Laboratoire des Sciences du Climat
et de l'Environnement

Upscaling et Downscaling dans la modélisation hydrologique

Auteur

Mathis Deronzier
Mines Saint-Étienne

Maîtres de stage

Emmanuel Mouche
C.E.A.
Mathieu Vrac
C.E.A.

Stage de recherche de master 2

Avril-Septembre
2021

Table des matières

1 Downscaling

Le downscaling (voir par exemple Vrac et al. (2012) et Ayar et al. (2016)) est une méthode statistique utilisé dans les sciences du climat permettant d'améliorer les modèles de prédiction. À partir des données obtenues par un modèle de simulation (modèle de circulation général, modèle climatique régional) et des données réelles on cherche à corriger les biais systématiques. Le nom "downscaling" vient du fait que l'on cherche souvent à faire des prédictions sur un point particulier du domaine prédit par le modèle de prédiction. Cette méthode est très utile dans la pratique où l'on a des simulateurs donnant des informations sur des maillage de grande distance de grille ($\sim 200km$). Dans notre cas, nous utilisons le downscaling pour prédire les variables climatiques de *précipitation* et d'*évapotranspiration* sur le bassin du **Little Washita** ($\sim 30 \times 30km^2$). Nous avons testé nos méthodes à partir de deux jeux de données : les données NARRs (voir section 4.1) ainsi que celles de l'IPSL.

Pour formuler rigoureusement l'approche du downscaling nous introduisons des hypothèses communément admises dans les sciences du climat. On suppose que les variables étudiées sont des variables aléatoires dépendantes du temps et de l'espace. On appelle $\mathcal{M}(\Omega, \mathbb{R})$ l'espace des variables aléatoires réelles et $S(\mathbb{R})$ la spère unité dans $\mathbb{R}^3$, on suppose de plus que l'on peut faire correspondre chaque point de la terre à un point de la sphère unité.

Définition 1. *Pour une variable quantitative V à valeur dans $\mathbb{R}$, on appelle $\mathcal{T}_V$ la fonction donnant les valeurs réelle de cette variable sur la terre à un moment donné, formellement (en considérant la terre comme une sphère $\mathcal{S}(\mathbb{R}^3)$) nous avons*

$$\mathcal{T}_V : \mathbb{R}_+ \times \mathcal{S}(\mathbb{R}^3) \quad \rightarrow \quad \mathcal{M}(\Omega, \mathbb{R}). \tag{1}$$

Alors, $\mathcal{T}_V(t, x)$ est la valeur de la variable au temps t au point de coordonnée x sur terre.

Définition 2. *On appelle simulateur de variable quantitative V à valeur dans $\mathbb{R}$, une fonction S_V satisfaisant :*

$$S_V : \mathbb{R}_+ \times \mathcal{S}(\mathbb{R}^3) \quad \rightarrow \quad \mathcal{M}(\Omega, \mathbb{R}). \tag{2}$$

On peut alors estimer la qualité des simulations en mesurant une distance entre la réalisation $\mathcal{T}_V([0, T])$ et celle de $S_V([0, T])$. Le travail du downscaling est de minimiser ces distances. Les géostatistiques essaie de munir ces fonctions de structures particulières pour améliorer les résultats des prédictions (voir par exemple Lindgren et al. (2011)).

1.1 Introduction à la problématique du downscaling

Le downscaling consiste à effectuer des transformations sur des variables aléatoire réelles. Pour étudier ces variables aléatoires on introduit ici les notions de **fonction de répartition** et de **fonction de répartition empirique** ainsi que celle de **fonction de densité**. Ces notions sont centrales dans les méthodes de downscaling que nous allons expliquer.

Définition 3. *Soit X une variable aléatoire réelle, on appelle **fonction de répartition de** X, $\mathcal{F}_X : \mathbb{R} \rightarrow [0, 1]$ la fonction vérifiant*

$$\mathcal{F}_X(x) = P(X \leqslant x). \tag{3}$$

Définition 4. *Soient $X_1, X_2, ..., X_n$, n réalisations indépendantes d'une variable aléatoire réelle X, on appelle **fonction de répartition empirique de** X, la fonction $\mathcal{F}_n : \mathbb{R} \rightarrow [0, 1]$ définie par*

$$\mathcal{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[X_i, +\infty)}(x). \tag{4}$$

Définition 5. *Soit X une variable aléatoire réelle, on note f_X la **fonction de densité de** X, $f_X : \mathbb{R} \rightarrow [0, \infty[$ la fonction vérifiant*

$$f_X(x) = \lim_{t \to 0} \frac{P(x \leqslant X \leqslant x + t)}{t}. \tag{5}$$

Remarquons qu'une variable aléatoire ne possède pas nécessairement de fonction de répartition (notamment les variables aléatoires à valeurs discrètes). On peut cependant les étudier dans la théorie des distributions.

Commençons par établir notre problématique dans le cas le plus simple où l'on cherche à prédire une variable V dans l'avenir alors que nous connaissons ses réalisations dans le passé à un endroit donné x. On peut alors considérer deux processus aléatoires à valeurs dans $\mathbb{R}$, $(X_t)_{t \in \mathbb{N}} = (\mathcal{S}_V(t, x))_{t \in \mathbb{N}}$ et $(Y_t)_{t \in \mathbb{N}} = (\mathcal{T}_V(t, x))_{t \in \mathbb{N}}$.

La problématique à laquelle nous cherchons de répondre est la suivante : connaissant $X_1, X_2, ..., X_n$ et $Y_1, Y_2, ..., Y_n$ les réalisations jusqu'au temps n ainsi que $X_{i_1}, X_{i_2}, ..., X_{i_m}$, on cherche une fonction $G : \mathbb{R} \to \mathbb{R}$ telle que les tirages $G(X_{i_1}), ..., G(X_{i_m})$ et $Y_{i_1}, ..., Y_{i_m}$ soient proches du point de vu de leur loi (nous éclaircirons ce point par la suite voir la section 2).

Autrement dit, en appelant X et Y les réalisations de $(X_t)_{t\in\mathbb{N}}$ et $(Y_t)_{t\in\mathbb{N}}$ sur $\{1, ..., n\}$ et X' et Y' les réalisations de $(X_t)_{t\in\mathbb{N}}$ et de $(Y_t)_{t\in\mathbb{N}}$ sur $\{i_1, ..., i_m\}$, on cherche à définir $G_{X,Y}$ à partir de X, Y tel que $G_{X,Y}$ minimise

$$d(\mathcal{F}_{G_{X,Y}(X')}, \mathcal{F}_{Y'}),$$

où d est une distance définie sur les fonctions. Nous voulons aussi que $G_{X,Y}$ respecte certaines propriétés. Une des propriété qui nous intéresse est celle de la consistance de notre transformation.

Définition 6. *Soient X et Y deux variables aléatoires réelles et $G_{X,Y} : \mathbb{R} \to \mathbb{R}$ une transformation, on dit que $G_{X,Y}$ est **consistante vis à vis de X et de Y** si elle vérifie*

$$\mathcal{F}_{G_{X,Y}(X)} = \mathcal{F}_Y. \tag{6}$$

Dans la suite les transformations que nous considérerons satisferont toujours l'équation (6).

1.2 Cumulative Distribution Function transform (CDFt)

Nous allons ici présenter l'algorithme principalement étudié et utilisé dans ce stage, l'algorithme CDF-t. Nous commencerons par présenter l'algorithme de quantile-quantile permettant de comprendre l'esprit des transformations G affectées aux processus aléatoires. Puis nous décrirons l'algorithme de CDFt-t.

1.2.1 Quantile-Quantile

Le quantile-quantile consiste simplement à définir $G_{X,Y}$ la transformation permettant de passer de la fonction de distribution de X à celle de Y.

