Laboratoire des Sciences du Climat
et de l'Environnement

Changement d'échelles dans les projections climatiques et leurs impacts hydrologiques: Cas des grandes plaines américaines

Auteur

Mathis Deronzier
Mines Saint-Étienne

Maîtres de stage

Emmanuel Mouche
C.E.A.
Mathieu Vrac
C.N.R.S.

Stage de recherche de master 2

Avril-Septembre
2021

Contents

1 Introduction

1.1 Contextualisation et motivations du stage

Alors que le rapport du GIEC 2021 sorti cet été projette une augmentation de la température mondiale moyenne de 1.5°C d'ici 2050 ainsi que des modifications des climats dans plusieurs régions du monde. Il semble aujourd'hui primordial de comprendre les conséquences locales d'un changement climatique global. Les changements climatiques locaux sont des enjeux politiques et économiques majeurs et leurs prévisions sont des problématiques centrales pour la population mondiale.

Les projections climatiques sont données par des modèles de climat travaillant à l'échelle de l'ordre d'une centaine de kilomètres, on cherche alors à prévoir des résultat sur des zones plus localisées. Le domaine du climat étudiant les questions de changement d'échelle du global au local est le *downscaling*. La question inverse du changement d'échelle du local au global est non moins intéressante pour les climatologues. En effet, comment savoir si les modèles à grande échelle sont vraiment représentatifs de la réalité? Alors que les modèles climatiques et hydrologiques travaillent sur des échelles globales, les lois de la physique sont elles, locales. On cherche alors à comprendre à partir des équations de la physique les interactions ou équations qu'elles engendrent à grande échelle, ce domaine de recherche s'appelle l'*upscaling*.

C'est dans ces problématiques de changement d'échelles que s'ancre ce stage. Pour concentrer et enrichir nos réflexions, nous nous intéresserons plus précisément à la question de l'impact du changement d'échelle sur les modélisations hydrologiques. Le domaine d'étude ayant été choisi pour modéliser et concrétiser ces problématiques a été le bassin du Little Washita. Ce bassin situé aux États-Unis dans l'état d'Oklahoma, possède de nombreuses caractéristiques qui le rendent intéressant. Sa superficie de $611 km^2$ est de l'ordre de grandeur d'une maille de modèle continentale. Il a déjà fait l'objet de nombreuses études (voir par exemple Maxwell et al. (2007), Rosero et al. (2011), Maquin (2016)).

Pour s'intéresser à la question du changement d'échelle il fallait pouvoir comparer des données à grande échelle et à petite échelle. Les données NARR (North American Regional Reanalysis) allant de l'année 1979 à 2014 semblaient constituer un jeux de donnée pertinent pour l'étude du Little Washita. En effet, sa longueur de maille de $32 km$ permettait à la fois de considérer l'upscaling hydrologique sur de ce bassin ainsi que le downscaling pour des jeux de données sur des échelles plus larges. Pour obtenir ces données à grande échelle nous avons opté pour deux types de données: les données l'IPSL qui possède une longueur de maille de $200 km$ ainsi que des données obtenues en faisant des dégradations spatiales sur les données NARR. Ainsi, si les données IPSL et NARR sont peut-être décorrélées on est absolument certain que les données NARR et leurs dégradées le sont. L'objectif aura été dans un premier temps de regarder différentes méthodes de downscaling pour améliorer les projections des données NARR à partir des séries à plus grande échelle. Une fois effectuée cette étape il s'agissait de réaliser une simulation hydraulique sur les années prédites et de comparer l'impact du downscaling sur les résultats des simulations hydrologiques. La problématique de l'upscaling sera en fait cachée dans la modélisation hydrologique du bassin du Little Washita, bien qu'elle n'ait pas été concrètement traitée dans les faits, elle a été centrale dans notre réflexion. Finalement nous avons donné les conclusions que nous avons tirées dans.

