Séance 5

Introduction à la descente d'échelle statistique et à la correction de bias

Soulivanh Thao
sthao@lsce.ipsl.fr

LSCE, ESTIMR

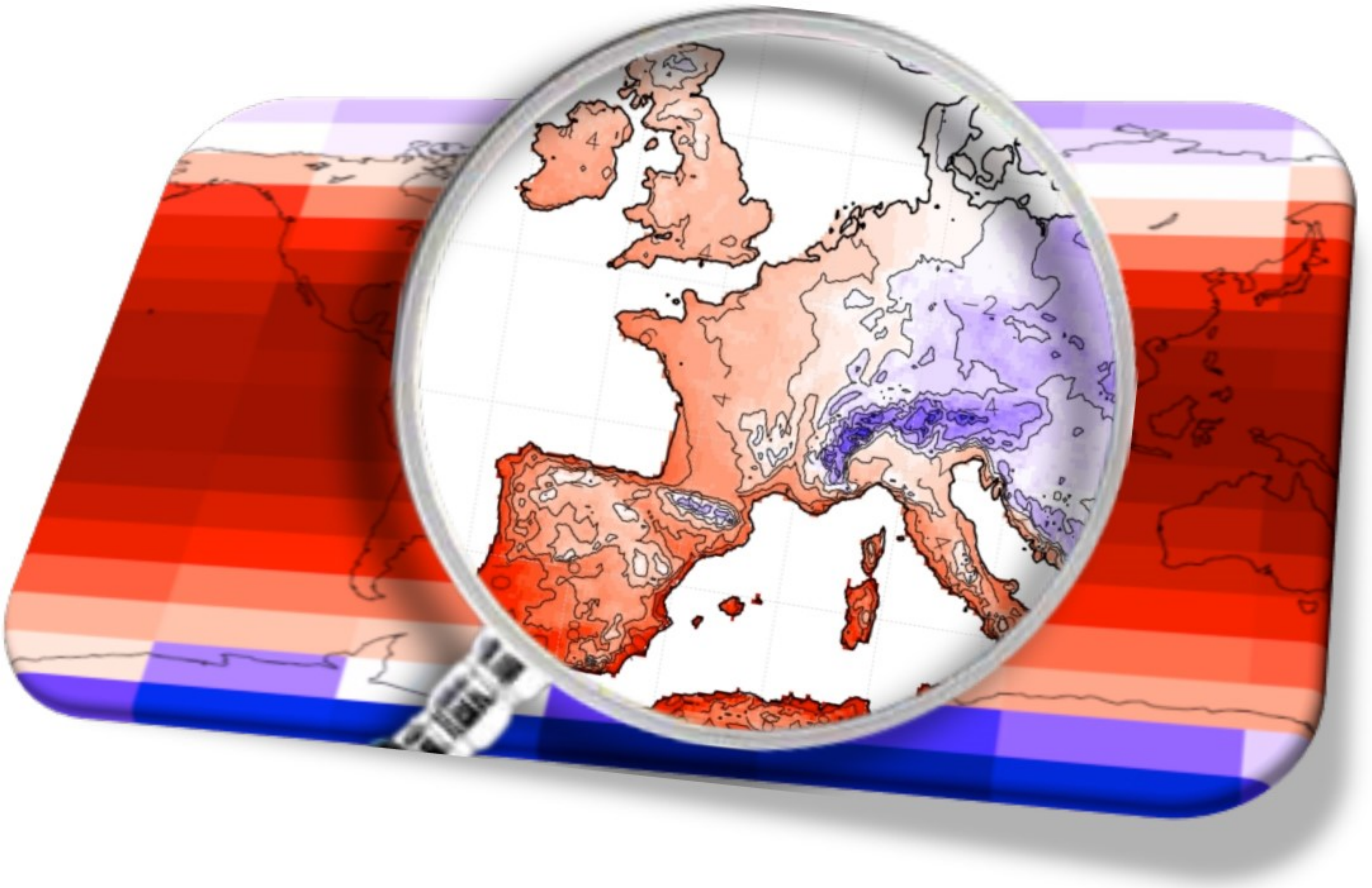
2020/10/01 (updated: 2021-01-05)

Downscaling

IPCC, 2013: Annex III: Glossary

> Downscaling is a method that derives local to regional scale (10 to 100 km) information from larger-scale models or data analyses.
>
> Two main methods exist: dynamical downscaling and empirical/statistical downscaling.
>
> The dynamical method uses the output of regional climate models, global models with variable spatial resolution or high-resolution global models.
>
> The empirical/statistical methods develop statistical relationships that link the large-scale atmospheric variables with local/regional climate variables.
>
> In all cases, the quality of the driving model remains an important limitation on the quality of the downscaled information.