Proposition 1. *Soit X et Y deux variables aléatoires réelles ayant des fonctions de répartition $\mathcal{F}_X$ et $\mathcal{F}_Y$ continues, alors $\mathcal{F}_Y^{-1}(\mathcal{F}_X(X))$ et Y suivent la même loi.*

Démonstration. Montrons que $\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X)$ et Y possède la même fonction de densité.

$$\mathcal{F}_{\mathcal{F}_Y^{-1}\circ\mathcal{F}_X(X)}(y) = \mathbb{P}(\mathcal{F}_Y^{-1}(\mathcal{F}_X(X)) \leqslant y) = \mathbb{P}(\mathcal{F}_X(X) \leqslant \mathcal{F}_Y(y)),$$

comme $\mathcal{F}_X(X)$ suit une loi uniforme sur $[0, 1]$ si F_X est continue cette égalité se réécrit

$$= \mathbb{P}(\mathcal{U}(0, 1) \leqslant \mathcal{F}_Y(y)) = \mathcal{F}_Y(y).$$

□

Le principe de l'algorithme **quantile-quantile** est de calculer la transformation $G = \mathcal{F}_Y^{-1} \circ \mathcal{F}_X$. Ainsi, $\mathcal{F}_{G(X)} = \mathcal{F}_Y$ et l'on définit alors $G_{X,Y} = G$. L'une des limites de cette méthode est que le support de $f_{G(X)}$ est inclus dans celui de f_Y, alors les valeurs prises par $G_{X,Y}(X')$ seront incluses dans le support de f_Y. Nous aimerions que X' ait aussi une influence sur le support des valeurs prises. C'est à dire que l'on considère que le support de X est aussi porteur d'information.

1.2.2 CDFt

L'algorithme de **Cumulative Distribution Function transfer** (CDFt) vise à remédier au problème des bornes en appliquant des transformations sur les lois X et Y.

CDFt avec régression linéaire :

Supposons que l'on ait des fonctions estimant la variance et de la moyenne de Y' à partir de X', s'exprimant sous la forme $\overline{f(X')} = \overline{Y'}$ ainsi que $\overline{g(X')} = \overline{\sigma(Y')}$ (on suppose ici X' et Y' comme des suites de variables aléatoires) [1]. On pose alors une condition supplémentaire sur $G_{X,Y}$, imposant la conservation de la moyenne et

1. On définit pas $\overline{X}$ la moyenne X et par $\sigma(X)$ son écart type

de la variance. Formellement, on voudrait définir $G_{X,Y}$ telle que quelque soient $\{i_1, ..., i_m\}$ un ensemble d'entiers consécutifs et X' et Y' les vecteurs des variables aléatoires des X_{i_j} et Y_{i_j} sur cet ensembles, on ait :

$$\overline{G_{X,Y}(X')} = \overline{Y'}, \tag{7}$$

ainsi que

$$\sigma(G_{X,Y}(X')) = \sigma(Y'), \tag{8}$$

et que G respecte la condition de consistance (6). Concrêtement, nous allons faire des transformations sur les variables aléatoires pour avoir ces conditions là. En appelant $G_{X,Y} = \mathcal{F}_Y^{-1} \circ \mathcal{F}_X$ on peut définir la transformation $G_{X,Y,X'}$

$$G_{X,Y}(X') = \overline{g(X')} \frac{\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X') - \overline{\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X')}}{\sigma(\mathcal{F}_Y^{-1} \circ \mathcal{F}_X(X'))} + \overline{f(X')}.$$

On peut vérifier facilement que $G_{X,Y}$ ainsi défini respecte bien les propriétés (6) (7) et (8). L'idée est alors de trouver les fonctions f et g par des méthodes de regression linéaires. Remarquons aussi que lorsque la loi est bornée cela ajoute une condition supplémentaire à $G_{X,Y}$ et l'on ne peut pas nécessairement avoir les conditions sur la moyenne et la variance. Il faut alors choisir parmi l'une des conditions (7) et (8), c'est par exemple le cas pour les précipitations qui ne peuvent être négatives.

Note 1. *Remarquons que l'efficacité des transformations que nous avons décrit repose en partie sur la stationnarité des lois suivies par les variables aléatoires aléatoires au cours du temps. Cette question a été abordée dans Maraun (2012), Christensen et al. (2008) ainsi que Nahar et al. (2017).*

1.3 Transport optimal

Remarquons que nous voulons à la fois prédire la précipitation et l'évapotranspiration, on peut alors considérer une seule variable aléatoire dans $\mathbb{R}^2$. On peut généraliser l'idée utilisée précédemment pour trouver une méthode permettant de corriger les biais statistiques introduits par les modèles de prédictions. Cette méthode a d'autant plus d'intérêt que la variable utile dans les modèles hydrologique est

$$\textit{pluie entrant dans le sol} = \textit{précipitation} - \textit{évapotranspiration}.$$

Comme l'objectif final est de prédire des résultats hydrologiques sur le bassin du Little Washita, il semble particulièrement pertinent de considérer la loi conjointe précipitation, évapotranspiration. La théorie généralisant cette idée est la théorie du **transport optimal**.

La problématique du transport optimal a premièrement été introduite en par Gaspard Monge en 1781 puis a été développée par Kantorovitch en 1971 et ses travaux pour l'allocution des ressources lui ont valu un prix nobel d'économie en 1975.

1.3.1 Problématique

Précédemment nous avons cherché à définir une transformation G de la variable aléatoire X telle $G(X)$ suive la même loi que Y. En considérant les fonctions de densité de X et de Y on a cherche à ce que la fonction de densité de $G(X)$ soit la même que celle de Y. On peut considérer une fonction de densité comme une mesure sur l'espace sur lequel on travail on a alors transformé une mesure f_X en une autre mesure f_Y. Comme ces deux mesures sont de mesure totale égale à 1. On peut dire que d'une certaine manière chaque "poids" de la mesure f_X a été déplacé vers un poids de la mesure f_Y. L'idée du transport minimal est de trouver la description des déplacements totales des poids, il vise à trouver des manières "naturelles" de déplacer ces poids.

Nous présenterons ici la formulation établie par Kantorovitch dans les années 70 qui a l'avantage d'inclure celle de Monge. Le livre Villani (2003) donne un cours assez complet sur les problématiques de transport optimal.

Considérons deux fonctions de répartitions pour des variables aléatoires U et V à valeurs dans X et Y, on appelle ces fonctions $\mathcal{F}$ et $\mathcal{G}$ et on appelle f et g leurs fonctions de densités. On cherche alors une mesure π sur $X \times Y$ satisfaisant

$$\int_Y d\pi(x, y) = f(x), \quad \int_X d\pi(x, y) = g(y),$$

de plus on veut que π satisfaisant l'équation précédente minimise la quantité

$$\mathcal{I}[\pi] = \int_{X\times Y} d(x,y)\, d\pi(x,y),$$

où d est une certaine distance définissant le coût de transport de x à y. Dans notre cas U et V sont des variables aléatoires à valeurs dans $\mathbb{R}^2$. Nous voyons que le choix de la distance a une influence majeure sur la mesure obtenue. Alors on peu voir $d\pi(x,y)$ comme la quantité déplacée de x à y.

1.3.2 Résolution du problème dans le cas fini et downscaling

La résolution de ce problème dans le cas fini a été traité de nombreuses fois. On utilisera les idées développées dans le papier Robin et al. (2019). Appelons $\pi \in \mathbb{R}^{m\times n}$ une matrice de transfert de poids dans le cas fini. On a $X_1, ..., X_n$ ainsi que $Y_1, ..., Y_m$ des réalisations de X et de Y et on cherche alors une matrice π telle que

$$\sum_{j=1}^{n} \pi_{i,j} = P(X = X_i) \text{ et } \sum_{i=1}^{m} \pi_{i,j} = P(Y = Y_j),$$

avec π minimisant

$$\mathcal{I}[\pi] = \sum_{i,j} d(X_i, Y_j)\pi_{i,j}.$$

Le papier utilise la norme euclidienne comme distance, l'obtention de cette solution peut se faire par un algorithme de simplexe, voir par exemple Huang and Chen (2012). Pour corriger le biais d'estimation on peut alors pour chaque x tirés récupérer (par interpolation linéaire ou méthode de krigeage) le $\pi(x,\cdot)$. D'après la construction de π, on peut alors normaliser la fonction $y \mapsto \pi(x,y)$ et tirer aléatoirement un point selon la loi ainsi trouvée (voir Robin et al. (2019) pour plus de détails).