1.2 Présentation du plan

Dans la première partie nous allons introduire rigoureusement le concept de downscaling et les outils utilisés pour mesurer la qualité de nos prédictions. Nous verrons l'algorithme *Cumulative Distribution Function transfert* (CDF-t), celui utilisé pour downscaler nos séries. Nous introduirons le transport optimal généralisant l'algorithme CDF-t. Nous verrons ensuite comment estimer la qualité de nos prédictions, alors on étudiera les distances de Kolmogorov Smirnov et celle de Cramér-von Mises permettant de quantifier la différence entre deux lois.

La seconde partie sera consacrée à la physique et la modélisation, elle présentera les principaux mécanismes de la modélisation hydrologique ainsi que les modèles d'Orchidée et d'HydroGéoSphère. Dans un premier temps nous présenterons les interactions sol-atmosphère ainsi que les caractéristique des sol et leur influence sur les débits. Nous présenterons ensuite les principales équations de la mécanique des fluides en milieu poreux, puis deux modèles utilisés pour les modélisations hydrologiques, Orchidée et HydroGéoSphère en expliquant plus en détail la problématique de l'upscaling.

La troisième et dernière partie présentera la démarche les résultats ainsi que les questionnements que nous avons eu au cours de ce stage. Nous commencerons par expliquer en détail la méthodologie que avons suivi, puis nous présenterons pas à pas les résultats que nous avons obtenus.

2 Changement d'échelle, downscaling et analyse des résultats

Cette section se concentre sur la partie statistique du stage, on cherchera à partir de données à grande échelle issues de modèles climatiques et de données réelles à améliorer les projections de la précipitation et de l'évapotranspiration sur le bassin versant du Little Washita.

Nous commencerons par formaliser rigoureusement le downscaling statistique, puis nous réintroduirons les concepts mathématiques essentiels à la compréhension de l'algorithme CDF-t. Nous verrons ensuite comment on peut généraliser le principe de l'algorithme CDF-t avec la méthode du transport optimal. Pour terminer nous verrons les méthodes de test pour évaluer la qualité de prédictions sur plusieurs année

Le downscaling statistique (voir par exemple Vrac et al. (2012) et Ayar et al. (2016)) est une méthode utilisée dans les sciences du climat permettant d'améliorer les projections des modèles climatiques. À partir des données obtenues par un modèle climatique (modèle de circulation général, modèle climatique régional) et des données observées, on cherche à observer et corriger les biais systématiques introduits par les modèles. Le nom "downscaling" vient du domaine d'application de cette méthode. On passe d'un modèle à grande échelle à des données observées à petite échelle. Cette méthode est très utile dans la pratique où l'on a des modèles de climat donnant des résultats sur des maillages de grande échelle ($\sim 200km \times 200km$). Dans notre cas, nous utilisons le downscaling pour projeter les variables climatiques de *précipitation* et d'*évapotranspiration* sur le bassin du Little Washita. Nous n'expliquerons que deux méthodes de downscaling mais d'autres méthodes existent, l'article de Ayar et al. (2016) donne un aperçu des différentes méthodes. Nous n'avons testé qu'une méthode de downscaling mais d'autre existent. Nous l'avons testée à partir du jeux de données: les données NARRs (de taille de grille de $30km \times 30km$) ainsi que celles de l'IPSL (de taille de grille de $200km \times 200km$), nous décrivons ce travail dans la section **??**.

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 réelles dépendantes du temps et de l'espace, on appelle $\mathcal{M}(\Omega, \mathbb{R})$ l'espace des variables aléatoires réelles et $S(\mathbb{R}^3)$ 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. Nous travaillerons par la suite sur la sphère unité que l'on considérera être la terre.

Définition 1. *Pour une variable quantitative V à valeur dans $\mathbb{R}$, on appelle $\mathcal{T}_V$ la fonction donnant les valeurs réelles 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 trouver des transformations sur les variables aléatoires $S_V(t,x)$ pour minimiser ces distances. Plusieurs méthodes peuvent-être utilisées pour réduire ces distances. Les géostatistiques essaient entre autres d'étudier la structure de covariance sur l'espace sur lequel on considère ces variables. Des méthodes de résolution d'équations différentielles permettent d'interpoler les résultats (voir par exemple Lindgren et al. (2011)). Nous ne donnons qu'un bref apercu de ces méthodes puisque ce ne sont pas celles-ci que nous allons étudier.