Downscaling



Correction de biais

Biais (statistiques) : Soit T un estimateur d'un parametre θ. Le bias de l'estimateur T est défini par

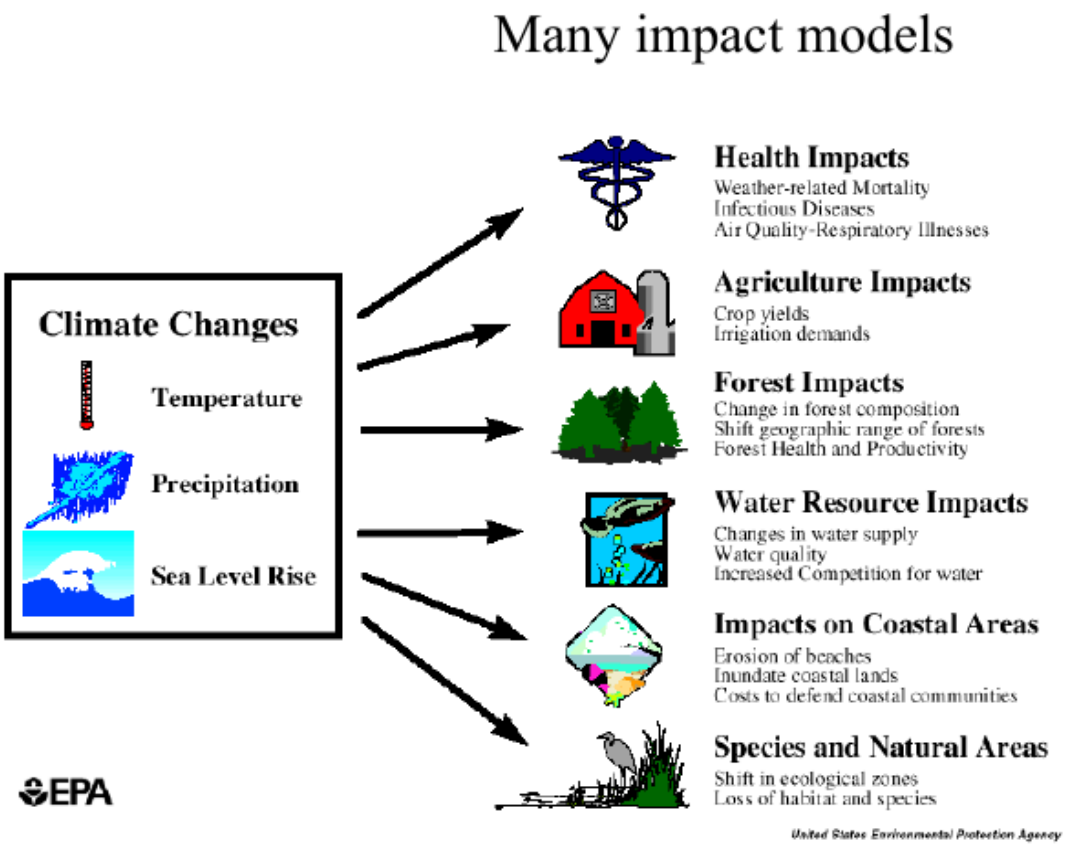
$$\mathbb{E}[T - \theta] = \mathbb{E}[T] - \theta$$

Maraun (2016)

> Transferring the bias concepts from statistics and forecast verification to a climate modelling context,
>
> a climate model bias can be defined as the systematic difference between a simulated climate statistic and the corresponding real-world climate statistic.
>
> A model bias derived from model and observational data is —as the statistics it is calculated from— only an estimate of the true model bias and therefore also affected by internal climate variability.

Remarque : le downscaling statistique peut être vu comme une forme de correction de bias où le bias est généré par la différence de résolution.

Pour l'étude des impacts



Downscaling statistique et correction de bias

Objectif trouver une relation statistique entre X, la variable à downscaller/corriger et Y, la variable de référence.

Deux familles de méthode:

- Perfect Prognosis : on travaille sur les réalisations x et y des variables X et Y,
- Model Output Statistics : on travaille sur loi de probabilité des variables X et Y

Perfect Prognosis

Trouver comment transformer une réalisation x_i en une réalisation y_i.

