724
MA{2}	0.62986
MA{3}	0.60004
MA{4}	0.42324
Beta(Prcp)	-0.19383
Beta(Tair)	0.51196
Beta(VPD)	2.5609
Beta(Tsoil)	-0.64583
Variance (σ^2)	14.0025
AIC	**916.8607**
BIC	**987.4455**

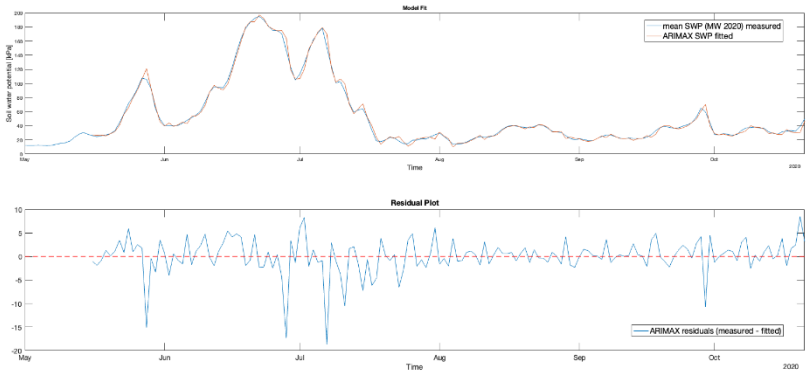
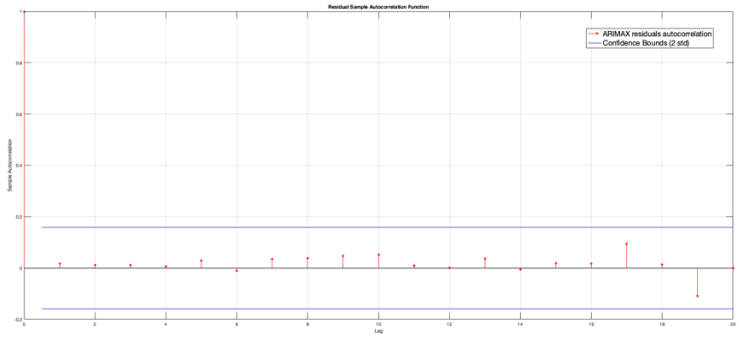


Figure 51 : Modèle ARIMAX(14,1,4).

Sur la Figure 51 associée au tableau la courbe en orange représente le potentiel hydrique du sol ajusté à partir du modèle ARIMAX(14,1,4), la courbe en bleu représente le potentiel hydrique du sol moyen mesuré en 2020, pour le peuplement MW.

Il est possible d'analyser les résidus obtenus à chaque point temporel. Ces résidus ne doivent pas être corrélés dans le temps, nous pouvons nous en assurer en traçant les coefficients d'autocorrélation (Figure 52). Leur valeur reste toujours en dessous des intervalles de confiance (en bleu sur la Figure 52).

Nous pouvons également tracer l'histogramme de densité des résidus et s'assurer que leur répartition tend vers une distribution normale (Figure 52).



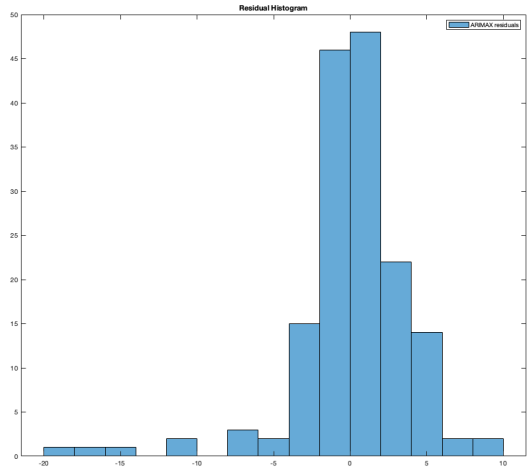
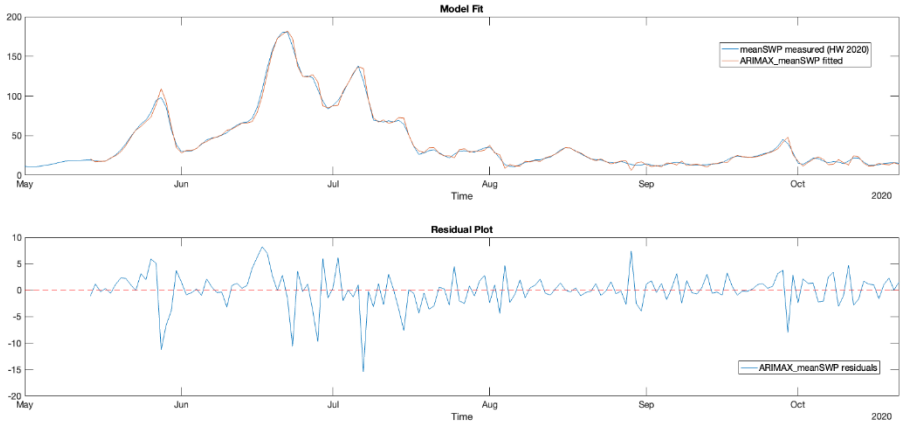


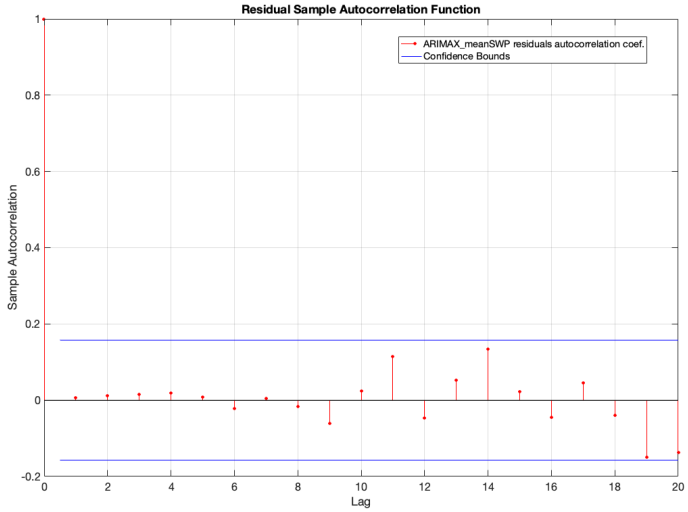
Figure 52 : Coefficients d'autocorrélation associés aux résidus du modèle ARIMAX(14,1,4) et leur histogramme de densité.

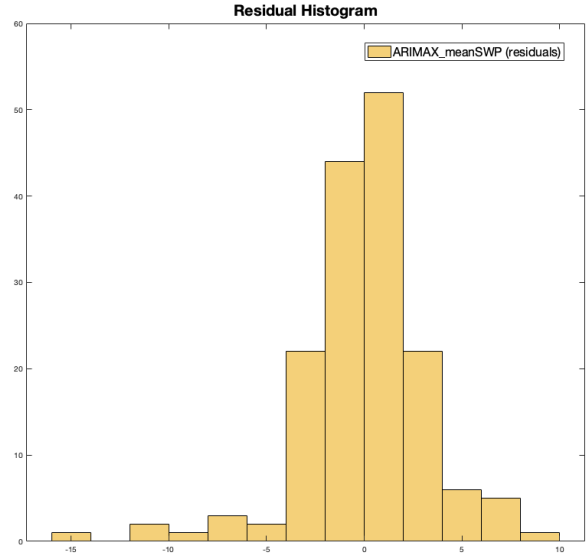
Pour le peuplement HW, le modèle ARIMAX(12,1,4) est défini par l'équation suivante :

$$(1-\phi_1 L-\cdots-\phi_{12}L^{12})(1-L)y_t=\beta_1 X_1+\cdots+\beta_4 X_4+(1+\theta_1 L+\cdots+\theta_4 L^4)\varepsilon_t$$

Paramètres	Valeurs
AR{1}	0.59455
AR{2}	-0.037964
AR{3}	-0.90501
AR{4}	0.57238
AR{5}	0.22763
AR{6}	-0.82912
AR{7}	0.38574
AR{8}	0.15496
AR{9}	-0.38
AR{10}	0.097078
AR{11}	0.02902
AR{12}	-0.20339
MA{1}	0.5492
MA{2}	0.50375
MA{3}	0.52083
MA{4}	0.39328
Beta(Prcp)	-0.18578
Beta(Tair)	0.20958
Beta(VPD)	3.4066
Beta(meanTsoil)	-0.31796
Variance	10.3719
AIC	**875.4933**
BIC	**940.2028**



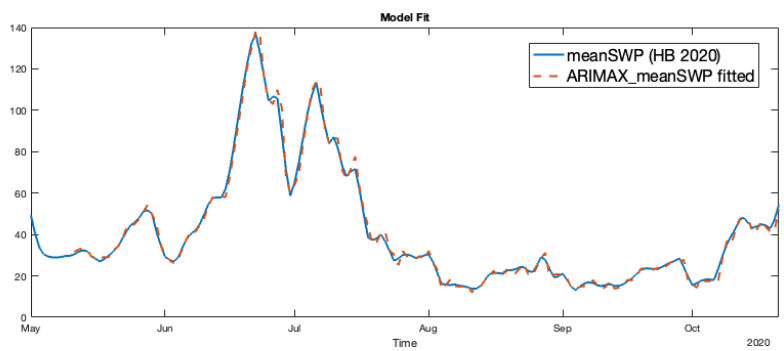


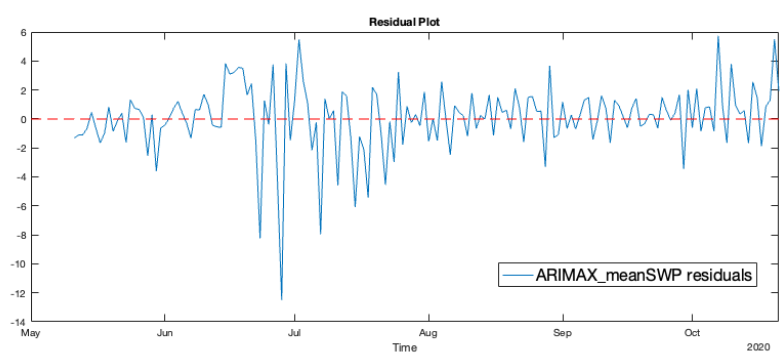


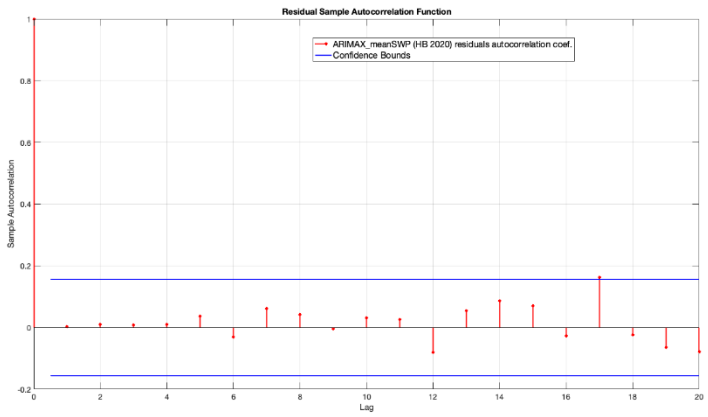
Enfin, pour le peuplement HB, le modèle ARIMAX(9,1,4) est défini par l'équation suivante :

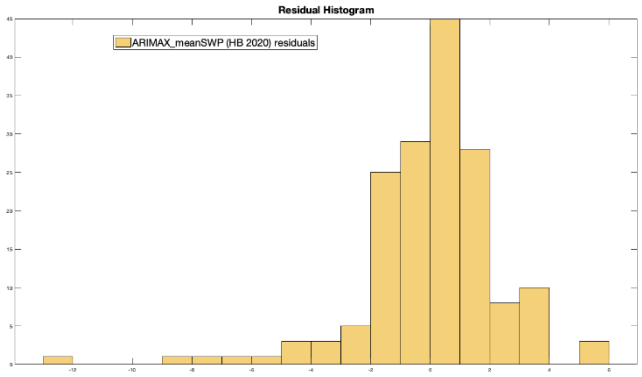
$$(1-\phi_1 L-\cdots-\phi_9 L^9)(1-L)y_t=\beta_1 X_1+\cdots+\beta_4 X_4+(1+\theta_1 L+\cdots+\theta_4 L^4)\varepsilon_t$$

Paramètres	Valeur
AR{1}	0.52817
AR{2}	-0.06214
AR{3}	-0.65323
AR{4}	0.50405
AR{5}	0.15664
AR{6}	-0.57591
AR{7}	0.3031
AR{8}	-0.036504
AR{9}	-0.13049
MA{1}	0.60046
MA{2}	0.53461
MA{3}	0.33325
MA{4}	0.09728
Beta(Prcp)	-0.069411
Beta(Tair)	0.070763
Beta(VPD)	5.01
Beta(meanTsoil)	-0.2354
Variance	5.3712
AIC	**777.1037**
BIC	**832.9013**







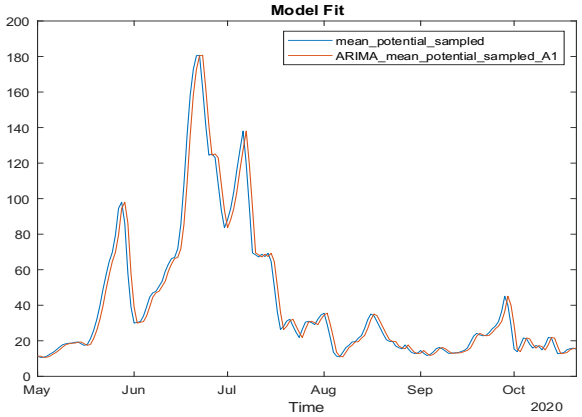


À titre informatif, nous pouvons comparer les trois modèles ARIMAX obtenus avec des modèles simples, d'ordre 1.