2.1 Introduction à la problématique du downscaling

Le downscaling consiste donc à effectuer des transformations sur des variables aléatoires réelles. Pour étudier ces variables aléatoires et construire les transformation on commence par introduire ici les notions de **fonction de répartition**, **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, $\mathcal{F}_X : \mathbb{R} \to [0,1]$ 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 $\mathcal{F}_n : \mathbb{R} \to [0,1]$ 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, $f_X : \mathbb{R} \to [0,\infty[$ 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 Golse (2020).

Commençons par établir notre problématique dans le cas le plus simple où l'on cherche à projeter 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 $\mathbb{R}$,

$$(X_t)_{t\in\mathbb{N}} = (\mathcal{S}_V(t,x))_{t\in\mathbb{N}} \text{ et } (Y_t)_{t\in\mathbb{N}} = (\mathcal{T}_V(t,x))_{t\in\mathbb{N}}.$$

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 $X_{i_1}, X_{i_2}, ..., X_{i_m}$ (pour des temps futurs), 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 dans la section 2.4).

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).

2.2 Cumulative Distribution Function transform (CDFt)

Nous allons ici présenter l'algorithme principalement étudié et utilisé dans ce stage, l'algorithme CDF-t (voir Vrac et al. (2012)). Nous commencerons par présenter l'algorithme quantile-quantile (voir) permettant de comprendre l'esprit des transformations G affectées aux processus aléatoires. Puis nous décrirons l'algorithme de CDFt-t.

2.2.1 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.*

On peut retrouver la démonstration de cette proposition dans les annexes section 4.1. Le principe de l'algorithme **quantile-quantile** est alors 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. Ce qui n'est pas le cas dans la transformation G défini par la proposition 1. C'est en partie cette observation qui motive l'étude d'autres transformations.

2.2.2 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. On considère les réalisations $X_1,..,X_n$ et $Y_1,...,Y_n$ ainsi que $X'_1,...,X'm$.

CDFt avec support égale:

On peut par exemple faire en sorte que la transformation sur les projections conserve le support de celles-ci en d'autres mots on veut

$$\min_{i\in I}(X'_i) = \min_{i\in I}(G_{X,Y}(X'_i)) \quad \text{et} \quad \max_{i\in I}(X'_i) = \max_{i\in I}(G_{X,Y}(X'_i)).$$

Il suffit de transformer la loi de Y sur $[0,1]$, de réaliser un quantile-quantile puis d'effectuer la transformation inverse. Concrêtement, nous posons

$$\widetilde{Y} = \frac{(Y - \min_{i\in I} Y_i)}{\max_{i\in I} Y_i - \min_{i\in I} Y_i}.$$

Alors quelque soit i dans $\mathcal{I}$ on a $\widetilde{Y}_i \in [0,1]$. En appelant $G = \mathcal{F}_{X,\widetilde{Y}}$ la transformation quantile-quantile de X à $\widetilde{Y}$ on a $G_{X,Y}(X'_i) \in [0,1]$, la transformation s'imposant naturellement est alors

$$G_{X,Y} = (\max_{i\in I} Y_i - \min_{i\in I} Y_i)G + \min_{i\in I} Y_i.$$

C'est l'une des premières idées qui vienne-t-en tête lorsqu'on parle de transformations que l'on peut appliquer sur nos variables. Notons que l'on a aussi une grosse base de données et que l'on peut aussi essayer d'utiliser des méthodes de regression linéaire ou de machine learning pour estimer les max et les min de Y' en fonction de X. Nous allons illustrer l'usage d'une de ces méthodes dans la prochaine partie 2.2.2.

CDFt avec méthode de prédiction

Supposons que l'on ait des fonctions estimant la variance et la moyenne de Y' à partir de X', s'exprimant sous la forme $\overline{f(X')} = \overline{Y'}$ ainsi que $\overline{g(X')} = \overline{\sigma(Y')}$ [1]. Ces fonctions peuvent être obenues par des méthodes de régression linéaire sur des sous échantillons des données de X et Y soit après les avoir ordonnées si les X_i et les Y_i sont indépendants soit en fonction de leur indice.