2 Analyse des résultats obtenus par downscaling

Concrètement nous allons faire de la validation croisée, nous allons apprendre sur 50% de nos données et faire nos prédictions. La partie à laquelle nous allons nous intéresser ici est la distance que nous utilisons pour évaluer nos prédictions. Contrairement à la manière habituelle de faire, consistant à estimer une distance entre chaque point prédit (souvent RMSE), en climatologie nous cherchons à comprendre la tendance générale. En effet, le paradigme d'évaluation en prévisions climatiques sur plusieurs années n'a pas l'ambition de prédire ponctuellement chaque prévision, mais il a pour objectif de décrire une tendance générale. On s'intéresse alors à des informations plus générales, c'est à dire que l'on travaille sur les lois de répartitions. Il faut alors réfléchir à des normes ou des distance pour évaluer la qualité de nos prédictions.

2.1 Tests basés sur les fonctions de répartition empiriques

Dans notre cas nous faisons des tests non-paramétriques, c'est à dire que l'on ignore tout des lois que nous comparons. Différentes méthodes pour tester l'égalité de lois sont connues, nous n'en développerons que deux. Le mémoire Éthier (2011) donne une présentations de principaux tests statistiques permettant d'évaluer si oui ou non à partir des réalisations $X_1, ..., X_n$, $Y_1, ..., Y_n$ de deux lois inconnues sont les mêmes.

Nous posons habituellement en statistiques deux hypothèses :

$$\mathcal{H}_0 : \mathcal{F}_X = \mathcal{F}_Y \quad et \quad \mathcal{H}_1 : \mathcal{F}_X \neq \mathcal{F}_Y,$$

où les égalités sur les lois sont en norme L^p. On suppose $\mathcal{H}_0$ et on définit une statistique sur $\|\mathcal{F}_X - \mathcal{F}_Y\|_{L^p(\mathbb{R})}$ permettant à partir de nos observations d'accepter ou de rejeter l'hypothèse $\mathcal{H}_0$. Les tests de Kolmogorov-Smirnov et Cramer-von Mises utilisent à-peu-près cette idée.

2.1.1 Quelques outils mathématiques

Proposition 2. *Soient $X_1, ..., X_n$ n réalisations d'une variable aléatoire réelle X et $\mathcal{F}_n$ sa fonction de répartition empirique nous avons*

$$E[\mathcal{F}_n] = F_X,$$

alors la fonction de répartition empirique est un estimateur sans biais de la lois de F.

Démonstration. C'est en effet évident puisque $\mathbb{1}_{[X,+\infty)}(x)$ suit une lois de Bernoulli de paramètre $\mathcal{F}(x)$ alors

$$E\Big[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[X_i,+\infty)}(x)\Big] = \frac{1}{n}\sum_{i=1}^{n}E[\mathbb{1}_{[X_i,+\infty)}] = \mathcal{F}_X(x).$$

□

Théorème 2.1. *(Glivenko-Cantelli) Soient $\mathcal{F}_X$ et $\mathcal{F}_n$ respectivement la fonction de répartition et la fonction de répartition empirique. Alors*

$$\|\mathcal{F}_X - \mathcal{F}_n\|_\infty \xrightarrow[n\to\infty]{prob} 0 \tag{9}$$

Démonstration. (Cas où $\mathcal{F}_X$ est continue) On commence par remarquer que quelque soit x dans $\mathbb{R}$,

$$F_n(x) \xrightarrow[n\to\infty]{p.s.} \mathcal{F}_X(x)$$

d'après la loi forte des grands nombres et la proposition (2). Pour q dans $\mathbb{Q}$, on définit

$$\Omega_q = \{\omega \in \Omega | \lim_{n\to\infty} \mathcal{F}_n(q) = \mathcal{F}_X(q)\},$$

d'après ce que nous avons dit, sa mesure pour la probabilité $P(\Omega_q) = 1$ comme $\mathbb{Q}$ est dénombrable nous avons

$$P\Big(\bigcap_{q\in\mathbb{Q}}\Omega_q\Big) = 1.$$

Alors, comme $\mathbb{Q}$ est dense dans $\mathbb{R}$ et que $\mathcal{F}_X$ et les $(\mathcal{F}_n)_{n\in\mathbb{N}}$ sont continues on peut assurer que

$$\|\mathcal{F}_X - \mathcal{F}_n\|_\infty \xrightarrow[n\to\infty]{prob} 0.$$

□

Le cas où $\mathcal{F}_X$ n'est pas continue est géré par Durrett (2019) (ex 7.2 chap 1). On voit d'après le théorème 2.1 que la fonction de répartition empirique est le bon estimateur de la fonction de répartition.

2.1.2 Kolmogorov-Smirnov

Le test d'ajustement de Kolmogorov-Smirnov est l'un des plus utilisé pour tester l'égalité de deux lois de probabilités. Dans le contexte de l'égalité de lois de probabilité, la statistique de test est

$$K_{n,m} = \sqrt{\frac{nm}{n+m}}\|\mathcal{F} - \mathcal{F}_n\|_\infty.$$

La suite de variables aléatoires $K_{n,m}$ converge vers une variable aléatoire K dont la fonction de survie est donnée par :

$$Q(x) = P(K > x) = \sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2(jx)^2) \tag{10}$$

On peut alors l'approximer avec les premiers termes de la série pour construire le test statistique. La démonstration de ce théorème peut être trouvée dans le livre Fisz (1963)(chap 12.5). Nous voyons cependant que ce test est sensible aux données aberrantes, nous privilégierons alors le test de **Cramér-von Mises**.

2.1.3 Cramér-von Mises

On considère ici deux fonctions de répartitions $\mathcal{F}$ et $\mathcal{G}$ continues. Nous voulons tester les hypothèses

$$\mathcal{H}_0 : \mathcal{F} = \mathcal{G} \quad \mathcal{H}_1 : \mathcal{F} \neq \mathcal{G},$$

Dans les tests proposés les statistiques sont construites à partir de deux échantillons indépendants (ce qui n'est pas notre cas). On définit aussi la fonction de répartition empirique $\mathcal{F}_n$.

Nous avons donc n réalisations de X et m réalisations de Y de lois de répartitions $\mathcal{F}$ et $\mathcal{G}$. La statistique du test est définie par

$$C_{n,m} = \frac{nm}{n+m} \int_{\mathbb{R}} \left[\mathcal{F}_n(x) - \mathcal{G}_m(x)\right]^2 \mathrm{d}\mathcal{H}_{m,n}(x), \tag{11}$$

avec,

$$\mathcal{H}_{n,m} = \frac{n}{n+m}\mathcal{F}_n + \frac{m}{n+m}\mathcal{G}_m. \tag{12}$$

$\mathcal{H}_{n,m}$ est alors la fonctions de répartitions empirique d'une variable aléatoire Z construite à partir des $n+m$ réalisations indépendantes $X_1, ..., X_n, Y_1, ..., Y_m$. On peut simplement réécrire la valeur $C_{n,m}$

$$C_{n,m} = \frac{mn}{(m+n)^2} \sum_{i=1}^{m+n} \left(\mathcal{F}_n(Z_i) - \mathcal{G}_m(Z_i)\right)^2 \tag{13}$$

Lemme 1. *On peut simplifier cette formule en supposant que les $(X_i)_{i\in\{1,...,n\}}$ et $(Y_i)_{i\in\{1,...,m\}}$ sont triés on a :*

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{4nm-1}{6(m+n)}. \tag{14}$$

où R_{X_i} est le rang de X_i dans $X_1, ..., X_n, Y_1, ..., Y_n$ autrement dit

$$R_{Z_i} = Card(\{z \in Z, z \leqslant Z_i\}).$$

Note 2. *La démonstration de cette formule se trouve dans l'indexe 1 et la formulation de cette égalité diffère de celle contenue dans le mémoire Éthier (2011)(sec 2.3.2) qui contient une erreur.*

Cette formulation a le bon goût de nous indiquer que la statistique ne dépend pas de la loi. On peut calculer simplement sa statistique sous l'hypothèse $\mathcal{H}_0$ pour la loi uniforme sur $[0, 1]$ et ainsi retrouver les quantiles présentés dans l'article de Büning (2002). Nous voyons que l'idée du test est aussi de pondérer la différence des fonctions de répartition empiriques par les observations (l'intégration selon $\mathcal{H}_{m,n}$). Ainsi, si le test de Kolmogorov-Smirnov est sensible aux outliers, celui-ci l'est beaucoup moins lorsque les échantillons sont suffisamment grands.

3 Upscaling des modèles hydrologiques

L'upscaling est un concept physique consistant à simplifier un modèle pour diminuer les calculs.