Ajustement obtenu (en orange) pour un modèle AR(1), pour le peuplement HW (potentiel mesuré, en bleu) :

$$(1 - \phi_1 L) y_t \;=\; \varepsilon_t$$

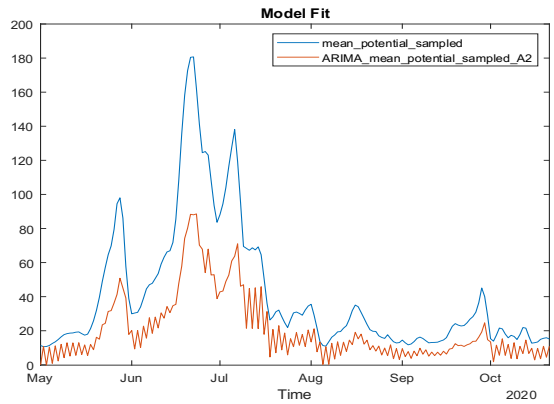
Paramètres	Valeur
Constant	0
AR{1}	1
Variance	100
AIC	**1198**
BIC	**1204**



Ajustement obtenu pour un modèle MA(1) pour HW :

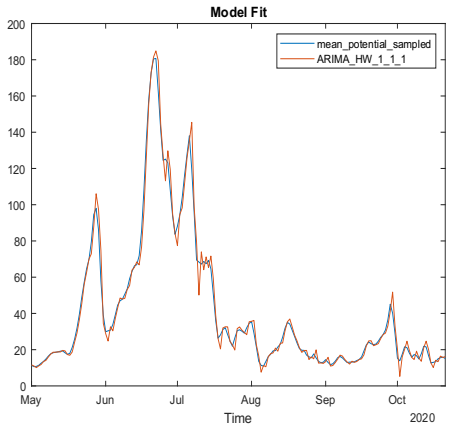
$$y_t \;=\; (1+\theta_1 L)\varepsilon_t$$

Paramètre	Valeur
Constant	0
MA{1}	0.95558
Variance	1173.5468
AIC	**1674**
BIC	**1681**



Exemple d'ajustement obtenu pour un modèle ARIMA(1,1,1) pour HW :

$$(1-\phi_1 L)(1-L)y_t \;=\; (1+\theta_1 L)\varepsilon_t$$



AIC = 1042 et BIC = 1052

En comparant deux modèles simples, AR(1) et MA(1), nous constatons qu'un modèle MA seul ne permet pas un bon ajustement du potentiel hydrique du sol. Un modèle autorégressif d'ordre 1 permet l'obtention d'un meilleur modèle, mais un décalage temporel est à compenser en augmentant l'ordre p.

La différenciation d'ordre 1 permet également de diminuer le coefficient AIC.

(5) Validation du modèle sur des données test *et* (6) Prédiction sur un horizon futur

- Résultats obtenus pour le peuplement HB

Le modèle ARIMAX(9,1,4) a d'abord été entraîné avec les données de 2020. Sur la Figure 53 le potentiel hydrique prédit (en vert) suit les variations du potentiel mesuré (en bleu) au début de l'horizon, qui a été fixé à la même année (de mai à novembre). Cette étape permet de tester le modèle.

Durant environ 60 jours le potentiel prédit est capable d'être dans le même état que le potentiel mesuré, mais ensuite, les variations ne sont plus prédites et le potentiel se maintient dans un état *moyen* alors que les mesures se situent entre 0 et 40 kPa (état fixé à *faible*), même si la tendance à diminuer est reproduite. Notons que le fait que le processus soit non-stationnaire implique une augmentation au cours du temps des deux intervalles de confiance représentés en pointillés[28].

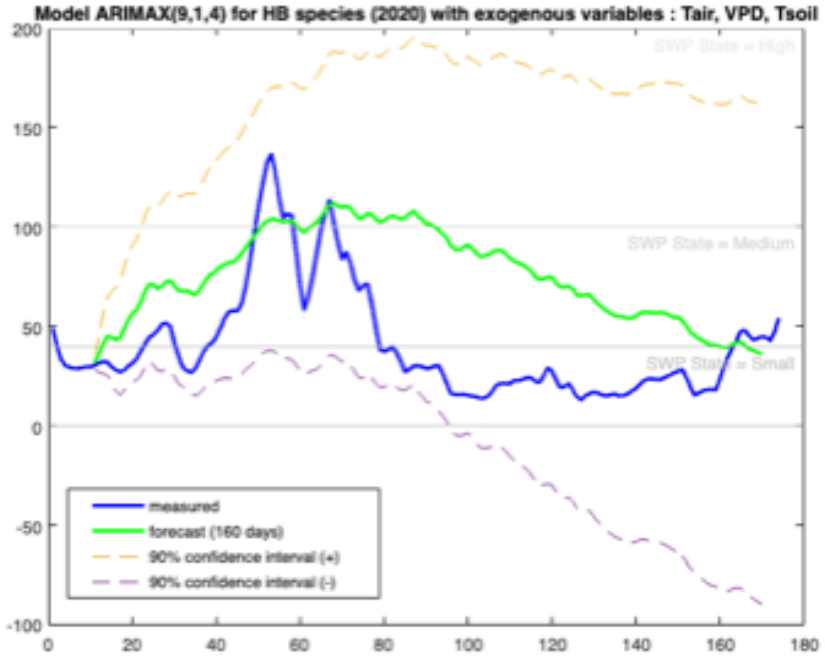


Figure 53 : Modèle ARIMAX(9,1,4) pour le peuplement HB (entraînement sur les données de 2020).

L'intervalle de validation est alors ensuite réduit à 60 jours, et deux modèles sont testés (

[28] Matlab, Forecast multiplicative ARIMA model (source)

Figure 54). Le premier contient trois variables exogènes : Tair, Tsol et VPD. Le second en contient quatre : les précipitations sont ajoutées.

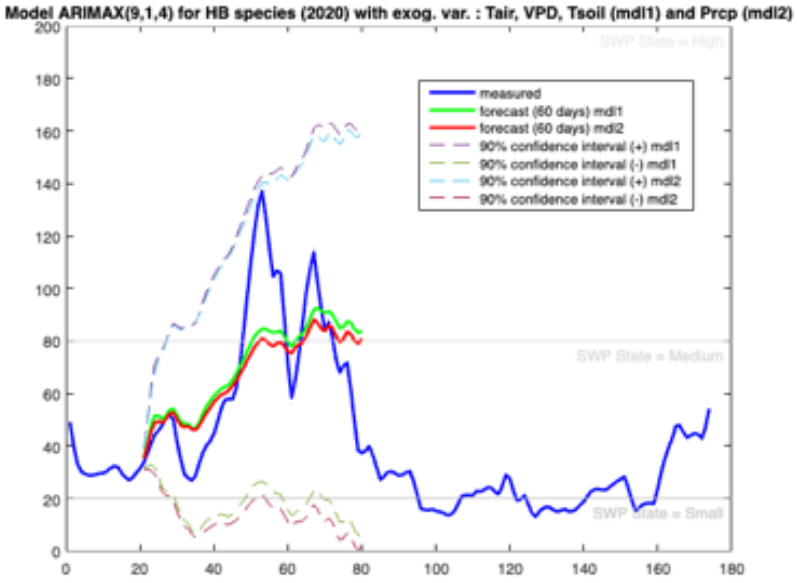


Figure 54 : Réduction de l'intervalle de prédiction à 60 jours.

Le potentiel hydrique prédit (en vert pour le modèle à 3 variables exogènes, en rouge pour le modèle à quatre variables exogènes) suit les variations du potentiel mesuré (en bleu). Dans les deux modèles, le potentiel prédit atteint les mêmes états (moyen puis élevé) que le potentiel mesuré. Les deux extrema locaux ne sont pas prédits, mais les modèles sont capables de prédire que le potentiel atteindra des valeurs supérieures à 80kPa. Il faut préciser que dans ce cas, les seuils fixés diffèrent de ceux de la Figure 53.

À la suite d'une première série de tests, nous avons implémenté une fonction qui permet d'appliquer une approche de prédiction par le biais de l'utilisation de fenêtres glissantes (*rolling windows*).

Dans une première approche, nous avons défini trois fenêtres glissantes de prédiction. Cela permet de diviser les données en trois intervalles, qui pourraient correspondre au printemps, à l'été et à l'automne (Figure 55). Ainsi, nous pouvons adapter les prédictions à des moments météorologiques similaires.

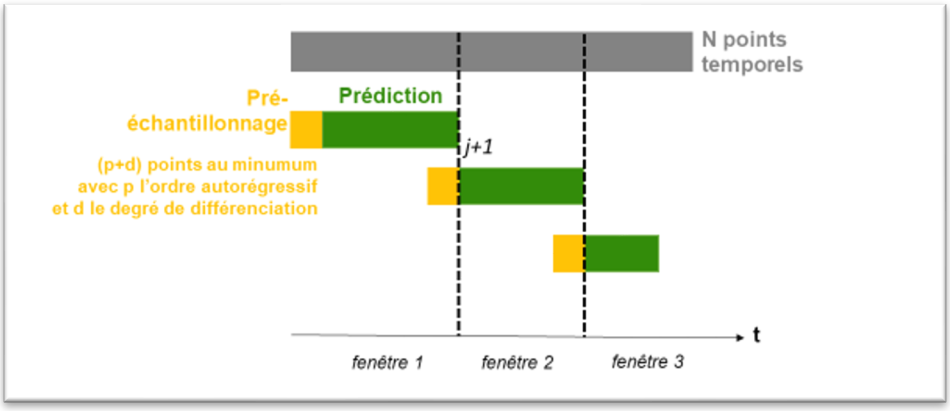


Figure 55 : Schéma représentant l'approche des fenêtres glissantes.

La fonction implémentée permet donc de définir trois fenêtres, chacune ayant leurs propres paramètres.

Trois fonctions ont été implémentées, car les coefficients autorégressifs (p) et ceux de la moyenne mobile (q) sont différents pour chaque peuplement. Ils auraient pu être intégrés aux paramètres d'entrée, mais le choix s'est porté sur trois fonctions différentes :

ARIMAX_HB_p9_d1_q4_4X(dataInputEstimation,preSampleEstimation_days,SampleEstimation);

ARIMAX_HW_p12_d1_q4_4X(dataInputEstimation,preSampleEstimation_days,SampleEstimation);

ARIMAX_MW_p14_d1_q4_4X(dataInputEstimation,preSampleEstimation_days,SampleEstimation);

La fonction ARIMAX_prediction_erreur() et la fonction ARIMAX_MW_p14_d1_q4_4X() sont détaillées à l'Annexe VII.

Stratégies de test – fenêtres glissantes

(A) Forecasting

À partir des données antérieures, nous allons prédire le potentiel hydrique du sol des années futures.

Pour chaque modèle estimé, la taille du pré-échantillonnage est de 40 jours.

Pour chaque prédiction, la taille des trois fenêtres de pré-échantillonnage est de 60 jours ou 40 jours.

La taille des fenêtres de prédiction peut varier.

Test	Année de l'estimation du modèle	Année de prédiction du potentiel
1	2017	*2017 (à proscrire)*
2		2018
3		2019
4		2020
5	2018	*2018 (à proscrire)*
6		2019
7		2020
8	2019	*2019 (à proscrire)*
9		2020
10	2020	*2020 (à proscrire)*

Dans cette approche, il n'est pas conseillé de prédire le potentiel la même année des données qui ont servi à entraîner le modèle.

(B) Backcasting

À partir des données futures, nous allons prédire le potentiel hydrique du sol des années antérieures.

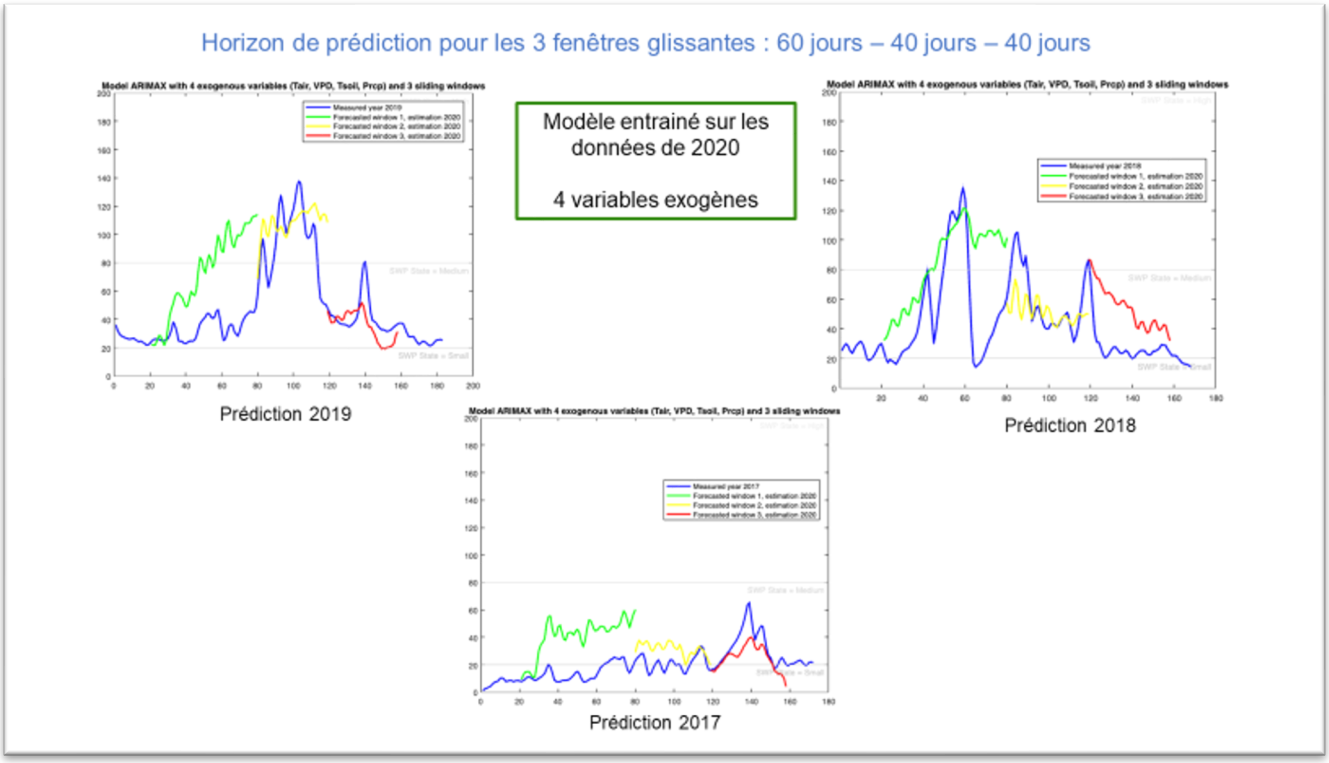
Test	Année de l'estimation du modèle	Année de prédiction du potentiel
11	2020	2019
12		2018
13		2017
14	2019	2018
15		2017
16	2018	2017