On aimerait alors que $G_{X,Y}$ conserve la moyenne ainsi que 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.

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

Note 1. *Remarquons que l'efficacité des transformations que nous avons décrites repose en partie sur la stationnarité des lois suivies par les variables aléatoires aléatoires au cours du temps. Les questions de stationnarité ont été abordées dans Maraun (2012), Christensen et al. (2008) ainsi que Nahar et al. (2017).*

Finalement, une des limites de l'algorithme de CDF-t est dans le cas où plusieurs valeurs de X' sont égales elle seront envoyées sur le même points. C'est vraiment problématique dans le cas où X' possède plusieurs réalisations égales et que $\mathcal{F}_Y$ est une fonction continue (sans Dirac).

Exemple 1. *Soit X une variable aléatoire qui possède comme probabilité*

$$P(X = 0) = \frac{1}{2} \text{ et } P(\frac{1}{2} \leqslant a \leqslant X \leqslant b \leqslant 1) = b - a.$$

et Y une variable aléatoire réelle suivant la loi uniforme sur $[0, 1]$. Intuitivement, on aimerait que pour chaque $i \in \{1, ..., n\}$ tel que $X_i = 0$ $G(X_i)$ suivent une loi uniforme sur $[0, 1/2]$ ce qui n'est pas possible avec les transformations que nous avons considérées.

C'est en partie à ce problème que répond le transport optimal.

2.3 Transport optimal

Nous voulons à la fois projeter 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 projection. 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 projeter 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.

2.3.1 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 cherché 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 optimal est de trouver les déplacements naturels des poids d'une fonction de densité à une autre.

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 de référence internationale sur les problématiques de transport optimal.

Considérons deux fonctions de répartitions pour des variables aléatoires U et V à valeurs dans A et B, 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 $A \times B$ satisfaisant

$$\int_B d\pi(x,y) = f(x), \quad \int_A d\pi(x,y) = g(y),$$

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

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

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. On peut interpréter $d\pi(x,y)$ comme la quantité déplacée de x à y.

2.3.2 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 Robin et al. (2019) 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 méthode de krigeage par exemple) 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). Nous n'avons pas utilisé cette méthode, bien qu'il aurait été très intéressant de voir ses résultats.

2.4 Analyse des résultats obtenus par downscaling

L'analyse de nos résultats se fera par validation croisée, nous allons apprendre sur 50% de nos données et faire nos projections sur les 50% restants. La question à laquelle nous allons répondre dans cette partie est la suivante:

Comment évaluer la qualité de nos prédictions alors que nous avons réalisé des prédictions sur plusieurs années?

Contrairement à la manière habituelle de faire, consistant à estimer une distance entre chaque point projeté (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 projections.

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.

2.4.1 Distance de Kolmogorov-Smirnov

Le test d'ajustement de Kolmogorov-Smirnov Büning (2002) 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:

$$Q(x) = P(K > x) = \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2(jx)^2) \tag{9}$$

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** Büning (2002) (section **??**).

2.4.2 Distance de Cramér-von Mises

On considère ici les deux fonctions de répartitions $\mathcal{F}_n$ et $\mathcal{G}_m$ continues des variables aléatoires X et Y. Nous voulons tester les hypothèses

$$\mathcal{H}_0 : \mathcal{F}_n = \mathcal{G}_m \quad \text{et} \quad \mathcal{H}_1 : \mathcal{F}_n \neq \mathcal{G}_m.$$

Dans les tests proposés les statistiques sont construites à partir de deux échantillons indépendants des tests dans les cas où les variables X et Y sont indépendantes peuvent être trouvés dans Éthier (2011). 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

$$C_{n,m} = \frac{nm}{n+m} \int_{\mathbb{R}} \big[\mathcal{F}_n(x) - \mathcal{G}_m(x)\big]^2 \mathrm{d}\mathcal{H}_{m,n}(x), \tag{10}$$

avec,

$$\mathcal{H}_{n,m} = \frac{n}{n+m}\mathcal{F}_n + \frac{m}{n+m}\mathcal{G}_m. \tag{11}$$