3.1 Généralités sur les interactions atmosphère - surface continentale - sol

Cette section reprend brièvement les principaux mécanismes entrant en compte dans les modélisations hydrologiques. Ces mécanismes seront présentés brièvement mais des sources seront données pour les lecteurs voulant étudier plus en détails ces mécanismes. Cette section s'inspire très largement de la thèse Maquin (2016) sur les modèles hydrologiques de colonne appliqués sur le bassin du Little Washita.

3.1.1 Les écoulements

Les processus hydrologiques sont classiquement étudiés à l'échelle du bassin versant (voir). En hydrologie, le bassin versant est une unité géographique définie par les limites topographiques que sont les lignes de crête. L'ensemble des écoulements converge vers les dépressions, formant ainsi un réseau hydrographique qui se dirige vers le point bas du bassin versant, l'exutoire.

Il possède son équivalent en mathématiques, soit E un espace métrique et $S : E \to E$ un endomorphisme sur E. On définit le bassin d'attraction d'un point a qu'on appelle $B(a)$ l'ensemble des points x dans E tels que la suite $(x_n)_{n\in\mathbb{N}} = (S^n(x))_{n\in\mathbb{N}}$ converge vers a. En considérant qu'il existe une fonction S définissant la trajectoire d'une goutte d'eau déposée au point x les deux définitions ont la même signification.

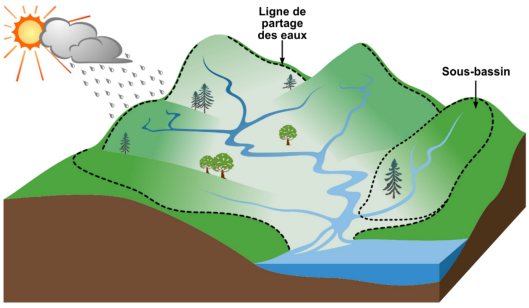


Figure 1 – Bassin versant (Source :http ://rqes-gries.ca/).

À l'échelle du bassin versant, on distingue deux types d'écoulements : les écoulements de subsurface, les écoulements de surface.

Les écoulements de subsurface :

La notion d'écoulement de subsurface se rapporte à l'écoulement de l'eau dans les pores du sol. L'écoulement de subsurface dépends de plusieurs paramètres comme les caractéristiques du sol(la porosité, la perméabilité), la saturation en eau du sol la topographie et le climat (précipitation, évaporation, transpiration). Ces écoulements sont traités par les équations de la mécanique des fluides (voir section à citer) et pour plus de détails De Marsily (1986).

Les écoulements de surface :

Les écoulements de surface, aussi qualifiés de ruissellement, sont la conséquence de deux phénomènes distincts. Le ruissellement peut apparaître lorsque le sol est saturé en surface. En effet, lorsque le sol est saturé, l'eau ne peut s'y infiltrer Cappus (1960). Cette condition de saturation à la surface du sol peut être la conséquence d'une nappe affleurant la surface, la zone satisfaisant cette propriété est appelée zone de suintement. Cela arrive naturellement lors d'épisode pluvieux pour les nappes peu profondes. Le ruissellement peut aussi être causé par de fortes précipitations, ainsi le débit surfacique peut devenir supérieur à la quantité d'infiltration et ainsi créer un ruissellement (voir la figure). La quantité d'infiltration décroit exponentiellement lors d'évènements pluvieux Horton (1933).

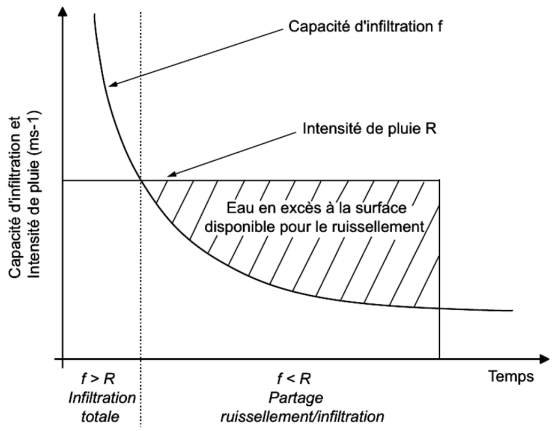


Figure 2 – Estimation du ruissellement en fonction du temps modèle de Horton

3.1.2 Transferts d'eau entre le sol et l'atmosphère

La végétation constitue le lien entre l'atmosphère et le sol. Les végétaux transfèrent de l'eau dans les deux sens, via les racines et la canopée. Il y a aussi des interactions directes entre le sol et l'atmosphère. Les trois principaux processus décrits sont l'évaporation du sol, la transpiration des végétaux ainsi que l'évaporation de l'eau interceptée par la canopée (mécanisme visant à conserver l'eau).

la transpiration :

La "transpiration" des plantes consiste en une libération de vapeur d'eau par les plantes dans l'atmosphère. Ce phénomène constitue une réponse passive à l'environnement atmosphérique dû à l'existence d'un gradient de pression positif de l'atmosphère à la canopée, on parle alors de demande atmosphérique. La description des processus d'évaporations ont été traités dans la thèse Maquin (2016).

Évaporation :

Sur les surfaces de sol non recouvertes de végétation (sol nu), l'eau présente dans le sol, à proximité de la surface, peut s'évaporer. Ce phénomène apparaît en présence d'un gradient de pression de vapeur d'eau entre le sol et l'atmosphère et d'un apport d'énergie. L'évaporation effective dépend de l'état hydrique de la surface du sol, l'énergie pour extraire l'eau du sol augmentant à mesure que le sol s'assèche et des propriétés conductrices du sol (voir Hillel (2003)).

Pertes par interception :

Lors d'un épisode pluvieux, une partie de l'eau incidente est interceptée par le feuillage. Il s'agit du phénomène dit d'interception. Cette eau présente sur la canopée peut ensuite s'évaporer directement. On désigne ce processus d'évaporation sur la canopée comme les pertes par interception. L'importance de ce flux d'eau dépend de l'ampleur du feuillage et de la capacité de stockage d'eau de la canopée, c'est-à-dire de l'épaisseur maximale de la lame d'eau par unité de surface de feuillage

l'évapotranspiration potentielle :

On désigne par « évapotranspiration potentielle » la quantité d'eau maximale que l'atmosphère peut extraire via les trois processus décrits précédemment. Elle correspond ainsi à la demande atmosphérique évoquée auparavant. L'évapotranspiration potentielle correspond à l'évaporation d'une surface saturée en eau. Elle dépend de paramètres atmosphériques comme l'humidité de l'air, le vent et la température. Ce taux potentiel a la propriété de majorer la somme des flux de transpiration, d'évaporation et des pertes par interception.

3.2 Les concepts hydrologiques

Nous allons ici introduire les principale notions à l'étude hydrologique des sols. Plusieurs caractéristiques définissent un sol, mais avant d'étudier en détail ce qui définit un sol, il est important de comprendre que l'étude hydro d'un sol est simplement un bilan d'eau dans celui-ci. Il faut alors commencer par déterminer ce qui rentre et ce qui sort. L'estimation de ces quantités est l'objet d'étude du downscaling qui cherche à prévoir la précipitation et l'évapotranspiration (l'eau drainée par les plantes recouvrant la surface).

3.2.1 Quelques définitions

Définition 7. *On appelle* ***porosité totale*** ω *la valeur définie par*

$$\omega = \frac{Volume\ des\ vides}{Volume\ total\ de\ la\ roche}. \tag{15}$$

On appelle aussi ***indice des vides*** e *la valeur définie par*

$$e = \frac{Volume\ des\ vides}{Volume\ du\ solide\ plein}. \tag{16}$$

On peut passe d'une formule à l'autre par la relation

$$e\omega = e - \omega,$$

Mais l'on utilise toujours la notion de porosité en hydrologie.

L'on peut trouver des méthodes de mesure de la porosité d'un sol dans l'ouvrage (Marsily). On dit aussi que le sol n'est pas saturé lorsque l'eau n'a pas pris tout l'espace disponible, on parle alors de satutation volumique.

Définition 8. *On parle de* ***saturation volumique*** θ*, la saturation définie par le rapport*

$$\theta = \frac{Volume\ d'eau\ contenu}{Volume\ total}, \tag{17}$$

on a $0 \leqslant < \theta \leqslant \omega$*. Et la* ***saturation volumique*** s

$$s = \frac{Volume\ d'eau\ contenu}{Volume\ total\ des\ pores}. \tag{18}$$

En fonction de la saturation volumique les échelles de temps et les forces misent en action ne sont pas les mêmes.