Pour le peuplement MW : test de 1 à 16

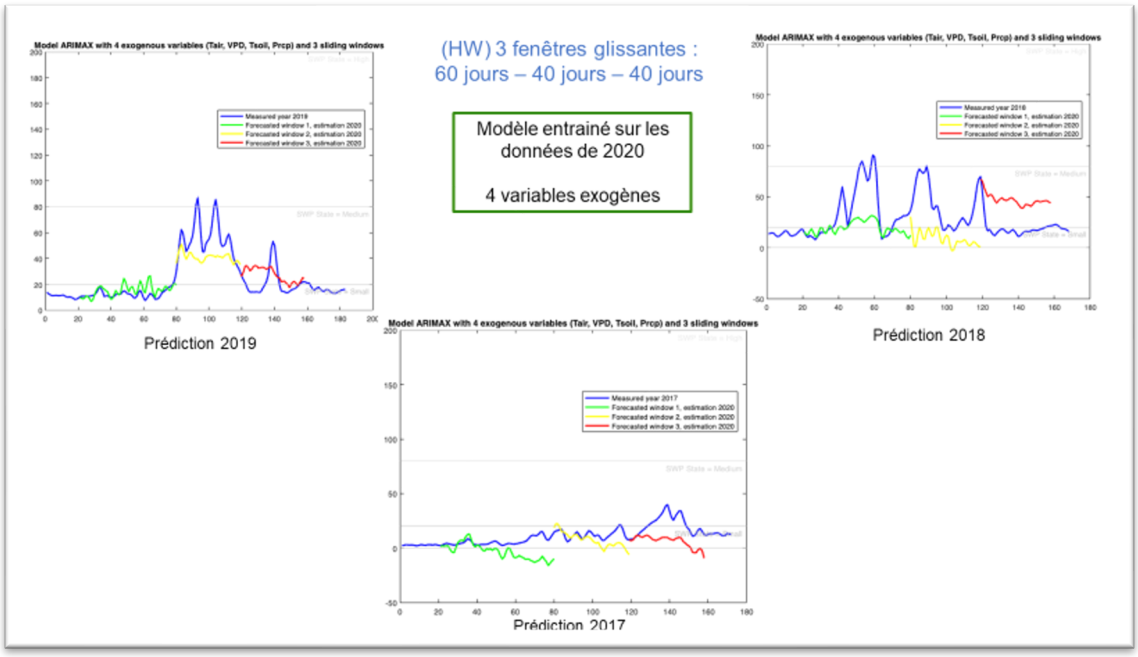
Pour le peuplement HW : test de 17 à 33

Pour le peuplement HB : test de 34 à 50

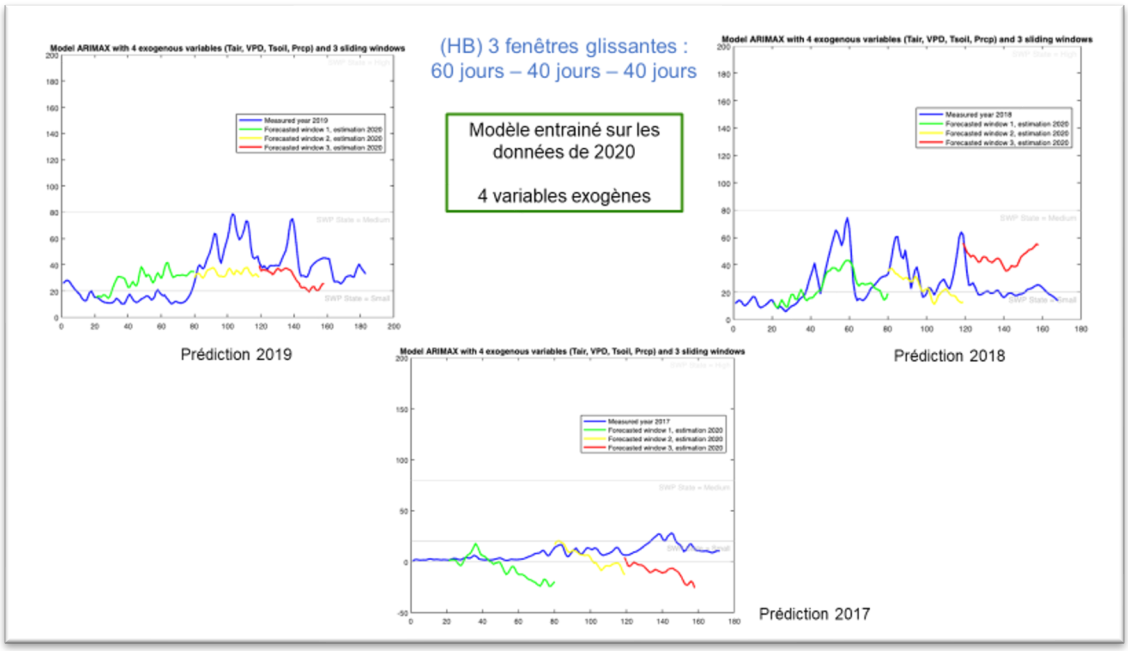
- Exemples de tests réalisés pour le peuplement MW, avec l'approche des fenêtres glissantes



- Exemple de tests réalisés pour le peuplement HW, avec l'approche des fenêtres glissantes



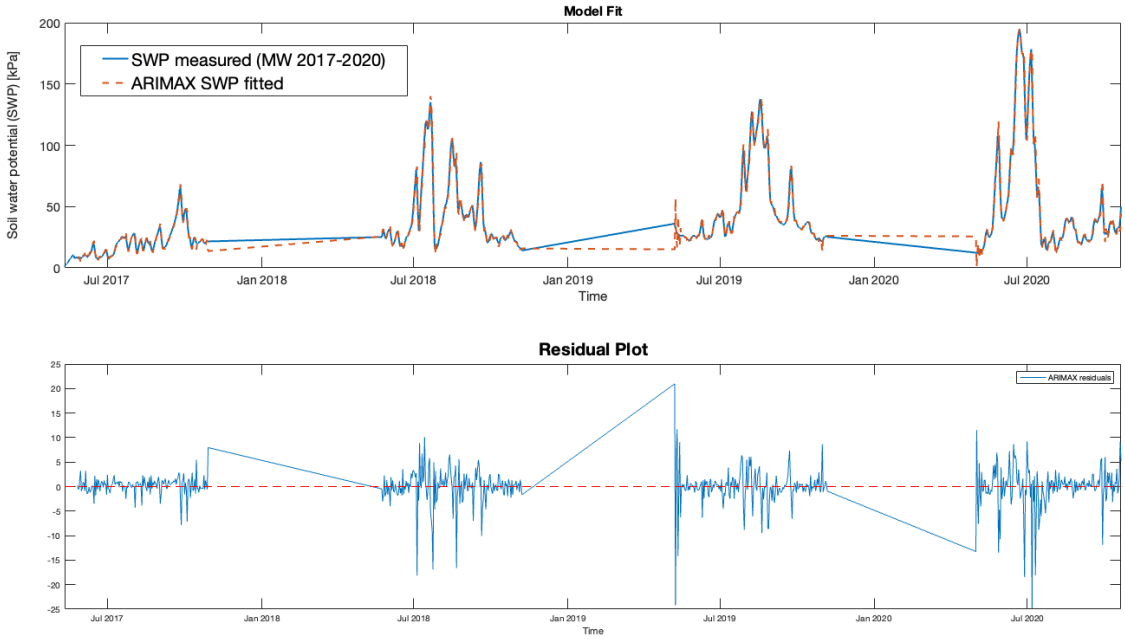
- Exemple de tests réalisés pour le peuplement HB, avec l'approche des fenêtres glissantes



- Estimation du modèle ARIMAX sur toutes les données : exemple pour le peuplement (MW)

Nous obtenons les coefficients AIC et BIC suivants : AIC > 3000 et BIC > 3000

Chaque période de données manquantes, de novembre à mai, ne permet pas un bon ajustement du modèle, ce qui entraîne un coefficient d'information de Akaike élevé.



Une stratégie d'analyse a été mise en place pour contourner ce problème.

Nous pourrions créer une temporalité fictive qui permette de concaténer les données sans créer d'espaces temporels entre chaque année de mesures. Cependant, nous l'avons souligné, un échantillon de données est nécessaire pour initialiser le modèle, cet échantillon devant être au minimum de taille (p+d) jours, pour un modèle ARIMAX(p,d,q).

Afin d'améliorer les modèles et pour profiter de la richesse des données disponibles, nous allons exploiter l'utilisation des variables exogènes, qui sont des variables météorologiques.

Une première approche qui s'appuie sur la similarité des variables exogènes a été testée et présentée dans le mémoire.

Stratégie des tests avec la fonction ARIMAX_forecast_slidingwindow_error()

La fonction est détaillée dans l'Annexe VII.

Quatre variables exogènes sont utilisées (Tair, Tsol, VPD, Prcp).

Différentes tailles de fenêtre peuvent être testées, à des moments différents.

L'année d'estimation du modèle peut varier.

Nous allons garder la taille de l'intervalle d'initialisation de la prédiction égale à 20 jours.

Pour la taille de l'intervalle d'initialisation de l'estimation du modèle, nous prendrons la taille par défaut (égale à 15 jours pour MW et HW, et égale à 10 jours pour HB).

Objectifs des tests :

(1) Évaluer l'impact de l'horizon de prédiction

(Quelle est la taille optimale des fenêtres de prédiction?)

Tracer : F1-score=f(taille en jours) et Accuracy=f(taille en jours)

(2) Évaluer l'impact de l'année sélectionnée pour estimer le modèle ARIMAX

(Hypothèse : l'estimation du modèle sur les données de 2020 permet d'obtenir des prédictions plus exacte)

Exemple d'appel de la fonction ARIMAX_forecast_slidingwindow_error()

Estimation du modèle en 2020, prédiction des données de 2019, peuplement MW considéré, les trois fenêtres de prédiction sont de 20 jours, la première fenêtre commence au jour 60

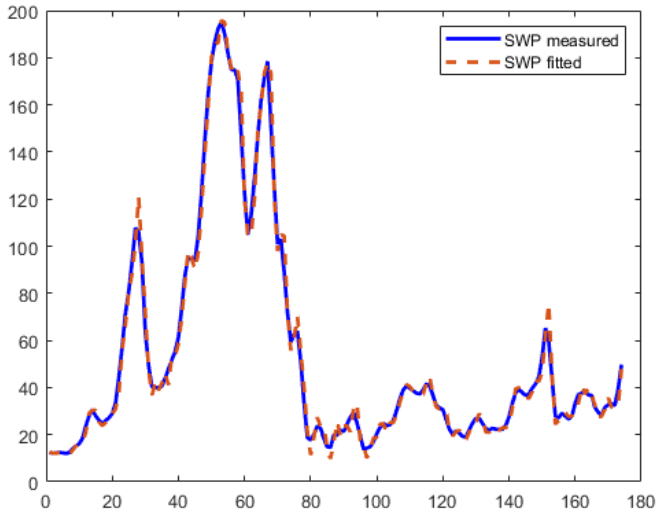
```
 Clf;
 [f1,f2,f3,MSE_w,RMSE_w,MAPE_w_percent,mdl1,mdl2,mdl3,accuracy,precision,recall,
f1Score] =
ARIMAX_forecast_slidingwindow_error(spMW_2020,spMW_2019,'MW',0,0,2020,2019,[20,2
0,20],[20,20,20],60,0,0)
```

Description des sorties

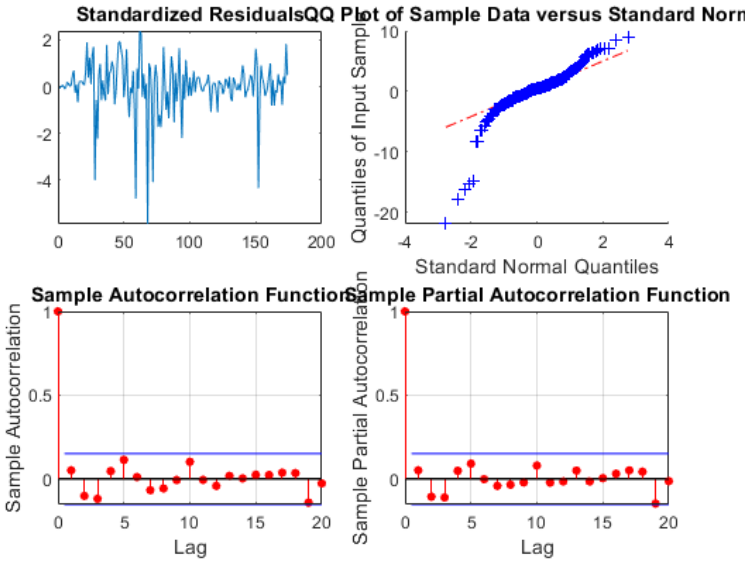
Exemple des graphiques obtenus

(A) Estimation du modèle ARIMAX(14,1,4) sur les données de 2020 pour le peuplement MW

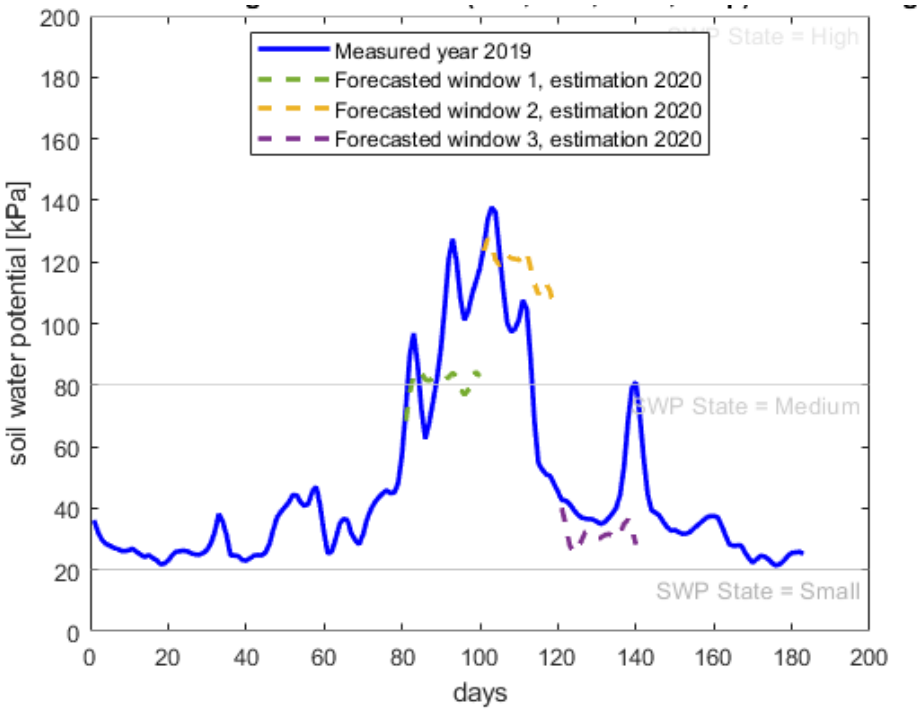
Sortie de la fonction ARIMAX_MW_p14_q4_4X (appelée dans la fonction ARIMAX_prediction_error)