$\mathcal{H}_{n,m}$ 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 $C_{n,m}$

$$C_{n,m} = \frac{mn}{(m+n)^2} \sum_{i=1}^{m+n} \big(\mathcal{F}_n(Z_i) - \mathcal{G}_m(Z_i)\big)^2 \tag{12}$$

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:*

$$C_{n,m} = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{4nm-1}{6(m+n)}. \tag{13}$$

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 alors 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 Présentation des modèles de climat

Cette section se concentre sur la partie physique du stage, nous développerons les concepts clés à la compréhension des modèles hydrologiques donnerons un apercu du fonctionnement des modèles climatiques en générale.

Nous commencerons par contextualiser l'étude hydrologique dans les modèles de climat en présentant brièvement les interactions entre ces différents modèles. Nous présenterons ensuite les principaux phénomènes physiques intervenant dans l'étude hydrologique d'un bassin. Nous irons ensuite plus dans les détails des équations de Navier-Stokes et présenterons la construction des principales équations de la mécanique des fluides en milieu poreux. Et pour finir nous présenterons deux modèles continentaux de simulation hydrologique très différents dans leur approche et essaierons de faire comprendre au lecteur les enjeux de l'upscaling dans la comparaison de ces deux modèles.

3.1 Présentation d'un modèle de climat

"*Un modèle climatique simule les interactions entre l'atmosphère, l'océan et les surfaces continentales. Grâce au modèle, les scientifiques ont des représentations numériques de la répartition géographique de différents paramètres, tels que la répartition des vents, des nuages, des masses d'eau...*" Jean Louis Dufresne. Les modèles interagissent entre eux selon un grand nombre d'équations mathématiquesw, et le maillage de simulation est alors un maillage multidimensionnel (voir la figure 1).

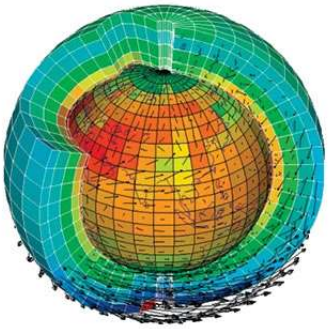

Figure 1: illustration d'un maillage multidimentionnel pour les modèles de climat

La complexité des interactions entre les différents noeuds peuvent rendre les calculs vraiment laborieux pour des mailles petites. D'où l'intérêt de considérer des mailles à grande échelle puis de downscaler les résultats obtenus. Nous présenterons dans la figure 2 les différentes interactions prises en compte dans l'élaboration du modèle de climat de l'IPSL.

Dans notre stage, nous nous sommes concentrés sur les modélisations continentales, à savoir, sur les interactions entre les modèles d'atmosphères ainsi que les mécanismes d'écoulement de l'eau dans les sols et les rivières. c'est ce que nous verrons dans la prochaine partie.