3.2.2 Les équations pour modéliser l'écoulement

On commence par rappeler les équations essentielles à la dynamiques des fluides. L'équation de de conservation de la matière :

$$div(\rho\overrightarrow{u}) + \frac{\partial\rho}{\partial t} = 0. \tag{19}$$

Où ρ est la masse volumique et $\overrightarrow{u}$ le vecteur vitesse du fluide. On écrit maintenant l'équation de Navier-Stokes

$$\frac{\partial p}{\partial x^i} - (\zeta + \frac{\mu}{3})\frac{\partial}{\partial x^i}(div\overrightarrow{u}) - \mu\nabla^2 u^i = \rho(F^i - \frac{du^i}{dt}), \tag{20}$$

ζ coefficient de viscosité du volume, (très souvent négligeable de vant mu) $[ML^{-1}T^{-1}]$,
μ coefficient de viscosité dynamique, $[ML^{-1}T^{-1}]$
∇^2 le laplacien,
F^i composante des forces à distance par unité de masse,
i un vecteur unitaire de l'espace $3D$.

En milieu poreux les équations de Navier-Stokes deviennent difficilement applicables car le milieu dans lequel s'écoule le fluide dépend lui-même de l'écoulement du fluide. On donne des hypothèse simplificatrices pour résoudre l'équation de Navier-Stokes.

Hypothèses simplificatrice :
On commence par supposer les écoulement permanents

$$\frac{\partial u^i}{\partial t} = 0,$$

on suppose aussi que

$$div(\rho \overrightarrow{u}) = \frac{\partial \rho}{\partial t} = 0,$$

Comme nous travaillons avec de l'eau on la suppose incompressible ce qui donne

$$div \overrightarrow{u} = 0.$$

Dans ces hypothèses l'équation de Navier-Stokes devient

$$\frac{\partial p}{\partial x^i} - \mu \nabla^2 u^i - \rho F^i = 0. \tag{21}$$

3.2.3 La loi de Darcy

Henry Darcy alors qu'il étudiait les fontaines de la ville de Dijon (1856) établit expérimentalement que le débit d'eau s'écoulant à travers un massif de sable peut se calculer

$$Q = KA\frac{\Delta h}{L}. \tag{22}$$

A est la section du massif sableux
Δh la perte de charge de l'eau entre le sommet et la base du massif sableux
K est une constante dépendant du milieu poreux, baptisée coefficient de perméabilité
L est l'épaisseur du massif sableux.

On appelle $U = Q/A$ **la vitesse de filtration** d'un sol. À partir des équations de Navier-Stokes on sait que les causes du déplacement du fluide sont dûs au gradient de pression ainsi qu'aux forces extérieures. La loi de Darcy peut alors s'exprimer sous la forme générale

$$\overrightarrow{U} = \frac{k}{\mu}(\overrightarrow{\nabla}\, p + \rho g\, \overrightarrow{\nabla}\, z), \tag{23}$$

On peut réécrire cette équation

$$\overrightarrow{U} = -K(\overrightarrow{\nabla}\, h + \overrightarrow{\nabla}\, z),$$

avec $K = k(\mu\rho g)^{-1}$ $[LT^{-1}]$ et $h = p(\rho g)^{-1}$ $[L]$. Ceux sont avec ces équations que les géologues travaillent, nous pouvons remarquer que cette simplification des équations peut être considérée comme un processus de downscaling.

3.2.4 Des équations de Navier Stokes aux équations de Darcy

3.3 Les modèles

3.3.1 Modèle de surface continentale

3.3.2 Modèle de colonne d'eau

4 Prédictions climatiques

L'étude que nous avons menée a été réalisée sur le bassin du Little Washita. Nous avons eu accès à aux données NARR de 1979 à 2014.

4.1 Les données NARR et la méthodologie

Les données NARR (North American Regional Reanalysis) couvrent l'entièreté du continent nord américain. La méthode de projection pour passer de $\mathcal{S}(\mathbb{R}^3)$ à $\mathbb{R}^2$ et ce qu'on appelle la **projection Lambert**. Cette projection d'après le théorème de Gauss n'est pas isométrique, cependant l'approximation que la géométrie ainsi obtenue est encore euclidienne est une approximation convenable pour le bassin du Washita pour les données NARRs (voir 5.2). La longueur de grille du maillage est d'une trentaines de kilomètres (voir figure 1).

Afin de déterminer l'efficacité du downscaling dans le cas où l'on ajoute du bruit à la variable que l'on cherche à prédire nous avons dégradé spatialement nos données en moyennant sur des échelles de grille différentes. Nous les avons ensuite downscalées et puis les avons injectées dans un modèle hydrologique (voir à citer) puis nous avons comparé nos résultats. Nous analysons nos résultats avec les methodes fournies dans la partie 2.

GRID DESCRIPTIONS

Regional North American Grid (Lambert Conformal)
used by NAM, SREF and RAP.

Nx	349
Ny	277
La1	1.000N
Lo1	214.500E = 145.500W
Res. & Comp. Flag	0 0 0 0 1 0 0 0
Lov	253.000E = 107.000W
Dx	32.46341 km
Dy	32.46341 km
Projection Flag (bit 1)	0
Scanning Mode (bits 1 2 3)	0 1 0
Latin1	50.000N
Latin2	50.000N
Lat/Lon values of the corners of the grid	
(1,1)	1.000N, 145.500W
(1,277)	46.635N, 148.639E
(349,277)	46.352N, 2.566W
(349,1)	0.897N, 68.318W
Pole point	
(I,J)	(174.507, 307.764)

The Dx, Dy grid increment (at 50 deg north) was selected so that the grid spacing would be exactly 32.000 km at 40 deg north; the intersection of 40N & 107W falls on point (174.507,108.664)

Figure 3 – Description du maillage NARR

4.2 Analyse de la structure spatiale des données NARRs

Comme nous faisons une dégradation, il est intéressant de voir la structure spatiale des données. Nous avons une grille spatiale pour laquelle nous avons vu que la géométrie pouvait être considérée euclidienne. Les données NARRs se présentent sous la forme d'un tenseurs $\mathcal{T}$ de dimension $T \times M \times N \times v$ où T est le nombre de jours sur lesquels ont a ces observations, $(M+1) \times (N+1)$ la dimension de la grille d'observation et v le nombre de variables observées. Nous faisons l'hypothèse la fonction S_V possède un noyau de covariance spatial K et nous chercherons à déterminer sa forme.

Définition 9. *Soit $f : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R}$ une fonction aléatoire, on définit un noyau de covariance sur cette fonction en faisant l'hypothèse que f est stationnaire en temps et espace, on a alors*

$$Cov(f(t,x), f(t,x+y)) = K(y), \quad \forall t \in \mathbb{R}^+,\ x, y \in \mathbb{R}^2.$$

On remarque de plus que la fonction K ainsi définie est symétrique, $K(y) = K(-y)$, $\forall y \in \mathbb{R}^2$. On peut alors simplement calculer la covariance empirique à partir de l'estimateur sans biais de la variance.

$$K(m,n) = \frac{1}{T(M-m+1)(N-n+1)-1} \sum_{t=1}^{T} \sum_{i=m}^{M} \sum_{j=n}^{N} (\mathcal{T}_{t,i,j} - \overline{\mathcal{T}}_{1,m,n})(\mathcal{T}_{t,i-m,j-n} - \overline{\mathcal{T}}_{2,m,n}),$$

où,

$$\overline{\mathcal{T}}_{1,m,n} = \frac{1}{T(M-m+1)(N-n+1)} \sum_{t=1}^{T} \sum_{i=m}^{M} \sum_{j=n}^{N} (\mathcal{T}_{t,i,j}), \quad \overline{\mathcal{T}}_{2,m,n} = \frac{1}{T(M-m+1)(N-n+1)} \sum_{t=1}^{T} \sum_{i=0}^{M-m} \sum_{j=0}^{N-n} (\mathcal{T}_{t,i,j}).$$

On peut donc tracer le noyau de covariance $K(i,j)$, $i \in \{-N/2, N/2\}$, $j \in \{-M/2, M/2\}$.