Graphiques des résidus et fonction d'autocorrélation pour vérifier la non-corrélation des résidus au cours du temps

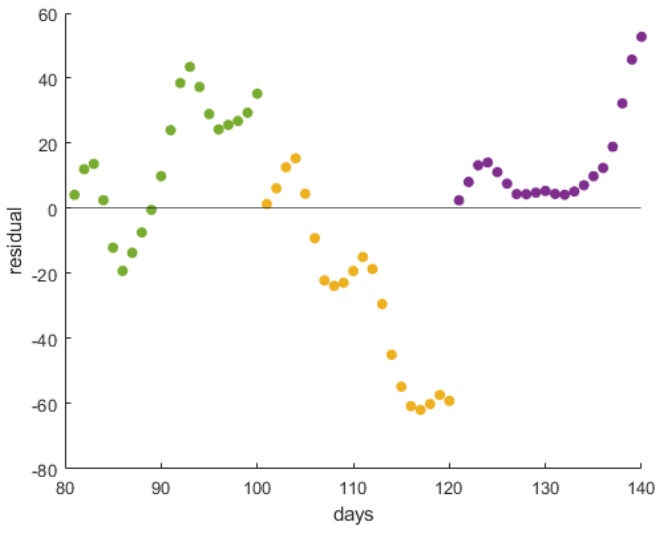


(B) Potentiel hydrique observé (ici en 2019) et prédictions obtenues sur chaque fenêtre (en pointillé sur le graphique), avec le modèle estimé en 2020.

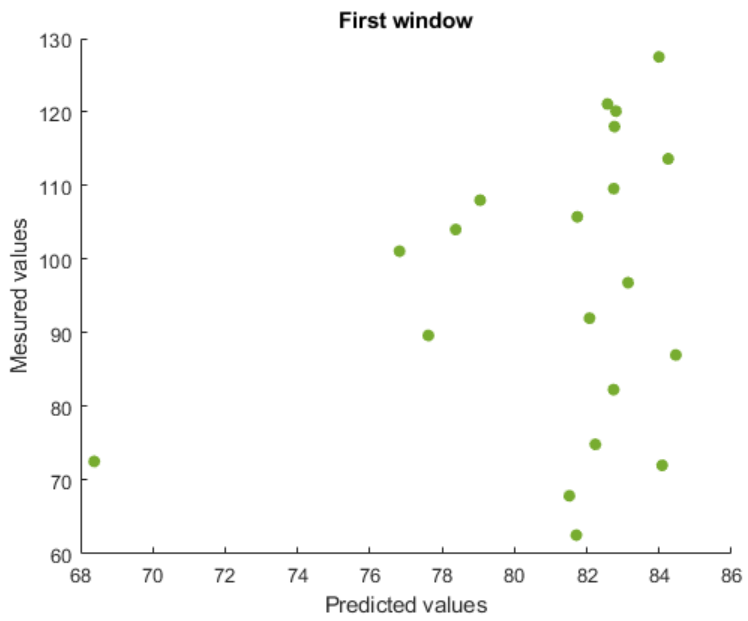


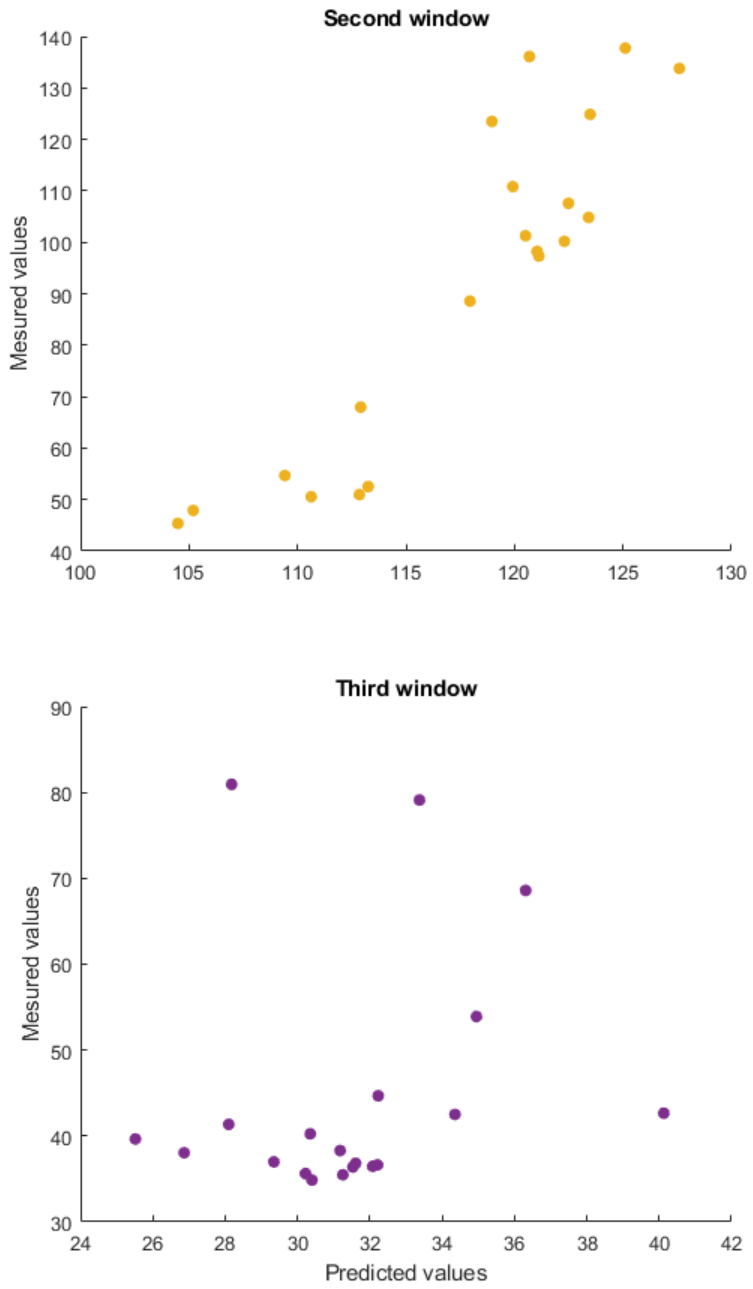
(C) Représentation graphique des résidus (valeur observée – valeur prédite) pour chaque point temporel contenu dans chacune des fenêtres de prédiction (même code couleur)

On observe que sur les 10 premiers jours de la prédiction, les résidus sont compris entre -20kPa et 20kPa.



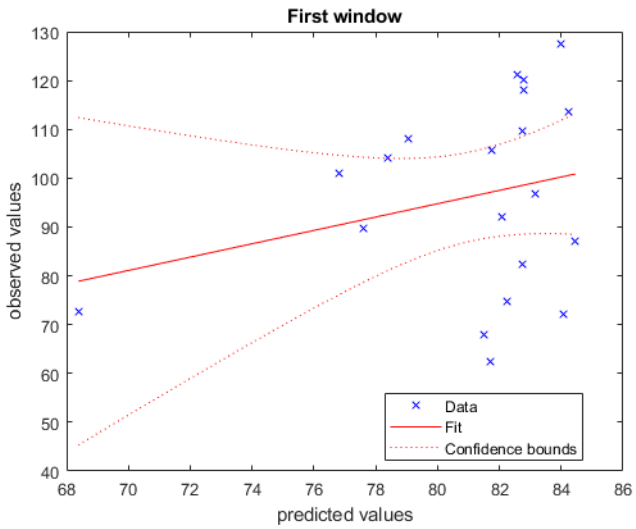
(D) Valeurs (du potentiel) prédites *vs* Valeurs observées (pour chaque fenêtre)

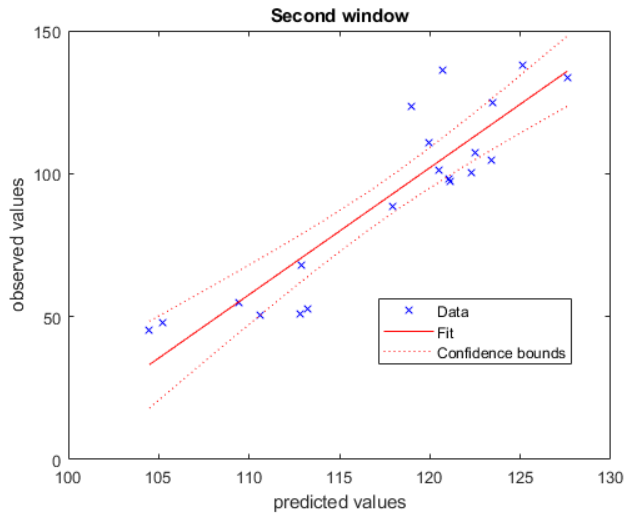


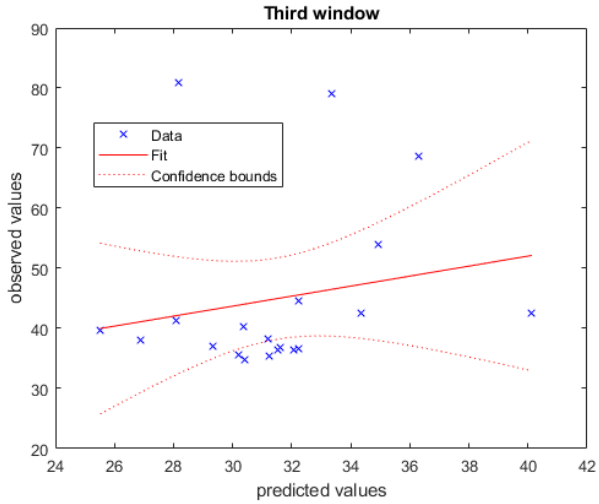


(E) Pour chaque fenêtre de prédiction, un ajustement linéaire est réalisé entre les valeurs prédites et les valeurs observées (sur chaque fenêtre).

Le détail des régressions est également affiché.


First window
observed values
predicted values
Data
Fit
Confidence bounds


Second window
observed values
predicted values
Data
Fit
Confidence bounds


Third window
observed values
predicted values
Data
Fit
Confidence bounds

Affichage dans la console des intervalles d'initialisation du modèle (preSamp) et d'estimation du modèle (forecast), en jours.

Cela permet de vérifier que la prédiction se fait sur des données qui n'ont pas servi à initialiser le modèle.

```
start_w1_preSamp = 60
end_w1_preSamp = 80
start_w1_forecast = 81
end_w1_forecast = 100
start_w2_preSamp = 80
end_w2_preSamp = 100
start_w2_forecast = 101
end_w2_forecast = 120
start_w3_preSamp = 100
end_w3_preSamp = 120
start_w3_forecast = 121
end_w3_forecast = 140
```

MSE, RMSE et MAPE (pour chaque fenêtre) :

```
MSE_w = 1x3
10^3 x
      0.5742    1.3424    0.3686

RMSE_w = 1x3
     23.9618   36.6381   19.1986

MAPE_w_percent = 1x3
     20.0237   48.1980   25.5378
```

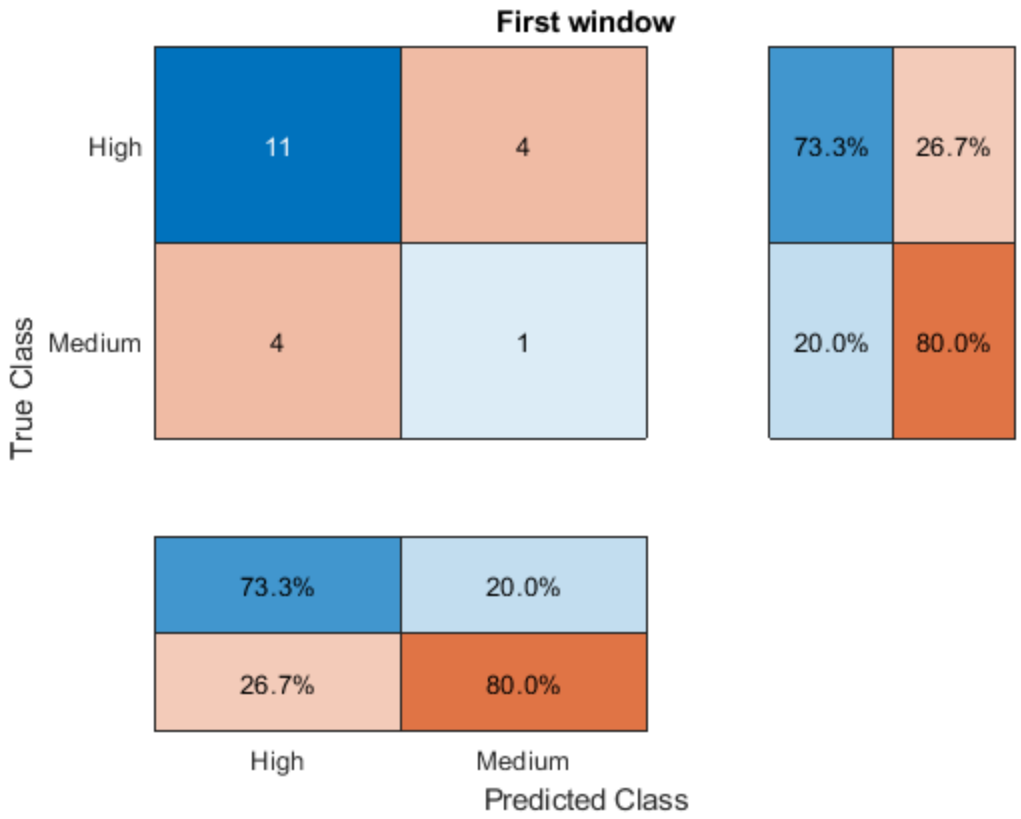
Détail des régressions linéaires (pour la première fenêtre):