Un exemple de modèle classique

$$Y = f(X) + \epsilon$$

- approche par fonction de transfert / régression, on cherche à modéliser

$$\mathbb{E}[Y|X = x] = f(x)$$

- approche par générateur de temps stochastique, on cherche à modéliser la variable aléatoire

$$(Y|X = x) = f(x) + \epsilon$$

  en essayant d'estimer la fontion f et de modéliser la distribution du bruit ϵ .

Les predicteurs sont supposés parfaits.

On a besoin de l'appariement temporelles entre les variables X_i et Y_i.

Régression linéaire simple

Soit Y la variable locale que l'on cherche à prédire à partir de X, la variable à large échelle.

On modélise Y par régression linéaire

$$Y_i = aX_i + b + \epsilon_i$$

avec

- a et b, les coefficents de la régression linéaire à estimer.
- ϵ, un bruit que l'on suppose souvent gaussien.

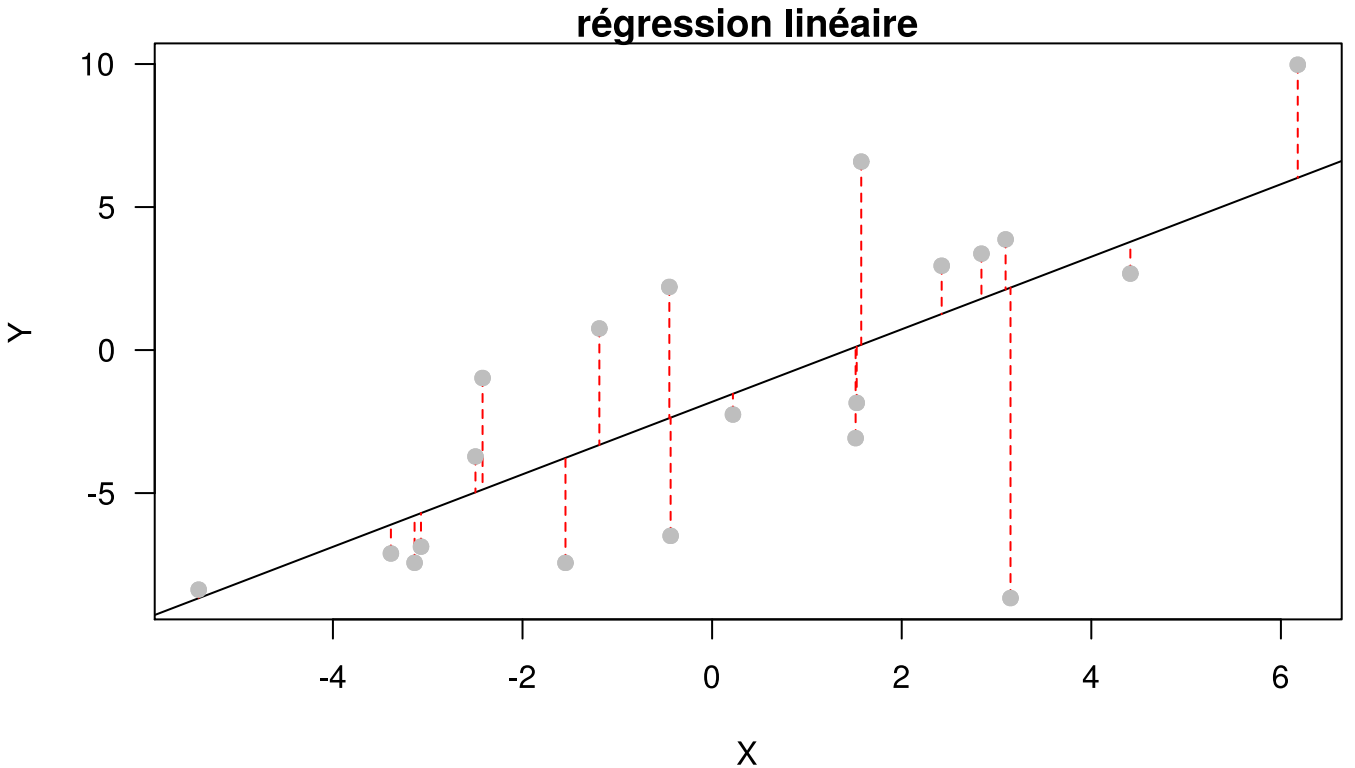
On cherche les coefficients $\hat{a}$ et $\hat{b}$ qui minimisent la fonction de coût

$$C(a, b) = \sum_i (Y_i - aX_i + b)^2$$

Pour une valeur X_{new}, la prédiction pour la valeur de Y_{new} correspondante est donnée par

$$\hat{Y_{new}} = \hat{a}X_{new} + \hat{b}$$

Exemple synthetique



Régression par plus proche voisin

En anglais, k-nearest-neighbour regression avec $k = 1$.