4.3 Présentation des résultats de prédictions climatiques

Nous allons travailler avec plusieurs séries différentes, comme nous l'avons dit précédemment nous avons travailler avec plusieurs séries temporelles. Nous allons décrire ces séries temporelles.

On commence par réintroduire le tenseur $\mathcal{T}$ de dimension $T \times M \times N$. Nous avons l'équation

$$\mathcal{T}_{t,m,n} = \mathcal{T}_V(t, lat_m, lon_n).$$

On applique cette méthode aux données NARRs et l'on obtient le profile de covariance suivant.

5 Indexes

5.1 Indexe 1 : La statistique de Cramér-von Mises

Soient $(X_i)_{i\in[\![1,n]\!]}$ et $(Y_i)_{i\in[\![1,m]\!]}$ des réalisations indépendante issues de variables aléatoires réelles X et Y. On appelle $\mathcal{F}_n$ et $\mathcal{G}_m$ les fonctions de répartitions empiriques définie à partir de ces réalisation et $\mathcal{H}_{m,n}$ la fonction de répartition empirique définie à partir de l'ensemble de ces réalisations $(Z_i)_{i\in[\![1,m+n]\!]} = X_1, ..., X_n, Y_1, ..., Y_m$. Par la suite on considéra que tous les éléments sont triés dans leur ensemble ($i \leqslant j \Rightarrow E_i \leqslant E_j$). Nous avons alors l'égalité suivante

$$C_{n,m} = \frac{nm}{n+m} \int_{\mathbb{R}} \big[\mathcal{F}_n(x) - \mathcal{G}_m(x)\big]^2 \mathrm{d}\mathcal{H}_{m,n}(x) = \frac{1}{nm(m+n)} \Big[n \sum_{i=1}^{n} (R_{X_i} - i)^2 + m \sum_{i=1}^{m} (R_{Y_i} - i)^2 \Big] - \frac{4nm-1}{6(m+n)}. \quad (24)$$

où R_{X_i} est le rang de X_i dans $X_1, ..., X_n, Y_1, ..., Y_n$ autrement dit

$$R_{X_i} = Card(\{j \in [\![1, m+n]\!], Z_j \leqslant X_i\}).$$

Notons que cette égalité transforme un problème d'analyse en un problème de dénombrement beaucoup plus simple. On rappelle la définition de l'intégrale par rapport à une fonction.

Définition 10. *Soient f une fonction continue par morceaux de $\mathbb{R}$ dans $\mathbb{R}$, soit g une fonction continue par morceaux on définit l'intégrale en appelant $x_{i,n} = i/n$*

$$\int_{\mathbb{R}} f(x)\, dg(x) = lim_{n\to\infty} \sum_{i\in\mathbb{Z}} f(x_{i,n}) \big(g\big(x_{i,n}\big) - g\big(x_{i,n-1}\big) \big).$$

Démonstration. Commencons par montrer l'égalité

$$\frac{nm}{n+m} \int_{\mathbb{R}} \big[\mathcal{F}_n(x) - \mathcal{G}_m(x)\big]^2 \mathrm{d}\mathcal{H}_{m,n}(x) = \frac{mn}{(m+n)^2} \sum_{i=1}^{m+n} \big(\mathcal{F}_n(Z_i) - \mathcal{G}_m(Z_i)\big)^2.$$

On pose $\delta = \inf\{|Z_i - Z_j|, Z_i \neq Z_j\}$, quel que soit $n \geqslant n_0$ tel que $1/n_0 < \delta$ on a :

$$\sum_{i\in\mathbb{Z}} \big(\mathcal{F}_n(\frac{i}{n}) - \mathcal{G}_m(\frac{i}{n})\big)^2 \Big(\mathcal{H}_{m,n}\big(\frac{i}{n}\big) - \mathcal{H}_{m,n}\big(\frac{i-1}{n}\big) \Big) = \sum_{i=1}^{m+n} \big(\mathcal{F}_n(Z_i) - \mathcal{G}_m(Z_i)\big)^2,$$

on obtient donc directement l'égalité voulue en passant à la limite.

Observons maintenant que $\mathcal{F}_n(X_i) = i/n$ et $\mathcal{G}_m(X_i) = (R_{X_i} - i)/m$ ainsi que $\mathcal{F}_n(Y_i) = (R_{Y_i} - i)/n$ et $\mathcal{G}_m(X_i) = i/m$. On peut alors réécrire $C_{n,m}$ en séparant la somme sur les X_i et Y_i

$$C_{n,m} = \frac{mn}{(m+n)^2}\Big[\sum_{i=1}^{n}\Big(\frac{i}{n} - \frac{R_{X_i} - i}{m}\Big)^2 + \sum_{i=1}^{m}\Big(\frac{R_{Y_i} - i}{n} - \frac{i}{m}\Big)^2\Big]$$

$$= \frac{mn}{(m+n)^2}\Big[\frac{1}{m^2}\sum_{i=1}^{n}\Big(R_{X_i} - i\frac{m+n}{n}\Big)^2 + \frac{1}{n^2}\sum_{i=1}^{m}\Big(R_{Y_i} - i\frac{m+n}{m}\Big)^2\Big]$$

Remarquons que $C_{n,m}$ est de la forme

$$C_{n,m} = \frac{mn}{(m+n)^2}\Big[\frac{C_1}{m^2} + \frac{C_2}{n^2}\Big],$$

et que C_1 et C_2 sont symétriques en n et m. On définit $\Sigma_1 = \sum_{i=1}^{n} R_{X_i}^2$, $\Sigma_2 = \sum_{i=1}^{m} R_{Y_i}^2$ et $\mathcal{S}_k = \sum_{i=1}^{k} i^2$ nous allons travailler sur l'expression

$$C_1 = \sum_{i=1}^{n}\Big(R_{X_i} - i\frac{m+n}{n}\Big)^2.$$

On la développe puis factorise pour obtenir

$$C_1 = \frac{m+n}{n}\sum_{i=1}^{n}(R_{X_i} - i)^2 - \frac{m}{n}\Sigma_1 + \frac{m(m+n)}{n^2}\mathcal{S}_n.$$

On obtient de la même manière

$$C_2 = \frac{m+n}{m}\sum_{i=1}^{m}(R_{Y_i} - i)^2 - \frac{n}{m}\Sigma_2 + \frac{n(m+n)}{m^2}\mathcal{S}_m.$$

D'après ce qu'on a dit précédemment on a donc :

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{\Sigma_1 + \Sigma_2}{(m+n)^2} + \frac{\mathcal{S}_n}{n(n+m)} + \frac{\mathcal{S}_m}{m(m+n)}.$$

On remarque $\Sigma_1 + \Sigma_2 = \mathcal{S}_{m+n}$ et que l'on a la première moitié de notre somme. Il ne reste plus qu'à développer l'expression

$$-\frac{\mathcal{S}_{m+n}}{(m+n)^2} + \frac{\mathcal{S}_n}{n(n+m)} + \frac{\mathcal{S}_m}{m(m+n)}$$

$$= -\frac{(m+n+1)(2m+2n+1)}{6(m+n)} + \frac{(n+1)(2n+1)}{6(m+n)} + \frac{(m+1)(2m+1)}{6(m+n)} = -\frac{4mn-1}{6(m+n)}$$

En regroupant nos deux résultats nous avons finalement :

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{4nm-1}{6(m+n)}$$

□

5.2 Indexe 2 : Projection conique conforme de Lambert

La plupart des fonctions de projection de $S(\mathbb{R}^3)$ dans $\mathbb{R}^2$ sont des surfaces développables sur lesquelles ont projette les points de la terre. Par exemple des cônes, des cylindres et des plans (projection stéréographique) sont les surfaces développable les plus connues. La **projection Lambert** est une projection cônique aussi appelée la projection orthomorphique. Ses caractéristiques sont décrites dans le livre Grafarend and Krumm (2014). Elle possède la caractéristique de préserver les angles et les distances pour deux latitudes choisies, pour les données NARR les latitudes choisies sont 33°N et 45°N. De plus les lignes de latitudes égales sont des cercles et celles de longitudes égales des lignes droites. Les coordonées que nous étudions sont entre 33°N et 36°N. On va montrer que les longueures étudiées dans cette zone de l'espace ne souffrent que de très peu de déformation.