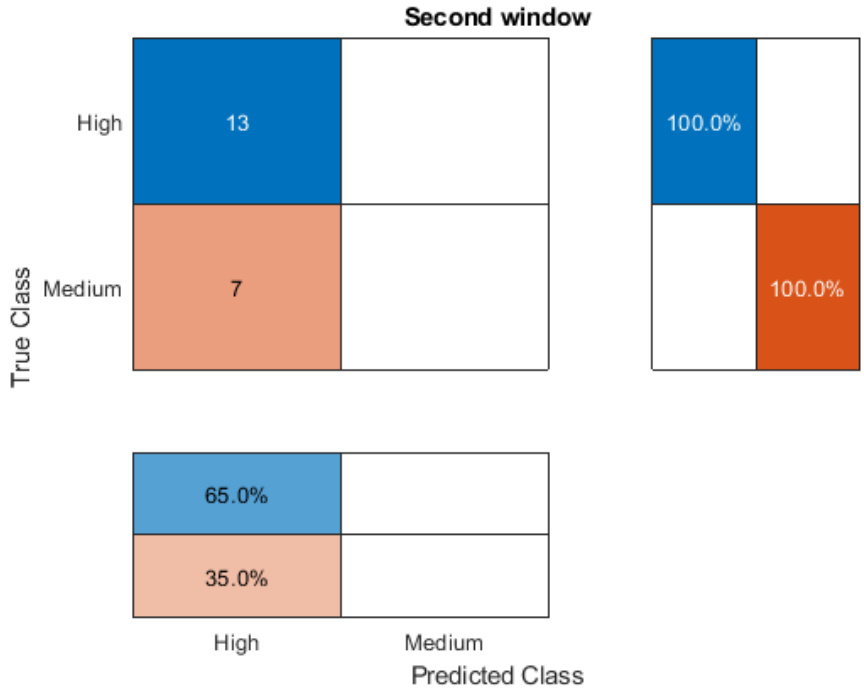
```
mdl1 =
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    ______    ________    _______

    (Intercept)    -14.612      97.62    -0.14968    0.88268
    x1              1.3672     1.2017      1.1377    0.27015


Number of observations: 20, Error degrees of freedom: 18
Root Mean Squared Error: 19.5
R-squared: 0.0671,  Adjusted R-Squared: 0.0153
F-statistic vs. constant model: 1.29, p-value = 0.27
```

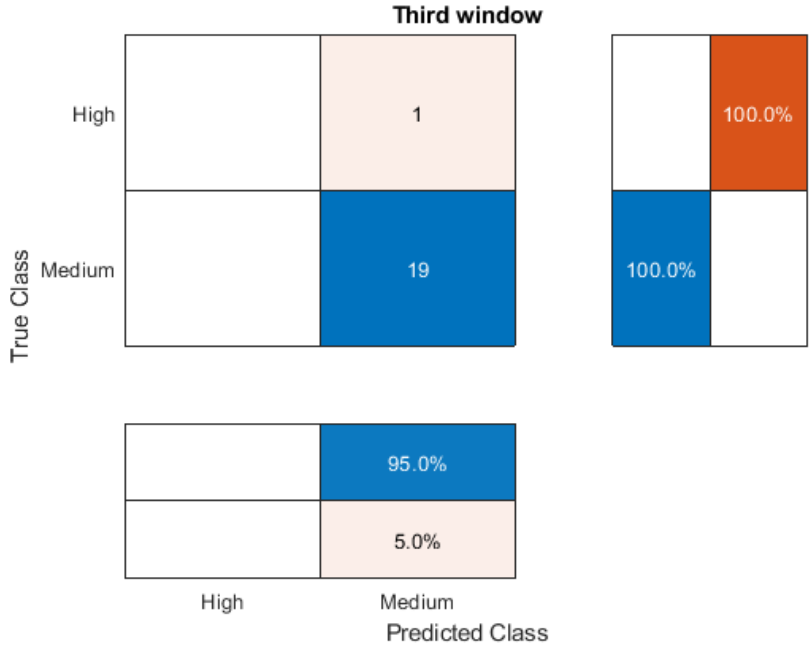
Résultats obtenus sur les états du potentiel hydrique : pourcentage de prédiction, précision, sensibilité et F1-score, pour chaque fenêtre de prédiction considérée :

```
accuracy = 1x3
      60    65    95

precision = 1x3
      0.4667    0.5000    0.5000

recall = 1x3
      0.4667    0.6500    0.9500

F1score = 1x3
      0.4667    0.5652    0.6552
```

Matrices de confusion




Second window
High
Medium
True Class
13
7
100.0%
100.0%
65.0%
35.0%
High
Medium
Predicted Class


Third window
High
Medium
True Class
1
19
100.0%
100.0%
95.0%
5.0%
High
Medium
Predicted Class

Conclusion : dans cet exemple, nous montrons que le pourcentage de prédictions exactes déterminé à partir des états du potentiel, et non à partir des valeurs numériques du potentiel hydrique, peut atteindre 95%, comme c'est le cas pour la troisième fenêtre.

Dans le cas de la seconde fenêtre, la différence se fait au niveau de la diminution brutale du potentiel à la fin de l'intervalle de prédiction considéré. L'état du potentiel prédit reste égal à celui observé sur la majorité de l'intervalle, mais la diminution du potentiel même si elle est prédite, est moins importante (le potentiel prédit reste dans un état High, alors que le potentiel observé a atteint l'état Medium à la fin de la fenêtre de prédiction).

Le calcul du score-f1 permet de comparer les performances du modèle sur chaque fenêtre, mais aussi permet une comparaison entre les stratégies testées.

Phase 6 : D'autres approches possibles

D'autres approches possibles, spécifiques aux propriétés du potentiel hydrique

Les modèles de type ARIMA sont adaptés aux phénomènes linaires.

Dans notre cas, il conviendrait d'explorer les modèles ARCH (*AutoRegressive Conditional Heteroskedasticity*).

On parle d'*hétéroscédasticité* lorsque les variations des résidus de la variable examinée sont différentes (elles varient au cours du temps). Précisons que les résidus sont définis comme la différence entre les valeurs observée et les valeurs modélisées.

Le test de Goldfeld et Quandt permet de vérifier l'homoscédasticité d'une série temporelle, ie. la répartition homogène des résidus. En pratique, c'est ce phénomène qui est souhaitable, car il faut s'assurer que les résidus correspondent bien à des aléas de mesure.

Or, l'hétéroscédasticité d'une variable est liée à la non-normalité de sa distribution.

Nous avons montré que le potentiel hydrique présente distribution non-gaussienne et non-stationnaire.

Cela demande donc un traitement particulier[29].

Cependant, nous allons tester les modèles ARCH(q) /GARCH(p,q) (Generalized ARCH)[30].

Ces modèles sont utilisés en économétrie pour modéliser des séries temporelles qui comportent des volatilités variables, ie., des périodes agitées suivies par des périodes de calme relatif.

Nous pourrions faire l'hypothèse que le potentiel hydrique se comporte ainsi.

Autre aspect spécifique du potentiel hydrique : une autocorrélation qui se maintient à des lags élevés. Selon Cowpertwait et Metcalfe, cela est la marque de processus appelés à mémoire longue (*long-memory*) :

[29] Balakrishna, N. 2022. Non-Gaussian Autoregressive-Type Time Series. Singapore: Springer. https://doi.org/10.1007/978-981-16-8162-2.

[30] Bauwens, L., Hafner, C., Laurent, S. (2012). Handbook of volatility models and their applications. John Wiley & Sons, Inc (https://onlinelibrary.wiley.com/doi/book/10.1002/9781118272039)

« Some time series exhibit marked correlations at high lags, and they are re-ferred to as long-memory processes. Long-memory is a feature of many geo-physical time series (…). Mudelsee (2007) shows that long-memory is a hydrological property that can lead to prolonged drought or temporal clustering of extreme floods» (Cowpertwait & Metcalfe)[31].

Ils proposent une approche de type FARIMA (Fractional ARIMA), que nous ne testerons pas dans le cadre de notre étude.

Approche NARX

Nous allons tester l'approche NARX (*Non-linear autoregressive with exogeneous variables*). C'est une approche qui exploite un réseau neuronal dit *récurrent*. Ce type de réseaux permet de capturer la dynamique de systèmes complexes.

En entrée du réseau, nous avons des séries chronologiques $x_i(t)$ qui représentent les variables exogènes du modèle, avec i compris entre 1 et le nombre de variables considérées. En sortie nous avons la prédiction y(t), dans notre cas $\Psi(t)$ le potentiel hydrique du sol. La prédiction s'écrit alors :

$$y(t) = f\big(x_i(t-1), \dots, x_i(t-d), y(t-1), \dots, y(t-d)\big).$$

Avec *f* une fonction non linéaire et *d* un délai temporel fixé.

L'équation de *y(t)* nous indique que les valeurs passées de y au temps (t-1) à (t-d) sont nécessaires à la prédiction de y au temps t. C'est l'aspect *récurrent* propre à ces réseaux. L'historique de y s'ajoute alors aux données temporelles des variables exogènes x_i.

Cette approche est accessible grâce aux fonctions implémentées dans la boîte à outils de Matlab *Deep Learning.*

Nous avons considéré les 6 variables exogènes d'intérêt (T_{air}, T_{sol}, VPD, RelH, SRad, Prcp).

Dans un premier temps le modèle est entraîné sur les données collectées en 2019, pour le peuplement MW.

[31] Cowpertwait, P.S.P, Metcalfe, A.V. (2009). *Long-memory processes*, in Introductory Time Series in R (pp 159-170)

Les paramètres optimaux trouvés sont les suivants (Figure 56) :

2 couches : une couche cachée et une couche de sortie

30 neurones dans la couche cachée

Délai de 8 jours introduit (d = 8 jours)

Algorithme de Levenberg-Marquardt (pour l'optimisation des poids *w* et des biais *b*)

80% des données d'entrée sont utilisées pour l'entraînement (training), 10% pour la validation et 10% pour le test.

Feedforward network

2 couches (une couche cachée et une couche de sortie)
30 neurones dans la couche cachée

Délai temporel : d = 8 jours

Fonction d'activation de la couche cachée : sigmoïde
Fonction d'activation de la couche de sortie : linéaire

6 variables exogènes $X_i(t)$: Tair, Tsol, VPD, Prcp, RelH, Srad

1 variable en sortie (le potentiel hydrique) : y(t) = f(y(t -1),…y(t-d))

w (poids)
b (biais)

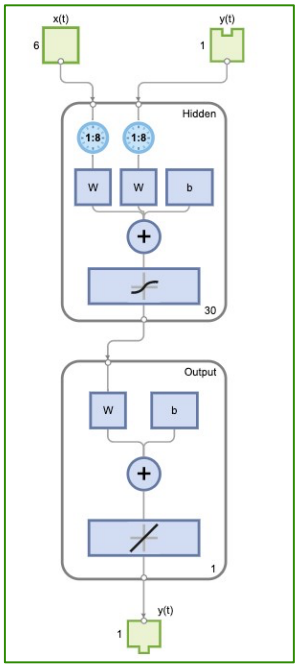


Figure 56 : Paramètres du modèle NARX testé.

Les résultats obtenus sont détaillés dans le mémoire.

Pour aller plus loin, nous pourrions appliquer des modèles de type LSTM (*Long Short - Term Memory*). Ils s'appuient sur l'apprentissage profond et permettent de stocker des informations - au moyen de portes d'oubli et de cellules - qui seront utilisées ultérieurement, introduisant alors une dépendance à plus long terme. Un des avantages de ce type d'approche est de pouvoir éviter le problème de la disparition du gradient (*vanishing gradient*) qui intervient dans les réseaux de neurones récurrents. En effet, lorsque l'optimisation des poids w^k, avec k désignant la couche neuronale cachée k, se fait grâce à la méthode de descente de gradient, il peut arriver quand la fonction de coût E(w) est de type sigmoïde σ ou tangente hyperbolique que la valeur du gradient devienne très petite, surtout dans le cas de la propagation de l'erreur pour des réseaux qui comportent un grand nombre de couches cachées.

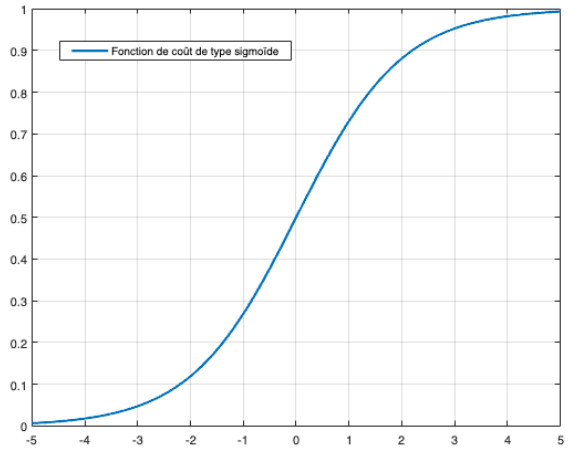
Soit la valeur du poids *w* de la couche (k+1) définie par : $w^{k+1} = w^k - \alpha^k \nabla_w E(w)$

Si la fonction de coût est de type sigmoïde alors $E(w) = \frac{1}{1+e^{-w}}$. Elle est continue et dérivable et sa dérivée est égale à : $\frac{dE}{dw} = E(w).(1 - E(w))$.

Ainsi, les valeurs du gradient $\nabla_w E(w)$ sont comprises entre 0 et ¼. Ces valeurs sont faibles, et peuvent tendre vers 0 assez rapidement, provoquant alors la disparition du gradient de E(w).