On définit la base d'apprentissage comme l'ensemble des couples (X_i, Y_i) observés. Pour une valeur X_{new}, la prédiction pour la valeur de Y correspondante est donnée par

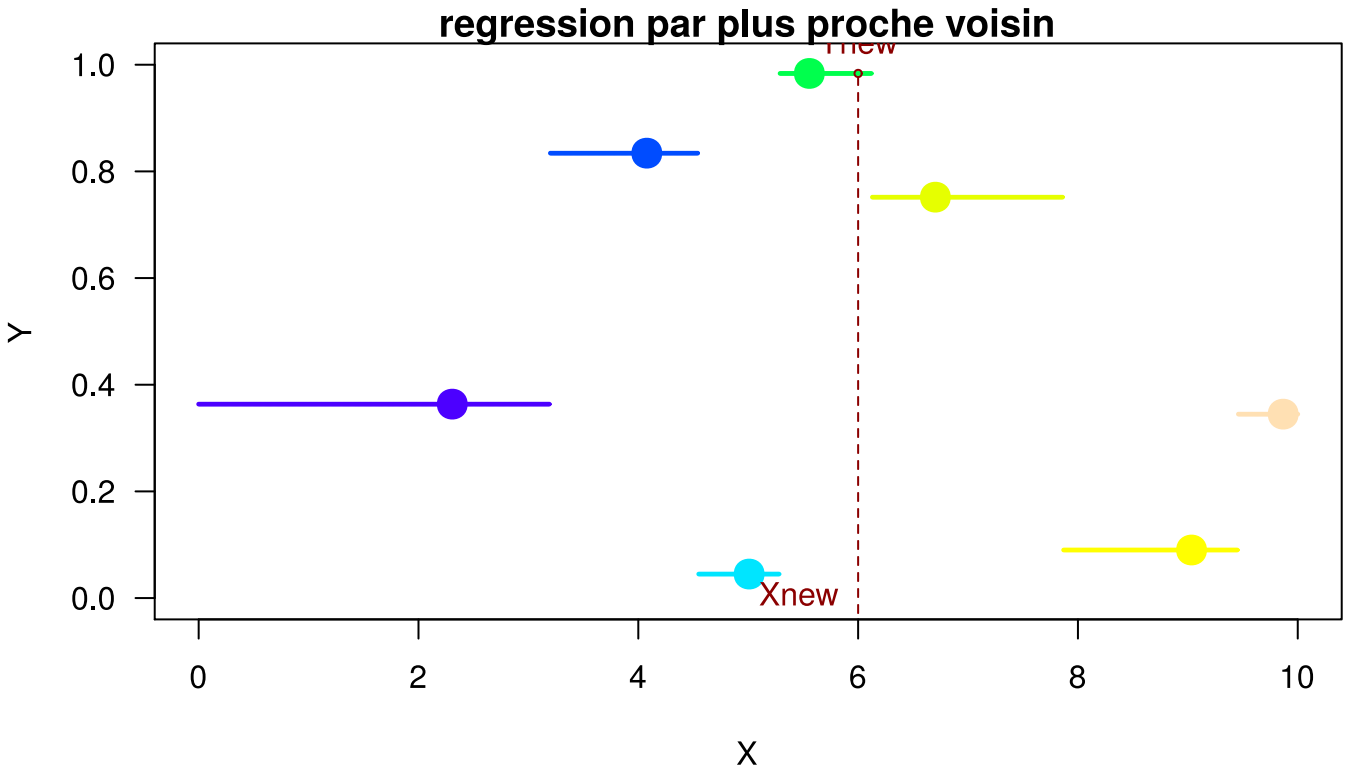
$$\hat{Y}_{new} = Y_{i^*}$$

avec

$$i^* = \arg\min_i \quad d(X_i, X_{new})$$

où d désigne une distance.

Exemple synthetique



Model Output Statistics:

Trouver comment transformer une variable aléatoire X en une variable aléatoire Y.

Soit $X \sim F$ et $Y \sim G$,

i.e. X et Y suivent les lois de probabilités qui sont définis respectivement par les fonctions de répartition F et G.

On cherche une transformation T tel que

$$T(X) \overset{d}{=} Y \quad \Leftrightarrow \quad T(X) \sim G$$

ou en tous cas que la loi de $T(X)$ se rapproche de la loi de Y.