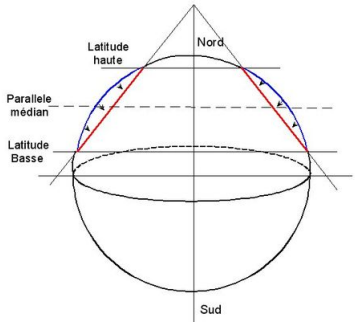


Figure 4 – Projection conique conforme de Lambert

Définition 11. *On définit les fonctions* $lat : S(\mathbb{R}^3) \to \mathbb{R}^2$ *et* $lon : S(\mathbb{R}^3) \to \mathbb{R}^2$ *qui associent à chaque point* x *de* $S(\mathbb{R}^3)$ *sa latitude et sa longitude.*

Définition 12. *On définit le cône convexe* $\zeta_{\theta,\theta+\epsilon}$ *comme l'ensemble des droites passant les points* x_1 *et* x_2 *de même longitudes de latitudes égale à* θ *et* $\theta + \epsilon$*. Autrement dit*

$$\zeta_{\theta,\theta+\epsilon} = \{D(x_1, x_2),\, x_1, x_2 \in S(\mathbb{R}^3),\, lon(x_1) = lon(x_2),\, lat(x_1) = \theta,\, lat(x2) = \theta + \epsilon\}.$$

où $D(x_1, x_2)$ *est la droite passant par* x_1 *et* x_2*.*

Il est évident que pour les lignes de latitude haute $(\theta + \epsilon)$ et basse (θ) les longueures sont conservées. On définit $\epsilon = \pi(45 - 33)/180$.

Proposition 3. *Pour toute courbe* $\gamma : [0, 1] \to S(\mathbb{R}^3)$ *continue dont les latitudes sont comprises entre* θ *et* $\theta + \epsilon$*, on a*

$$\min\Big(\frac{2\sin(\epsilon/2)}{\epsilon}, \frac{\cos(\theta+\epsilon)}{\cos(\theta)}\Big) \leqslant \frac{\|P(\gamma)\|_{\|\cdot\|_{\mathbb{R}^2}}}{\|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^3)}}} \leqslant 1.$$

Où P *est la projection de Lambert conservant les longueures pour les latitudes* θ *et* $\theta + \epsilon$*.*

Ce résultat permettra de conclure que la géométrie des lieux peux être considérée comme euclidienne si ϵ est suffisamment petit.

Démonstration. (Esquisse) On montrera cette inégalité pour les courbes de latitudes constantes et pour celles de longitudes constantes et la densité des fonctions de longitude ou de latitude par morceaux constantes dans l'ensemble des courbes permettra de conclure cette inégalité pour toutes les courbes.

On définit trois ensembles de courbes, $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_m$ tels que

$$\mathcal{C}_1 = \{\gamma \in \mathcal{C},\, lon(\gamma(t)) = c\, \forall t \in [0, 1],$$

$$\mathcal{C}_2 = \{\gamma \in \mathcal{C},\, lat(\gamma(t)) = c\, \forall t \in [0, 1]\},$$

$$\mathcal{C}_m = \{\gamma \in \mathcal{C},\, \exists t_1 < ... < t_n, \forall \leqslant i < n, \gamma_i : t \mapsto \gamma(t_i + t(t_{i+1} - t_i)) \in \mathcal{C}_1 \cup \mathcal{C}_2\}.$$

On commence par étudier les courbes γ dans $\mathcal{C}_1$ injectives, soit $C_\theta = x \in S(\mathbb{R}^3), lon()$ on peut alors sans perte de généralité se placer dans le cas du cercle unité dans $\mathbb{R}^2$ (le cercle de longitude constante). On a alors l'égalité

$$\|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^3)}} = \|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^2)}} = \|\gamma(1) - \gamma(0)\|_{S(\mathbb{R}^2)}.$$

Soit U le cercle unité alors quel que soit $\theta \in \mathbb{R}$ et $0 \leqslant \epsilon \leqslant \pi$, et en appelant $x_1 = (\cos(\theta), \sin(\theta))$ et $x_2 = (\cos(\theta + \epsilon), \sin(\theta + \epsilon))$ on l'égalité suivante

$$\frac{\|x_1 - x_2\|_{\mathbb{R}^2}}{\|x_1 - x_2\|_{S(\mathbb{R}^2)}} = \frac{2\sin(\epsilon/2)}{\epsilon}.$$

En remarquant que la fonction $f : [0, \pi] \to [0, 1]$, $x \mapsto 2\sin(x/2)/x$ est une fonction décroissante on peut en conclure que pour tout couple de points (x_1, x_2) dans l'arc de cercle $C_{\theta,\epsilon} = \{(\cos(\alpha + \epsilon), \sin(\theta + \alpha)), \alpha \in [0, \epsilon]\}$ on a

$$\frac{2\sin(\epsilon/2)}{\epsilon} \leqslant \frac{|x_1 - x_2\|_{\mathbb{R}^2}}{\|x_1 - x_2\|_{S(\mathbb{R}^2)}}$$

on obtient donc l'inégalité,

$$\frac{2\sin(\epsilon/2)}{\epsilon} \leqslant \frac{\|P(\gamma)\|_{\|\cdot\|_{\mathbb{R}^2}}}{\|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^3)}}}.$$

Étudions maintenant les courbes γ dans $\mathcal{C}_2$ de latitude égale à $\theta + \epsilon$ et injectives. On sait que la courbe γ ainsi que sa projection $P(\gamma)$ décrivent un arcs de cercle dans $\mathbb{R}^3$. Le rapport entre la longueur de l'arc de cercle définit par $P(\gamma)$ et γ dans $\mathbb{R}^3$ est majoré grossièrement par $\cos(\theta + \epsilon)/\cos(\beta)$.

$$\frac{\cos(\theta + \epsilon)}{\cos(\beta)} \leqslant \frac{\|P(\gamma)\|_{\|\cdot\|_{\mathbb{R}^2}}}{\|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^3)}}}$$

Pour chaque courbe γ dans $\mathcal{C}_m$ on a alors

$$\min\Big(\frac{2\sin(\epsilon/2)}{\epsilon}, \frac{\cos(\theta + \epsilon)}{\cos(\theta)}\Big) \leqslant \frac{\|P(\gamma)\|_{\|\cdot\|_{\mathbb{R}^2}}}{\|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^3)}}} \leqslant 1,$$

la densité de $\mathcal{C}_m$ dans l'ensemble des courbes continues permet de conclure. □

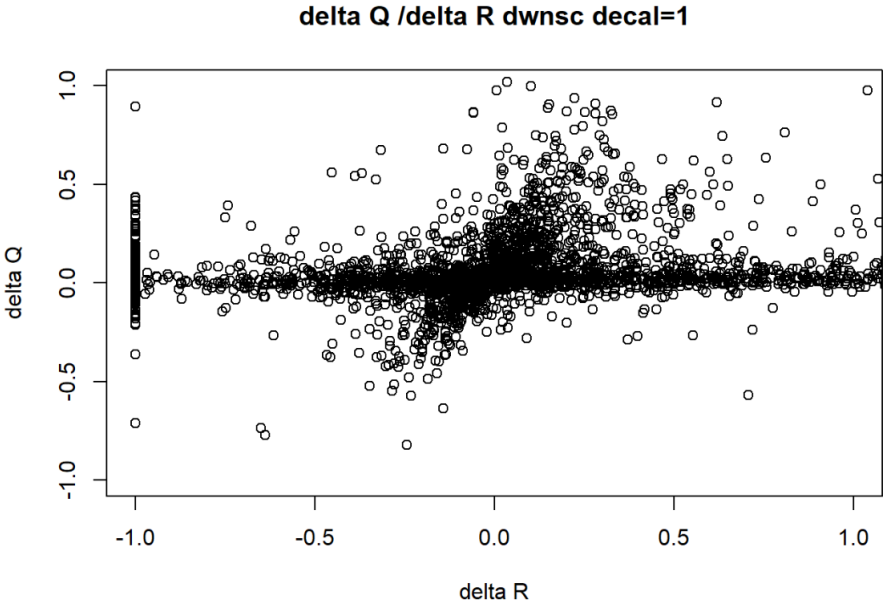
Finalement, on peut voir que dans notre cas où $\theta = \pi 33/180$ et $\epsilon = \pi 12/180$. On a que pour tout $\gamma : [0, 1] \to S(\mathbb{R}^3)$

$$0.84 \leqslant \frac{\|P(\gamma)\|_{\|\cdot\|_{\mathbb{R}^2}}}{\|\gamma\|_{\|\cdot\|_{S(\mathbb{R}^3)}}} \leqslant 1.$$

Regardons la grille donnée par le site sur lequel sont stockées les données NARRs.