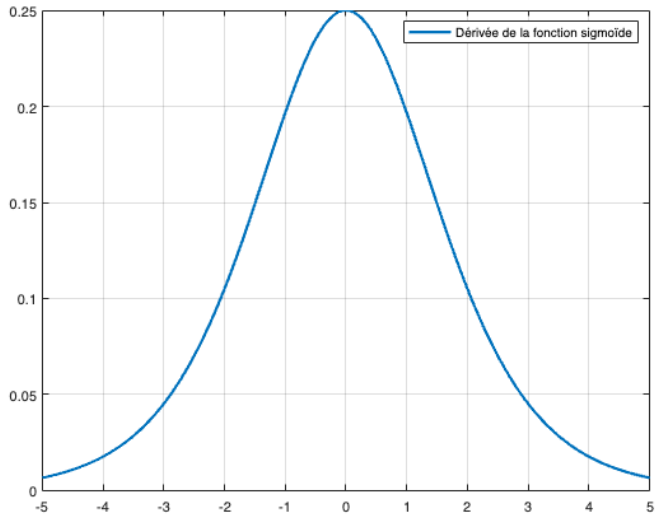
Une visualisation de la fonction et de sa dérivée sous Matlab permet d'illustrer cette explication.

```
w = -5:0.01:5;
E = 1./(1+exp(-w));
plot(w,E,'LineWidth',2)
grid on
legend('Fonction de coût de type sigmoïde')
legend("Position",[0.18114,0.82608,0.31964,0.035714])
```



La fonction sigmoïde a deux asymptotes qui tendent vers $\sigma(x) = 0$ quand x tend vers $-\infty$ et $\sigma(x) = 1$ quand x tend vers $+\infty$ et son point d'inflexion est en (0;0,5).

```
dE=E.*(1-E);
plot(w,dE,'LineWidth',2)
grid on
legend('Dérivée de la fonction sigmoïde')
```



Conclusion

L'analyse des séries temporelles grâce à la démarche de Box-Jenkins permet le développement de modèles prédictifs de type autorégressifs à moyenne mobile qui exploitent l'intégration de variables exogènes dans l'objectif d'améliorer la prédiction. Pour choisir les variables à intégrer, nous avons montré qu'une étude préalable s'appuyant sur des méthodes de classification telles que les arbres de décision associés à l'analyse de la causalité de Granger est pertinente.

Un code a été implémenté sous Matlab pour permettre la mise en place de toutes les étapes méthodologiques décrites dans le mémoire associé à ce rapport technique. Les données brutes collectées à la SBL, les données transformées et le code sont disponibles dans les archives du DOT-Lab.

Annexe

I . Description détaillée de l'étape d'échantillonnage et du calcul de l'incertitude associée à chaque observation moyennée

Pour procéder à l'échantillonnage des données, nous avons créé la fonction **sampling_frequency** (voir le fichier sampling_frequency.m).

Nous allons appliquer la fonction *sampling_frequency* pour obtenir $Temp_{avg}$ et SM_{avg} échantillonnés selon différentes fréquences.

Par défaut, $Temp_{avg}$ et SM_{avg} seront les valeurs échantillonnées sur 1 heure (en considérant donc 8 mesures par heure).

Input : Les données nettoyées des données aberrantes (en température).

L'avantage d'implémenter notre propre fonction d'échantillonnage est que nous pouvons calculer l'incertitude liée à la mesure échantillonnée.

Nous pouvons dégager plusieurs types d'incertitude[32] :

- l'incertitude de type A
- l'incertitude de type B
- l'incertitude-élargie

L'incertitude de type A est liée à la mesure (erreur aléatoire), au temps t_i, se définit comme étant l'écart-type $(sd)_{ti}$ de la moyenne (sd, pour standard deviation en anglais), ici celle de TMPA et de TMPB, et SMSC et SMSD, pour un temps t_i donné.

Nous allons ici utiliser l'incertitude liée à la répétabilité de la mesure, qui est définie comme :

$$U_a = {sd_{t_i}}/{\sqrt{n}}$$

avec n le nombre de mesures effectuées.

(On diminue alors l'incertitude grâce à l'augmentation du nombre de mesures).

[32] https://fr.wikipedia.org/w/index.php?title=Incertitude_de_mesure&oldid=cur#Incertitude_et_tol%C3%A9rance

L'incertitude de type B est liée à l'appareil de mesure (pas forcément liée à l'opérateur ici, car les sondes sont posées de façon permanente).

Elle comprend l'incertitude liée à l'erreur de justesse de l'appareil (U_{b1}) et celle liée à sa résolution (U_{b2}) i.e. la plus petite graduation disponible pour la mesure.

Pour la sonde de température, nous avons :

U_{b1}= 0.6 °C (donnée par le constructeur)

$$U_{b2} = {}^{0,1}/_{2\sqrt{3}}$$

Pour la sonde du potentiel hydrique :

U_{b1} non connue

$$U_{b2} = {}^{0,1}/_{2\sqrt{3}}$$

Nous devons ensuite calculer l'incertitude-composée U_c, telle que :

$$U_c = \sqrt{(U_a{}^2 + U_{b1}{}^2 + U_{b2}{}^2)}$$

Enfin, nous pouvons déterminer l'incertitude dite élargie U, qui se définit à partir de l'incertitude-composée, multipliée par un facteur d'élargissement, k_{elg} :

$U = k_{elg} * U_c$

Ce facteur dépend du nombre de mesures, et du niveau de confiance accordée à la mesure (correspond au coefficient t de la table de Student).

Pour 8 mesures, avec un niveau de confiance de 95%, k_{elg} sera égal à 1,86 (pour un niveau de confiance de 90%, k_{elg} =1,397, donc comparable à un niveau de confiance de 90% avec 2 mesures répétées.)

Dans notre démarche d'échantillonnage, nous considérons que la mesure est effectuée n fois, avec n le nombre de mesures considérées dans l'intervalle temporel de l'échantillonnage, multiplié par 2, pour considérer les 2 sondes.

Par défaut, nous allons sélectionner un échantillonnage d'une heure, ce qui correspond à (n = 4*2 = 8) mesures répétées, pour chaque station.

Pour le moyennage par bloc, nous avons n =32 (4*2*4)

Pour le moyennage par peuplement, nous avons :

n=96 (pour MW et HB)

n=64 (pour HW)

Les résultats seront donc :

potentiel_mesuré ± incertitude élargie (en kPA)

temperature_mesurée ± incertitude élargie (en ^{o}C)

D'après les notices techniques des sondes (Spectrum Technologies, Inc) modèle de type WatchDog 1000 Micro Station, nous avons les informations suivantes, concernant la précision :

External (soil) temperature sensor #3667 / accuracy : ± 0.6 ^{o}C (pour des températures comprises entre -40 et +85 ^{o}C)

Résolution : 0.1 ^{o}C

Watermark Soil Moisture sensor #6460 / précision : N/A

Résolution: 0.1 kPa

Nous n'avons donc pas d'information sur la tolérance de l'appareil.

Dans le cas de la mesure de température, nous ajouterons 0.6 ^{o}C, comme incertitude liée à l'appareil de mesure.

II. Interpolation par BSpline (cubique)

B-Spline signifie *Basis-Spline*. Au lieu de former la spline comme une combinaison de courbes de Bézier, courbes polynomiales paramétriques, l'idée est d'utiliser un autre ensemble de polynômes comme base, les polynômes dits de Bernstein. Néanmoins, l'approche est similaire : elle permet de définir une équation paramétrique, dans l'objectif d'interpoler une courbe existante.

Une combinaison linéaire de ces polynômes de base $B_D(t)$, d'ordre D, forme la B-Spline.

Mathématiquement, cela se traduit par l'équation suivante :

$$c(t) = \sum_{k=0}^{n} P_k \, B_{k,D}(t)$$

P_k étant le k-ième point de contrôle et le facteur du k-ième polynôme de base $B_{k,D}(t)$

La suite des points (P_0, …., P_k) forme le polygone de contrôle de Bézier.

Pour k=3, nous pouvons écrire que la somme des coefficients associés aux points de contrôle est égale à 1 (propriété des polynômes de Bézier) :

$$1^3 = 1 = (1 - t + t)^3 = (1 - t)^3 + 3(1 - t)^2 t + 3(1 - t)^2 + t^3$$

$$P(t) = (1 - t)^3 P_0 + 3(1 - t)^2 t P_1 + 3(1 - t) t^2 P_2 + (t)^3 P_3 \;\; avec\; t \in [0; 1]$$

P(t) est une forme paramétrique, représentant une courbe de Bézier de degré 3, définie par 4 points (P_0, P_1, P_2, P_3).

La courbe paramétrique ne passe pas forcément par les quatre points, mais les autres points permettent de donner une information sur la direction de la courbe.

III. Filtre Savitzky-Golay

Le filtre Savitzky-Golay est un filtre passe-bas caractérisé par une moyenne mobile pondérée par un polynôme[33]. Il consiste à faire concorder un polynôme de degré n sur un ensemble de données, en considérant un voisinage (k-m)….(k+m) de chaque point k constituant cet ensemble, avec m la demi-largeur de la fenêtre de la moyenne mobile considérée. Chaque point k est alors remplacé par la valeur concordante (la valeur du *fit*) de ce même point k.

Dans notre contexte, le coefficient polynomial optimal a été trouvé égal à 6 (n=6) et la demi-largeur de la fenêtre de la moyenne mobile égale à 1,5 jour (équivalent à une moyenne mobile centrée sur 3 jours).

Les filtres SG ont la particularité de préserver la largeur et la hauteur des pics d'un signal, ce qui est notre cas.

De plus, le fait d'avoir une grande densité de points, impliquant donc des points très proches, améliore la qualité de ce type de filtre.

[33] Savitzky, A., Golay, M.J.E. (1964) Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, *36* (8), 1627-1638

IV. Description de l'obtention des données météo avec le logiciel BioSIM

Deux sites ont été considérés : la Station de biologie des Laurentides (SBL) et la station située à Ste Émilie (générées pour un autre projet).

Coordonnées utilisées pour la SBL :

Latitude = 45.9865

Longitude = -73.9963

Élévation = 366m

Coordonnées pour la station située à Ste Émilie :

Latitude = 46.369

Longitude = -73.685

Élévation = 339m

Obtention des 4 stations les plus proches, avec une pondération associée à chaque station.

SBL : station de Saint-Hippolyte (pondération : 99.9 %)

Ste Émilie :

Normals

No	KeyID	Name	Latitude	Longitude	Elevation	Shore	Distance (km	Delta elevation (m)	Delta shore distance (km)	Virtual distance including elevation and shore (km)	Weight (%)
4	cqsl	Saint-Michel	46.6825	-73.9137	375	244.6	39	36	7	39.8	9.5
1	7017080D	St-Come (QC)	46.2833	-73.75	244	245.8	10.8	-95	8.2	16.5	55.2
2	7016902D	Ste-Beatrix (QC)	46.2	-73.6	198	239	19.9	-141	1.4	24.4	25.2
3	7016960D	St-Charles De Mandeville (QC	46.35	-73.35	167.6	214.6	25.8	-171	-23	38.6	10.1

St Michel : 9.5%

St Côme : 55.2%

Ste Béatrix : 25.2%

St Charles de Mandeville : 10.1%

Application des pondérations, et obtention des données générées pour la localisation demandée.

Bases de données utilisées :

Data type sources : from observations

Pour les données normalisées : Quebec++ 1981-2020 (adjusted from 2003-2017)

Pour les données journalières : Canada 1980 - 2020

Pour les données par heure : Canada 1980 – 2020 (HourlyDB)

Conifer canopy : *sélectionnée*

Nb de répétition : 1

Random seed

Nb de stations environnantes : 4 (par défaut)

Paramètres du modèle :

Model to execute : Climatic (daily and hourly)

Les données sont ensuite exportées sous Excel.

Remarque : le format des données temporelles est à adapter pour les analyses. Une fonction a été implémentée sous Matlab dans ce but.

Données générées :

Tmin: Hourly minimum temperature [°C]

Tair: Hourly mean temperature [°C]

Tmax: Hourly maximum temperature [°C]

Prcp: Hourly total precipitation [mm]

Tdew: Hourly mean dew point temperature [°C]

RelH: Hourly mean relative humidity [%]

SRad: Hourly solar radiation [W/m²]

V. Détermination de l'importance des prédicteurs dans une approche de classification de type arbre de décision

La Figure 57 illustre un arbre comprenant trois nœuds – un nœud parent et deux nœuds enfants - et deux branches.

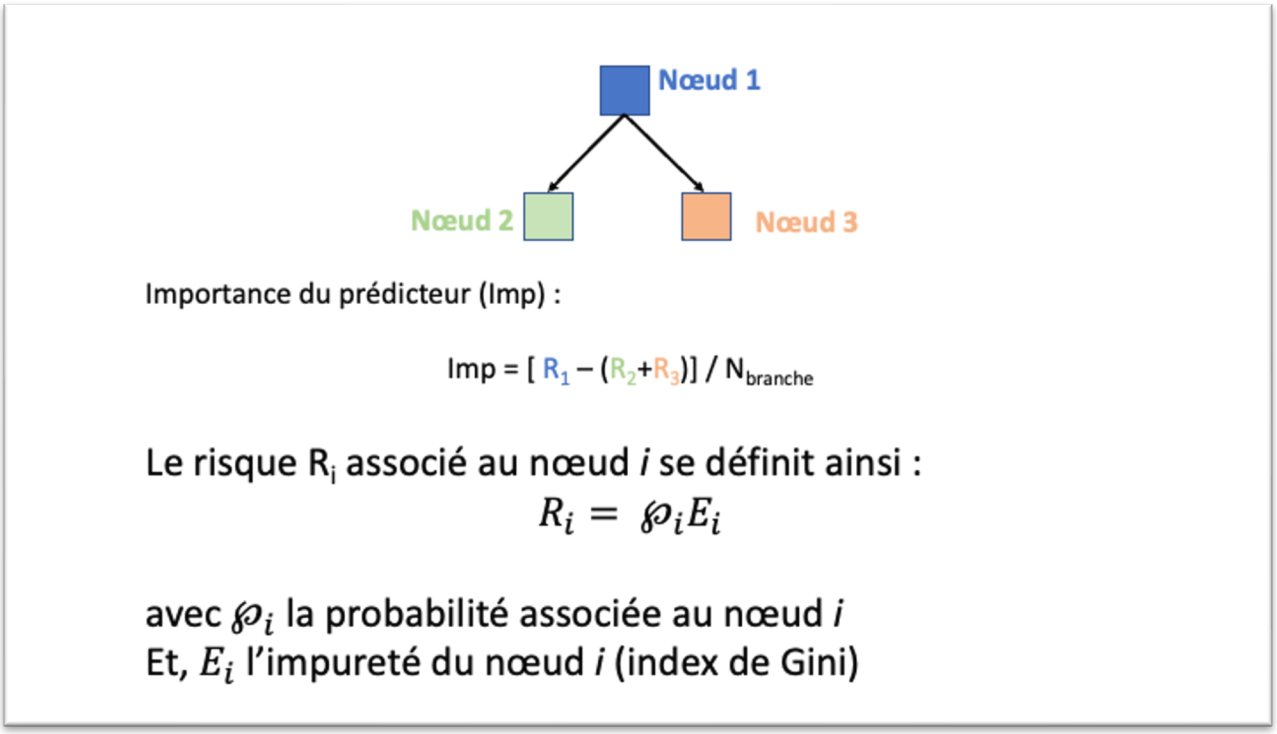


Figure 57 : Calcul de l'importance d'un prédicteur.