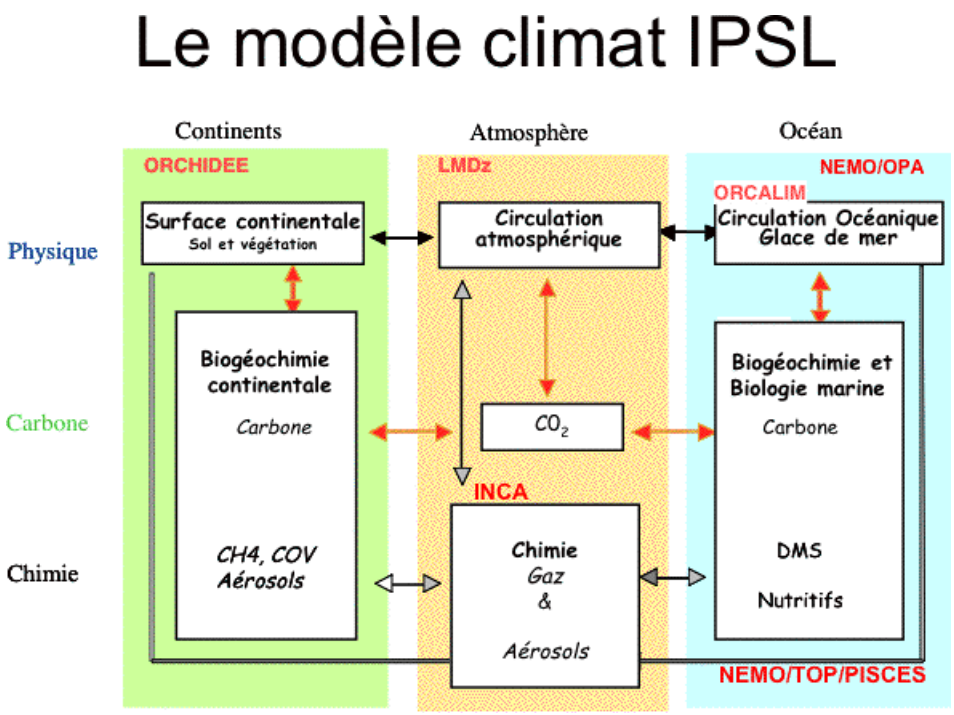

Figure 2: Présentation sommaire des différentes interactions dans le modèle de climat IPSL

3.2 Généralités sur les interactions atmosphère - surface continentale - sol

Nous allons présenter ici les principales interactions entre la surface continentale et l'atmosphère, cette étude a seulement pour objectif de donner un apercu des différentes problématiques rencontrées dans l'élaboration des modèles climatiques. En effet, nous n'avons pas eu à nous soucier de toutes ces interactions puisque les résultats renvoyés par les données NARR était l'évapotranspiration, une variable regroupant l'information de l'ensemble des interactions décrites ci-dessous. Cette section, de plus s'inspire très largement de la thèse Maquin (2016)[chap 1] qui fournit d'avantage de détails et traite de la modélisation de l'ensemble des interactions dont nous allons parler.

3.2.1 Les écoulements

Les processus hydrologiques sont classiquement étudiés à l'échelle du bassin versant[2]. 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. la figure 3 définie un bassin versant.

[2]Un bassin versant comme il est défini en hydrologie possède son équivalent en mathématiques, soit (E, d) 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 en considérant le plus grand bassin d'attraction contenant un point x choisi.

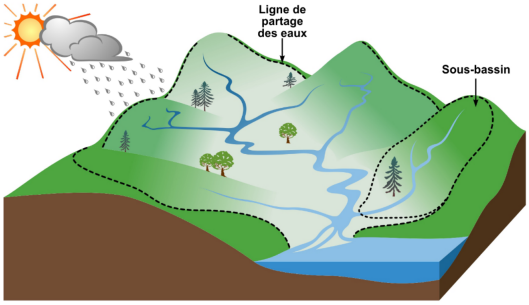

Figure 3: Bassin versant (Source :http://rqes-gries.ca/).

À l'échelle du bassin versant, on distingue deux types d'écoulements: les écoulements de subsurface, les écoulements de surface. Les premiers sont des écoulements ayant lieux dans les pores du sols dans la région non saturée et les seconds en 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 dépend de plusieurs paramètres comme les caractéristiques du sol (sa porosité, sa perméabilité) c'est à dire la capacité qu'a l'eau à s'infiltrer dans le sol ainsi que 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 **??**), pour plus de détails se référer à Maquin (2016) et De Marsily (1986).

Les écoulements de surface:

Les écoulements de surface, aussi qualifiés de ruissellement apparaissent lorsque le sol est saturé en surface et que le débit d'eau sur le sol est supérieur à sa capacité d'infiltration. Deux phénomènes distincts sont responsables du ruissellement, lorsque le sol est saturé, l'eau ne peut plus 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 aussi 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. La quantité d'infiltration décroit exponentiellement lors d'évènements pluvieux Horton (1933).