5.3 Indexe 3 : Classification des populations de débit

L'étude la différence des bébits ΔD en fonction de la différence des précipitations ΔR nous donne à envisager deux classes débit.



D'une part il semble qu'il y ait une population de débits indépendants des précipitations et une classe de débits linéairement liés aux précipitations. Il parait alors pertinent de classifier ces débits à partir de deux droites. Soit X un ensemble de points $x_1, x_2, .., x_n \in \mathbb{R}^2$, on cherche deux droites D_1 et D_2 minimisant la valeur

$$\sum_{i=1}^{n} d(x_i, D_1 \cup D_2),$$

où d est la distance définie par

$$d(x, E) = \min_{e \in E} |x - e|^2.$$

On peut définir les droites de $\mathbb{R}^2$ par un point de $\mathcal{U}_1(\mathbb{R}^2) \times \mathbb{R}^2$, où $\mathcal{U}_1(\mathbb{R}^2)$ est le cercle unité. Alors on peut redéfinir un problème avec contraintes

$$\begin{array}{rccc} F: & (\mathcal{U}_1(\mathbb{R}^2) \times \mathbb{R}^2) \times (\mathcal{U}_1(\mathbb{R}^2) \times \mathbb{R}^2) & \to & \mathbb{R} \\ & ((u_1, v_1), (u_2, v_2)) & \mapsto & \sum_{i=1}^{n} \min(d(x_i, \mathbb{R}u_1 + v_1), d(x_i, \mathbb{R}u_2 + v_2)) \end{array} \tag{25}$$

On commence par développer la distance à une droite, c'est une formule de projection classique. On appelle f la fonction $f : \mathcal{U}_1(\mathbb{R}^2) \times \mathbb{R}^2 \times \mathbb{R}^2$, $f((u, v), x) = d(x, \mathbb{R}u + v)$.

$$\begin{array}{rcl} f_x(u, v) & = & d(x, \mathbb{R}u + v) \\ & = & \left|x - v - u\langle x - v, u\rangle\right|^2 \\ & = & \left|x - v\right|^2 - \left\langle x - v, u\right\rangle^2 \\ & = & \left|x\right|^2 - 2\langle x, v\rangle + \left|v\right|^2 - \left\langle x, u\right\rangle^2 + 2\langle x, u\rangle\langle v, u\rangle - \left\langle v, u\right\rangle^2 \end{array} \tag{26}$$

On cherche maintenant à calculer $\overrightarrow{\nabla} f_x$, le gradient étant une application linéaire on a

$$\overrightarrow{\nabla} f_x(u, v) = \overrightarrow{\nabla}\left|x\right|^2 - \overrightarrow{\nabla} 2\langle x, v\rangle + \overrightarrow{\nabla}\left|v\right|^2 - \overrightarrow{\nabla}\left\langle x, u\right\rangle^2 + \overrightarrow{\nabla} 2\langle x, u\rangle\langle v, u\rangle - \overrightarrow{\nabla}\left\langle v, u\right\rangle^2,$$

et finalement on obtient

$$\overrightarrow{\nabla} f_x(u, v) = 2 \begin{pmatrix} (v - x)\langle u, x - v\rangle \\ v - x + u\langle u, x - v\rangle \end{pmatrix}. \tag{27}$$

On peut donc en déduire une formule pour la fonction F définie dans (25)

$$\begin{array}{rcl} F((u_1, v_1), (u_2, v_2)) & = & \sum_{i=1}^{n} \min(f_{x_i}(u_1, v_1), f_{x_i}(u_2, v_2)) \\ & = & \sum_{i=1}^{n} \frac{f_{x_i}(u_1, v_1) + f_{x_i}(u_2, v_2) - \left|f_{x_i}(u_1, v_1) - f_{x_i}(u_2, v_2)\right|}{2} \end{array} \tag{28}$$

Alors on peut obtenir une expression pour $\overrightarrow{\nabla} F$,

$$\overrightarrow{\nabla} F(u_1, v_1, u_2, v_2) = \begin{pmatrix} \sum_{i=1}^{n} \mathcal{I}^-{}_{x_i}((u_1, v_1), (u_2, v_2)) \overrightarrow{\nabla} f_{x_i}(u_1, v_1) \\ \sum_{i=1}^{n} \mathcal{I}^+{}_{x_i}((u_1, v_1), (u_2, v_2)) \overrightarrow{\nabla} f_{x_i}(u_2, v_2) \end{pmatrix} \tag{29}$$

où

$$\mathcal{I}^-{}_{x_i}((u_1, v_1), (u_2, v_2)) = \mathbb{1}_{]-\infty, 0]}(f_{x_i}(u_1, v_1) - f_{x_i}(u_2, v_2))$$

et

$$\mathcal{I}^+{}_{x_i}((u_1, v_1), (u_2, v_2)) = \mathbb{1}_{[0, \infty[}(f_{x_i}(u_1, v_1) - f_{x_i}(u_2, v_2)).$$

On peut appliquer l'algorithme d'Usawa pour trouver minimum sur $(\mathcal{U}_1(\mathbb{R}^2) \times \mathbb{R}^2) \times (\mathcal{U}_1(\mathbb{R}^2) \times \mathbb{R}^2)$ en considérant le plongement de F dans $(\mathbb{R}^2 \times \mathbb{R}^2) \times (\mathbb{R}^2 \times \mathbb{R}^2)$ avec les contraintes $|u_i| = 1$.

Références

Ayar, P. V., Vrac, M., Bastin, S., Carreau, J., Déqué, M., and Gallardo, C. (2016). Intercomparison of statistical and dynamical downscaling models under the euro-and med-cordex initiative framework : present climate evaluations. *Climate dynamics*, 46(3-4) :1301–1329.

Büning, H. (2002). Robustness and power of modified lepage, kolmogorov-smirnov and cramér-von mises two-sample tests. *Journal of Applied Statistics*, 29(6) :907–924.

Cappus, P. (1960). Etude des lois de l'écoulement-application au calcul et à la prévision des débits. *La houille blanche*, pages 493–520.

Christensen, J. H., Boberg, F., Christensen, O. B., and Lucas-Picher, P. (2008). On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, 35(20).

De Marsily, G. (1986). Quantitative hydrogeology. Technical report, Paris School of Mines, Fontainebleau.

Durrett, R. (2019). *Probability : theory and examples*, volume 49. Cambridge university press.

Éthier, F. (2011). *À propos de divers tests statistiques pour l'égalité des lois*. PhD thesis, Université du Québec à Trois-Rivières.

Fisz, M. (1963). Probability theory and mathematical statistics.

Grafarend, E. W. and Krumm, F. W. (2014). *Map projections*. Springer.

Hillel, D. (2003). *Introduction to environmental soil physics*. Elsevier.

Horton, R. E. (1933). The role of infiltration in the hydrologic cycle. *Eos, Transactions American Geophysical Union*, 14(1) :446–460.

Huang, W.-L. and Chen, S.-P. (2012). Optimal aggregate production planning with fuzzy data. *International Journal of Industrial and Manufacturing Engineering*, 6(8) :1633–1638.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(4) :423–498.

Maquin, M. (2016). *Développement d'un modèle hydrologique de colonne représentant l'interaction nappe-végétation-atmosphère et applications à l'échelle du bassin versant*. PhD thesis, Université Paris-Saclay (ComUE).

Maraun, D. (2012). Nonstationarities of regional climate model biases in european seasonal mean temperature and precipitation sums. *Geophysical Research Letters*, 39(6).

Nahar, J., Johnson, F., and Sharma, A. (2017). Assessing the extent of non-stationary biases in gcms. *Journal of Hydrology*, 549 :148–162.

Robin, Y., Vrac, M., Naveau, P., and Yiou, P. (2019). Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, 23(2) :773–786.

Villani, C. (2003). *Topics in optimal transportation*. Number 58. American Mathematical Soc.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., and Somot, S. (2012). Dynamical and statistical downscaling of the french mediterranean climate : uncertainty assessment. *Natural Hazards and Earth System Sciences*, 12(9) :2769–2784.