L'importance d'un prédicteur se définit comme le rapport de la différence entre le risque associé au nœud parent et le risque total associé aux deux nœuds enfants et le nombre total de branches.

Le risque se calcule à partir de la probabilité associée au nœud considéré et l'impureté de ce nœud, appelée l'index de Gini.

L'impureté au nœud *i* se détermine ainsi : $1 - \sum_i \wp^2 (i)$

Si nous considérons deux classes par exemple, C_1 et C_2, qui correspondent aux deux nœuds enfants,

alors $\wp(1)$ = (nombre d'éléments appartenant à la classe C_1) / (nombre total d'éléments considérés pour l'entraînement de l'arbre).

VI. Test de Ljung-Box (Greta Ljung et George Box)

Ce test statistique est un test dit porte-manteau, c'est-à-dire qu'il définit de façon précise l'hypothèse nulle testée, mais l'hypothèse alternative est moins bien définie. Appliqué aux séries temporelles, le test de Ljung-Box permet de tester l'autocorrélation des résidus d'un modèle, souvent de type autorégressif. Il teste en effet si un groupe d'autocorrélations est différent de zéro. Au lieu de tester l'aspect aléatoire des résidus à chaque lag, c'est sur tous les lags considérés que le test est effectué.

Dans ce test, l'hypothèse nulle H_0 est donc la suivante :

Les données sont distribuées de façon indépendante (les corrélations dans la population dont est extrait l'échantillon sont égales à 0, illustrant le fait que les corrélations observées dans les données résultent du caractère aléatoire de la formation de l'échantillon).

Dans notre cas, le test est appliqué aux résidus obtenus après ajustement du modèle sur les données réelles.

L'hypothèse alternative H_1 se définit ainsi : les données ne sont pas distribuées de façon indépendante, elles montrent une corrélation au niveau de la série.

Pour valider un modèle ARIMA, les résidus ne doivent pas être auto-corrélés.

VII. Détails des fonctions implémentées pour la prédiction du potentiel hydrique du sol

fonction ARIMAX_forecast_slidingwindow_error()

Paramètres d'entrée :

Données pour l'estimation du modèle ARIMAX

Données à prédire (pour la validation)

Peuplement ciblé (qui permet d'appeler le modèle ARIMAX correspondant)

Nombre de jours nécessaires pour l'initialisation du modèle : si égal à 0, alors par défaut, le nombre de jours est égal à (p+d), p étant le coefficient de la partie autorégressive et d le degré de différenciation

Nombres de jours utilisés pour l'estimation du modèle : si égal à 0, alors tous les jours du jeu de données utilisé pour l'estimation seront considérés

Année de l'estimation (pour l'affichage sur le graphique)

Année de la prédiction (pour l'affichage sur le graphique)

Vecteur définissant la taille (en jours) de chaque fenêtre de prédiction

Vecteur définissant le nombre de jours utilisés pour l'initialisation du modèle pour la prédiction

Translation de la première fenêtre de prédiction (taille en jours)

Translation de la seconde fenêtre de prédiction (taille en jours)

Translation de la troisième fenêtre de prédiction (taille en jours)

Remarque : cette translation a été ajoutée dans la nouvelle version de la fonction, elle permet de pouvoir prédire le potentiel à partir de n'importe quel jour dans l'intervalle de temps considéré.

La seconde fenêtre commencera une journée après la fin de la fenêtre 1, si la translation est nulle.

Idem pour la troisième fenêtre (par rapport à la seconde fenêtre).

Paramètres de sortie :

Données du potentiel hydrique prédites sur la première fenêtre de prédiction

Données du potentiel hydrique prédites sur la seconde fenêtre

Données du potentiel hydrique prédites sur la troisième fenêtre

Mean Square Error (MSE), pour chaque fenêtre

Root Mean Square Error (RMSE), pour chaque fenêtre

Mean Absolute Percent Error (MAPE), pour chaque fenêtre

Régression linéaire entre les données prédites et les données observées, pour chaque fenêtre

Pourcentage de prédiction, calculé sur les états du potentiel (Small, Medium, Élevé)

Remarque : les seuils de chaque état peuvent être changés dans le code de la fonction.

Précision pour chaque fenêtre

Recall (sensibilité) pour chaque fenêtre

F1-score pour chaque fenêtre (moyenne harmonique de la précision et de la sensibilité)

Graphiques en sortie :

Ajustement (fit) du modèle ARIMAX (sur les données d'estimation du modèle)

Résidus du modèle (différence entre les valeurs observées et les valeurs prédites)

Prédictions faites sur chaque fenêtre (en pointillé) et valeurs observées du potentiel

Valeurs prédites *vs* Valeurs observées

Régression linéaire (sur les données numériques du potentiel hydrique)