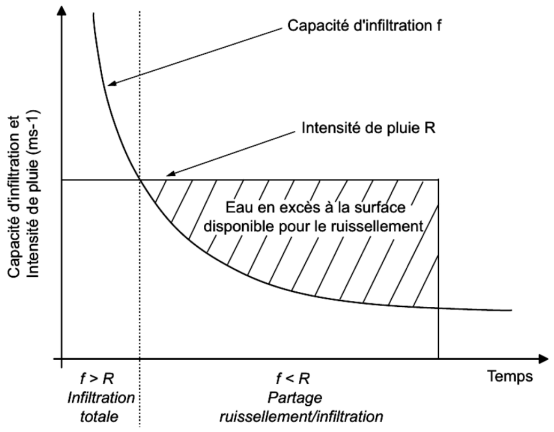

Figure 4: Estimation du ruissellement en fonction du temps, modèle de Horton

3.2.2 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. Attention l'évapotranspiration potentielle ne dépend pas de la couverture végétale (on la pondère généralement par la densité de la végétation présente sur le sol).

4 Annexes

4.1 Annexe: Preuves et légitimité des outils utilisés dans le downscaling

Proposition 2. *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.*

Proof. 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

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

□

Proposition 3. *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.

Proof. 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 4.1. *(Glivenko-Cantelli) Soient $\mathcal{F}_X$ et $\mathcal{F}_n$ respectivement la fonction de répartition et la fonction de répartition empirique. Alors*

$$\|\mathcal{F}_X - \mathcal{F}_n\|_\infty \xrightarrow[n\to\infty]{prob} 0. \tag{14}$$

Proof. (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 (3). 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 4.1 que la fonction de répartition empirique est le bon estimateur de la fonction de répartition.

4.2 Annexe: La statistique de Cramér-von Mises

Soient $(X_i)_{i\in[\![1,n]\!]}$ et $(Y_i)_{i\in[\![1,m]\!]}$ des réalisations indépendante issues des variables aléatoires réelles X et Y. On appelle $\mathcal{F}_n$ et $\mathcal{G}_m$ les fonctions de répartitions empiriques définies à partir de ces réalisation et $\mathcal{H}_{m,n}$ la fonction de répartition empirique définie à partir des 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 (c.à.d. $i \leqslant j \Rightarrow E_i \leqslant E_j$). Nous avons alors l'égalité suivante

$$C_{n,m} = \frac{nm}{n+m}\int_{\mathbb{R}} \big[\mathcal{F}_n(x) - \mathcal{G}_m(x)\big]^2 \mathrm{d}\mathcal{H}_{m,n}(x) = \frac{1}{nm(m+n)}\Big[n\sum_{i=1}^{n}(R_{X_i} - i)^2 + m\sum_{i=1}^{m}(R_{Y_i} - i)^2\Big] - \frac{4nm-1}{6(m+n)}. \tag{15}$$

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

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

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 7. *Soient $f : \mathbb{R} \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$ deux fonctions continues par morceaux on définit l'intégrale de f par rapport à g comme étant*

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

Proof. 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:

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

□

References

Ayar, P. V., Vrac, M., Bastin, S., Carreau, J., Déqué, M., and Gallardo, C. (2016). Inter-comparison 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.

Golse, F. (2020). *Distributions, analyse de Fourier, équations aux dérivées partielles*. Les éditions de l'école polytechnique.

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).

Maxwell, R. M., Chow, F. K., and Kollet, S. J. (2007). The groundwater–land-surface–atmosphere connection: Soil moisture effects on the atmospheric boundary layer in fully-coupled simulations. *Advances in Water Resources*, 30(12):2447–2466.

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.

Rosero, E., Gulden, L. E., Yang, Z.-L., De Goncalves, L. G., Niu, G.-Y., and Kaheil, Y. H. (2011). Ensemble evaluation of hydrologically enhanced noah-lsm: Partitioning of the water balance in high-resolution simulations over the little washita river experimental watershed. *Journal of Hydrometeorology*, 12(1):45–64.

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.