Matrices de confusion (pour chaque fenêtre de prédiction)

```
 function
[f1,f2,f3,MSE_w,RMSE_w,MAPE_w_percent,mdl_output,accuracy_output,precision_out
put,recall_output,F1_score_output] =
ARIMAX_forecast_slidingwindow_error(dataInputEstimation,dataInputForecast,spec
ies,preSampleEstimation_days,SampleEstimation,year_estimation,year_measured,ho
rizon_days,preSampleHorizon_days,translation_w1,translation_w2,translation_w3)
 % This function estimates and evaluates an ARIMAX model using three
 % different models, depending on the species considered
 % Four exogenous variables are considered (Tair, VPD, Tsoil,Prcp)
 % Functions called : ARIMAX_HB_p9_d1_q4_4X, ARIMAX_HW_p12_d1_q4_4X,
ARIMAX_MW_p14_d1_q4_4X, ARIMAX_MW_dd_4X (with 4 exogenous
 % variables) and level_potential_3levels
 % The forecasting is made on three different windows
 % Input :
 % input data for the model estimation (with exogenous variables)
 % input data for the forecasting step (with the exogenous variables
 % observed during the pre-sampling interval)
 % name of the species (MW, HW, HB or dd if the dry-down period is
 % considered)
 % preSample for the model estimation (number of days) : 0 for (p+d) days
 % sampleEstimation (if == 0 then all the data will be considered for the
 % estimation of the model)
 % year of the data used for the model estimation (for the graph legend)
 % year of the data used for the model validation (for the graph legend)
 % number of days for the forecasting horizon : vector with a number of days
 % for each forecasting window
 % vector for the preSample for the forecasting horizon (number of days)
 % number of days for each translation
 % This translation allows to start the prediction at a specific moment
 % Output : summary of the model, graphs, forecasted data f
 % Mean square error for each window (vector of 3 values)
 % Root mean square error for each window (vector of 3 values)
 % Mean absolute percent error for each window (vector of 3 values)
 % Percentage of accurate predictions based on the potential states
 % Precision of the model
 % Recall
 % F1-score
 % Graph of the residuals
 % Linear regression (predicted vs observed data)
 % B.Courcot mars 2023



 % Création du modèle mdl
 % Appel de la fonction pour HB avec ARIMAX(9,1,4) et 4 variables exogènes
```

```
 % Appel de la fonction pour HW avec ARIMAX(12,1,4) et 4 Xi
 % Appel de la fonction pour MW avec ARIMAX(14,1,4) et 4 Xi
 % Appel de la fonction spécifique aux périodes de sécheresse-flash

 if species=='HB'
     mdl =
ARIMAX_HB_p9_d1_q4_4X(dataInputEstimation,preSampleEstimation_days,SampleEstim
ation);
 elseif species=='HW'
     mdl =
ARIMAX_HW_p12_d1_q4_4X(dataInputEstimation,preSampleEstimation_days,SampleEsti
mation);
 elseif species=='MW'
     mdl =
ARIMAX_MW_p14_d1_q4_4X(dataInputEstimation,preSampleEstimation_days,SampleEsti
mation);
 else
     mdl=
ARIMAX_MW_dd_4X(dataInputEstimation,preSampleEstimation_days,SampleEstimation)
;
 end


 % Préparation des figures
 g1=figure;
 g2=figure;
 g3=figure;
 g4=figure;
 g5=figure;
 g6=figure;
 g7=figure;
 g8=figure;
 g9=figure;
 g10=figure;
 g11=figure;

 % Quatre variables exogènes Xi : Tair, VPD, Tsoil, Prcp


 %Fenêtre 1 : estimation du modèle et prédiction

 % Définition des intervalles d'estimation et de prédiction
 start_w1_preSamp = translation_w1 % on affiche le début de l'intervalle
d'initialisation de l'estimation
 end_w1_preSamp = start_w1_preSamp+preSampleHorizon_days(1) % on récupère le
1er paramètre d'entrée pour la taille de l'échantillonnage
```

```
 start_w1_forecast = end_w1_preSamp+1 % la prédiction commence au jour suivant
la fin de l'échantillonnage
 end_w1_forecast = start_w1_forecast + horizon_days(1)-1

 % Appel de la fonction Matlab forecast(), avec en entrée le modèle mdl
 [f1,YMSE1]=forecast(mdl,horizon_days(1),'Y0',dataInputForecast.mean_potential
_sampled(start_w1_preSamp:end_w1_preSamp),'XF', ...

[dataInputForecast.Tair(start_w1_forecast:end_w1_forecast),dataInputForecast.V
PD(start_w1_forecast:end_w1_forecast), ...

dataInputForecast.mean_temp_sampled(start_w1_forecast:end_w1_forecast),dataInp
utForecast.Prcp(start_w1_forecast:end_w1_forecast)]);


 % Fenêtre 2 : estimation du modèle et prédiction

 start_w2_preSamp = translation_w2 + end_w1_forecast -
preSampleHorizon_days(2)
 end_w2_preSamp = start_w2_preSamp+preSampleHorizon_days(2)

 start_w2_forecast = end_w2_preSamp +1
 end_w2_forecast = start_w2_forecast +horizon_days(2)-1

 [f2,YMSE2]=forecast(mdl,horizon_days(2),'Y0',dataInputForecast.mean_potential
_sampled(start_w2_preSamp:end_w2_preSamp),'XF', ...

[dataInputForecast.Tair(start_w2_forecast:end_w2_forecast),dataInputForecast.V
PD(start_w2_forecast:end_w2_forecast), ...

dataInputForecast.mean_temp_sampled(start_w2_forecast:end_w2_forecast),dataInp
utForecast.Prcp(start_w2_forecast:end_w2_forecast)]);

 % Fenêtre 3

 start_w3_preSamp = translation_w3 + end_w2_forecast -
preSampleHorizon_days(3)
 end_w3_preSamp = start_w3_preSamp+preSampleHorizon_days(3)


 start_w3_forecast = end_w3_preSamp+1
 end_w3_forecast = start_w3_forecast+horizon_days(3)-1

 [f3,YMSE3]=forecast(mdl,horizon_days(3),'Y0',dataInputForecast.mean_potential
_sampled(start_w3_preSamp:end_w3_preSamp),'XF', ...
```

```
[dataInputForecast.Tair(start_w3_forecast:end_w3_forecast),dataInputForecast.V
PD(start_w3_forecast:end_w3_forecast), ...

dataInputForecast.mean_temp_sampled(start_w3_forecast:end_w3_forecast),dataInp
utForecast.Prcp(start_w3_forecast:end_w3_forecast)]);


 % Nous allons définir les différents intervalles pour tracer les graphs
 % Pour le potentiel observé, tout l'intervalle est considéré

 X_obs = 1:height(dataInputForecast);
 Y_obs = dataInputForecast.mean_potential_sampled;


 % On définit le potentiel observé pour chaque fenêtre considérée
 Y_obs_w1=dataInputForecast.mean_potential_sampled(start_w1_forecast:end_w1_fo
recast);
 Y_obs_w2=dataInputForecast.mean_potential_sampled(start_w2_forecast:end_w2_fo
recast);
 Y_obs_w3=dataInputForecast.mean_potential_sampled(start_w3_forecast:end_w3_fo
recast);

 %**********Première fenêtre*******************

 X1forecast=start_w1_forecast:end_w1_forecast;

 Y1_forecast = f1;

 %**********Deuxième fenêtre*******************

 X2forecast=start_w2_forecast:end_w2_forecast;

 Y2_forecast = f2;

 %**********Troisième fenêtre*******************

 X3forecast=start_w3_forecast:end_w3_forecast;

 Y3_forecast = f3;


 % 90% confidence intervals window 1
 Y1_forecast_plus = f1 + 1.64*sqrt(YMSE1);
 Y1_forecast_minus = f1 - 1.64*sqrt(YMSE1);
```

```
% 90% confidence intervals window 2
Y2_forecast_plus = f2 + 1.64*sqrt(YMSE2);
Y2_forecast_minus = f2 - 1.64*sqrt(YMSE2);

% 90% confidence intervals window 3
Y3_forecast_plus = f3 + 1.64*sqrt(YMSE3);
Y3_forecast_minus = f3 - 1.64*sqrt(YMSE3);

%**************************************************

% Seuils pour les 3 états du potentiel
% Paramètres par défaut :
% First zone between 0 and 20 kPa (Small)
% Second zone between 20 and 80kPa (Medium)
% Third zone between 80 and 200 (High)
y1 = 0;
y2 =20;
y3=80;
y4=200;

%*****************Figures************************

figure(g1);

% Plot forecasted and observed potential
plot(X_obs,Y_obs,'b','LineWidth',2);

hold all

plot(X1forecast,Y1_forecast,'Color',[0.4660 0.6740
0.1880],'LineWidth',2,'LineStyle','--'); % couleur verte w1

plot(X2forecast,Y2_forecast,'Color',[0.9290 0.6940
0.1250],'LineWidth',2,'LineStyle','--'); % couleur orange w2

plot(X3forecast,Y3_forecast,'Color',[0.4940 0.1840
0.5560],'LineWidth',2,'LineStyle','--');      % couleur violet w3


% Plot confidence intervals
%plot(X1forecast,Y1_forecast_plus,'--','Color','#EDB120')

%plot(X1forecast,Y1_forecast_minus,'--','Color','#A2142F')


% add lines for SM_State (light grey)
```

```
 h1=yline(y1,'-','Color',[.7,.7,.7]);

 h2=yline(y2,'-','SWP State =
Small','Color',[.7,.7,.7],'LabelVerticalAlignment','bottom');

 h3=yline(y3,'-','SWP State =
Medium','Color',[.8,.8,.8],'LabelVerticalAlignment','bottom');

 h4=yline(y4,'-','SWP State =
High','Color',[.9,.9,.9],'LabelVerticalAlignment','bottom');


 % Legend
 year_measured=num2str(year_measured);
 yearLegend = sprintf('Measured year %s',year_measured);

 year_estimation=num2str(year_estimation);
 yearLegend1 = sprintf('Forecasted window 1, estimation %s',year_estimation);
 yearLegend2 = sprintf('Forecasted window 2, estimation %s',year_estimation);
 yearLegend3 = sprintf('Forecasted window 3, estimation %s',year_estimation);

 %legend(yearLegend,yearLegend1,yearLegend2,yearLegend3,'90% confidence
interval (+)','90% confidence interval (-)','Location','best')

 xlabel('days')
 ylabel('soil water potential [kPa]')
 legend(yearLegend,yearLegend1,yearLegend2,yearLegend3,'Location','best')


 title('Model ARIMAX with 4 exogenous variables (Tair, VPD, Tsoil, Prcp) and 3
sliding windows')

 hold off

 %*********************************************************
 % Calcul des residu_w1
 residu_w1=Y_obs_w1-Y1_forecast;

 % Calcul des residu_w2
 residu_w2=Y_obs_w2-Y2_forecast;

 % Calcul des residu_w3
 residu_w3=Y_obs_w3-Y3_forecast;


 %*******************************************************
```

```
% Calcul de MSE_w1 (first window)
% MSE = SSE/(nb points)
SSE_w1_sq=(Y_obs_w1-Y1_forecast).^2;
SSE_w1=sum(SSE_w1_sq,1);
MSE_w1=SSE_w1/length(Y_obs_w1);

RMSE_w1=sqrt(MSE_w1);

% Calcul de MSE_w2 (second window)
SSE_w2_sq=(Y_obs_w2-Y2_forecast).^2;
SSE_w2=sum(SSE_w2_sq,1);
MSE_w2=SSE_w2/length(Y_obs_w2);

RMSE_w2=sqrt(MSE_w2);


% Calcul de MSE_w3 (third window)
SSE_w3_sq=(Y_obs_w3-Y3_forecast).^2;
SSE_w3=sum(SSE_w3_sq,1);
MSE_w3=SSE_w3/length(Y_obs_w3);

RMSE_w3=sqrt(MSE_w3);

% Régression linéaire (valeurs prédites vs valeurs observées)
mdl1=fitlm(Y1_forecast,Y_obs_w1,'linear');
mdl2=fitlm(Y2_forecast,Y_obs_w2,'linear');
mdl3=fitlm(Y3_forecast,Y_obs_w3,'linear');

mdl_output={mdl1,mdl2,mdl3};

%*******************************************************
% Calcul de MAPE_w1_percent
mape1=sum(abs((Y_obs_w1-Y1_forecast)./Y_obs_w1),1);
MAPE_w1_percent=(100*mape1)/length(Y_obs_w1);

% Calcul de MAPE_w2_percent
mape2=sum(abs((Y_obs_w2-Y2_forecast)./Y_obs_w2),1);
MAPE_w2_percent=(100*mape2)/length(Y_obs_w2);

% Calcul de MAPE_w3_percent
mape3=sum(abs((Y_obs_w3-Y3_forecast)./Y_obs_w3),1);
MAPE_w3_percent=(100*mape3)/length(Y_obs_w3);

%*****************
% Output
```

```
 MSE_w=[MSE_w1,MSE_w2,MSE_w3];

 RMSE_w=[RMSE_w1,RMSE_w2,RMSE_w3];

 MAPE_w_percent=[MAPE_w1_percent,MAPE_w2_percent,MAPE_w3_percent];


 %*******************************************************
 % Graphiques des résidus

 figure(g2)
 yline(0);
 hold all
 scatter(X1forecast,residu_w1,'filled','MarkerFaceColor',[0.4660 0.6740
0.1880]); % couleur verte w1

 scatter(X2forecast,residu_w2,'filled','MarkerFaceColor',[0.9290 0.6940
0.1250]); % couleur orange w2

 scatter(X3forecast,residu_w3,'filled','MarkerFaceColor',[0.4940 0.1840
0.5560]);      % couleur bleue w3
 xlabel('days');
 ylabel('residual');
 hold off

 % Graphiques des valeurs du potentiel prédites vs valeurs observées
 figure(g3)
 scatter(Y1_forecast,Y_obs_w1,'filled','MarkerFaceColor',[0.4660 0.6740
0.1880])
 xlabel('Predicted values')
 ylabel('Mesured values')
 title('First window')

 figure(g4)
 scatter(Y2_forecast,Y_obs_w2,'filled','MarkerFaceColor',[0.9290 0.6940
0.1250])
 ylabel('Mesured values')
 title('Second window')

 figure(g5)
 scatter(Y3_forecast,Y_obs_w3,'filled','MarkerFaceColor',[0.4940 0.1840
0.5560])
 xlabel('Predicted values')
 ylabel('Mesured values')
 title('Third window')
```

```
%*****************************************************
%Graphiques des régressions linéaires
figure(g6);
plot(mdl1);
xlabel('predicted values');
ylabel('observed values');
title('First window');

figure(g7);
plot(mdl2);
xlabel('predicted values');
ylabel('observed values');
title('Second window');

figure(g8);
plot(mdl3);
xlabel('predicted values');
ylabel('observed values');
title('Third window');

%*********************************************************
% Prédiction sur les états du potentiels, sur chaque fenêtre
% Les états doivent être des variables catégorielles

% Appel de la fonction [data_output] =
level_potential_3levels(data_input,S0_inf,S0_sup,S1_inf,S1_sup)
% Cette fonction permet de déterminer l'état du potentiel selon la mesure
% du potentiel hydrique
% Les seuils sont précisés dans les paramètres d'entrée de la fonction

[state_pot_w1_obs] = level_potential_3levels(Y_obs_w1,0,20,20,80);

[state_pot_w1_forecast] = level_potential_3levels(Y1_forecast,0,20,20,80);

state_array_obs_w1=string(state_pot_w1_obs);

state_array_forecast_w1=string(state_pot_w1_forecast);

compare_w1=strcmp(state_array_obs_w1,state_array_forecast_w1);

somme_true_w1=sum(compare_w1==1);

% Pourcentage d'exactitude de la prédiction (fenêtre 1)
pourcentage_accuracy_state_w1=100*(somme_true_w1)/height(state_array_obs_w1);
```

```
%********
[state_pot_w2_obs] = level_potential_3levels(Y_obs_w2,0,20,20,80);

[state_pot_w2_forecast] = level_potential_3levels(Y2_forecast,0,20,20,80);

state_array_obs_w2=string(state_pot_w2_obs);

state_array_forecast_w2=string(state_pot_w2_forecast);

compare_w2=strcmp(state_array_obs_w2,state_array_forecast_w2);

somme_true_w2=sum(compare_w2==1);

% Pourcentage d'exactitude de la prédiction (fenêtre 2)
pourcentage_accuracy_state_w2=100*(somme_true_w2)/height(state_array_obs_w2);


%******************

[state_pot_w3_obs] = level_potential_3levels(Y_obs_w3,0,20,20,80);

[state_pot_w3_forecast] = level_potential_3levels(Y3_forecast,0,20,20,80);

state_array_obs_w3=string(state_pot_w3_obs);

state_array_forecast_w3=string(state_pot_w3_forecast);

compare_w3=strcmp(state_array_obs_w3,state_array_forecast_w3);

somme_true_w3=sum(compare_w3==1);

% Pourcentage d'exactitude de la prédiction (fenêtre 3)
pourcentage_accuracy_state_w3=100*(somme_true_w3)/height(state_array_obs_w3);

%************calcul precision, recall et f1-score**********************
% On va déterminer les matrices de confusion à partir des valeurs observées
% et prédites, pour chaque fenêtre
% Appel de la fonction Matlab confusionmat()
[m1]=confusionmat(state_array_obs_w1,state_array_forecast_w1);

[m2]=confusionmat(state_array_obs_w2,state_array_forecast_w2);

[m3]=confusionmat(state_array_obs_w3,state_array_forecast_w3);

m={m1;m2;m3};
```

```
 % À partir des trois matrices, nous allons déterminer la précision, le
 % recall et le score F1
 nb_windows=3;

 for i=1:nb_windows
     conf_mat=cell2mat(m(i));

     term_diagonal=diag(conf_mat);
     sum_rows=sum(conf_mat,2);
     precision=term_diagonal./sum_rows;
     overall_precision=mean(precision,'omitnan');

     sum_col=sum(conf_mat,1);
     recall=term_diagonal./sum_col';
     overall_recall=mean(recall,'omitnan');


F1_score=2*((overall_precision*overall_recall)/(overall_precision+overall_reca
ll));

     result_precision{i}=overall_precision;
     result_recall{i}=overall_recall;
     result_f1score{i}=F1_score;

 end % for

 %*****************Matrices de confusion : graphiques*******************
 figure(g9)
 cm_w1 =
confusionchart(state_array_obs_w1,state_array_forecast_w1,'Title','First
window','RowSummary','row-normalized','ColumnSummary','column-normalized');

 figure(10)
 cm_w2 =
confusionchart(state_array_obs_w2,state_array_forecast_w2,'Title','Second
window','RowSummary','row-normalized','ColumnSummary','column-normalized');

 figure(11)
 cm_w3 =
confusionchart(state_array_obs_w3,state_array_forecast_w3,'Title','Third
window','RowSummary','row-normalized','ColumnSummary','column-normalized');

 %*********************************OUTPUT********************
```

```
 accuracy_output=[pourcentage_accuracy_state_w1,pourcentage_accuracy_state_w2,
pourcentage_accuracy_state_w3];

 precision_output=[result_precision{1},result_precision{2},result_precision{3}
];

 recall_output=[result_recall{1},result_recall{2},result_recall{3}];

 F1_score_output=[result_f1score{1},result_f1score{2},result_f1score{3}];


 end %function
```

Fonction ARIMAX_MW_p14_q4_4X, appelée par la fonction ARIMAX_forecast_slidingwindow_error()

```
 function ARIMAX_mean_potential_sampled_spMW =
ARIMAX_MW_p14_d1_q4_4X(data,preSampleEstimation,SampleEstimation)
 % Evaluation of the ARIMAX model for MW
 % 4 exogenous variables : Tair, VPD, Tsoil, Prcp
 % (Use the names given in the project)
 % Input :
 % inputdata (format : time serie)
 % number of days for the preSampling interval
 % if preSampleEstimation==0 then the minimum number of points is taken equal
to (p+d)
 % Otherwise, the number of days can be specified
 % Number of days for the model estimation
 % if SampleEstimation==0 then all the input data will be considered for the
 % estimation
 % Output : estimation of the ARIMAX model

 mean_potential_sampled_spMW = data.mean_potential_sampled;

 %Quatre variables exogènes
 Tair = data.Tair;
 VPD = data.VPD;
 mean_temp_sampled = data.mean_temp_sampled;
 Prcp = data.Prcp;

 % Création du modèle mathématique ARIMAX(12,1,4)
 % Appel de la fonction Matlab arima()
 ARIMAX_mean_potential_sampled_spMW =
arima('Constant',0,'ARLags',1:14,'D',1,'MALags',1:4,'Distribution','Gaussian')
;

 validIndices =
find(~any(isnan([mean_potential_sampled_spMW,Tair,VPD,mean_temp_sampled,Prcp])
,2));

 if  preSampleEstimation==0 & SampleEstimation==0
     preSampleNumber = ARIMAX_mean_potential_sampled_spMW.P; % p value
     preSampleResponse =
mean_potential_sampled_spMW(validIndices(1:preSampleNumber));
     estimateResponse =
mean_potential_sampled_spMW(validIndices(preSampleNumber+1:end));
 elseif preSampleEstimation~=0 & SampleEstimation==0
     preSampleNumber = preSampleEstimation;
```

```
     preSampleResponse =
mean_potential_sampled_spMW(validIndices(1:preSampleNumber));
     estimateResponse =
mean_potential_sampled_spMW(validIndices(preSampleNumber+1:end));
 else
     preSampleNumber = preSampleEstimation;
     preSampleResponse =
mean_potential_sampled_spMW(validIndices(1:preSampleNumber));
     estimateResponse =
mean_potential_sampled_spMW(validIndices(preSampleNumber+1:SampleEstimation));
 end

 % Estimation du modèle sur l'intervalle
%[preSampleResponde+1:end_data_input]
 % Appel de la fonction Matlab estimate()
 ARIMAX_mean_potential_sampled_spMW =
estimate(ARIMAX_mean_potential_sampled_spMW,estimateResponse,'Y0',preSampleRes
ponse,'X',[Tair,VPD,mean_temp_sampled,Prcp],'Display','off');

 %summarize(ARIMAX_mean_potential_sampled_spMW)

 % On détermine les résidus
 % Appel de la fonction Matlab infer()
 residual =
infer(ARIMAX_mean_potential_sampled_spMW,data.mean_potential_sampled);

 % On définit 2 figures
 g1=figure;
 g2=figure;

 % ****** Figure residus ********

 clf
 figure(g1);
 subplot(2,2,1)
 plot(residual./sqrt(ARIMAX_mean_potential_sampled_spMW.Variance))
 title('Standardized Residuals')
 subplot(2,2,2)
 qqplot(residual)
 subplot(2,2,3)
 autocorr(residual)
 subplot(2,2,4)
 parcorr(residual)

 hvec = findall(gcf,'Type','axes');
 set(hvec,'TitleFontSizeMultiplier',1.2,...
```

```
    'LabelFontSizeMultiplier',1.2);

%************************************


potential_fit = data.mean_potential_sampled - residual;


%*********Figure fit *********************

figure(g2);
plot(data.mean_potential_sampled,'b','LineWidth',2)
hold on
plot(potential_fit ,'--','Color','#D95319','LineWidth',2)
legend(['SWP measured'],['SWP fitted'])
hold off

%***************************************

end %function
```