Remarque: Il est aussi possible de ne corriger que certaines propriétés statistiques de la distribution de X,

e.g l'espérance ou la variance.

Correction de la moyenne

La variable large échelle X est simplement transformée pour avoir la même espérance (moyenne) que la variable Y.

Ainsi pour une valeur X_{new}, la prédiction pour la valeur de Y_{new} correspondante est donnée par

$$\hat{Y}_{new} = X_{new} - \mathbb{E}[X] + \mathbb{E}[Y]$$

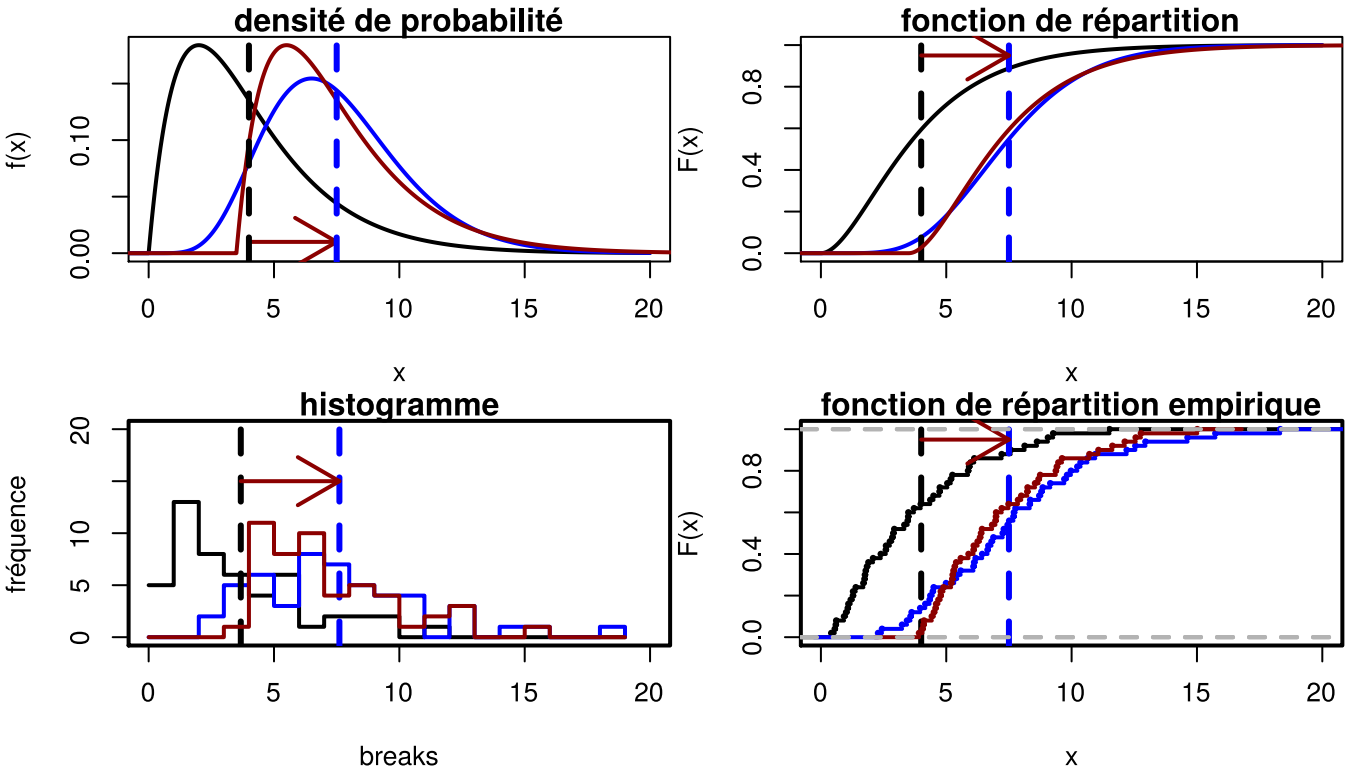
En pratique, on remplace l'espérance par la moyenne empirique

$$\hat{Y}_{new} = X_{new} - \bar{X} + \bar{Y}$$

en définissant la moyenne empirique de X par

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Exemple synthetique



La correction quantile-quantile

Soit $X \sim F$ et $Y \sim G$

On cherche une transformation T tel que

$$T(X) \overset{d}{=} Y \quad \Leftrightarrow \quad T(X) \sim G$$

Si F et G sont des fonctions de répartition continues et strictement monotones, alors

$$G(Y) \overset{d}{=} U(0,1) \overset{d}{=} F(X)$$

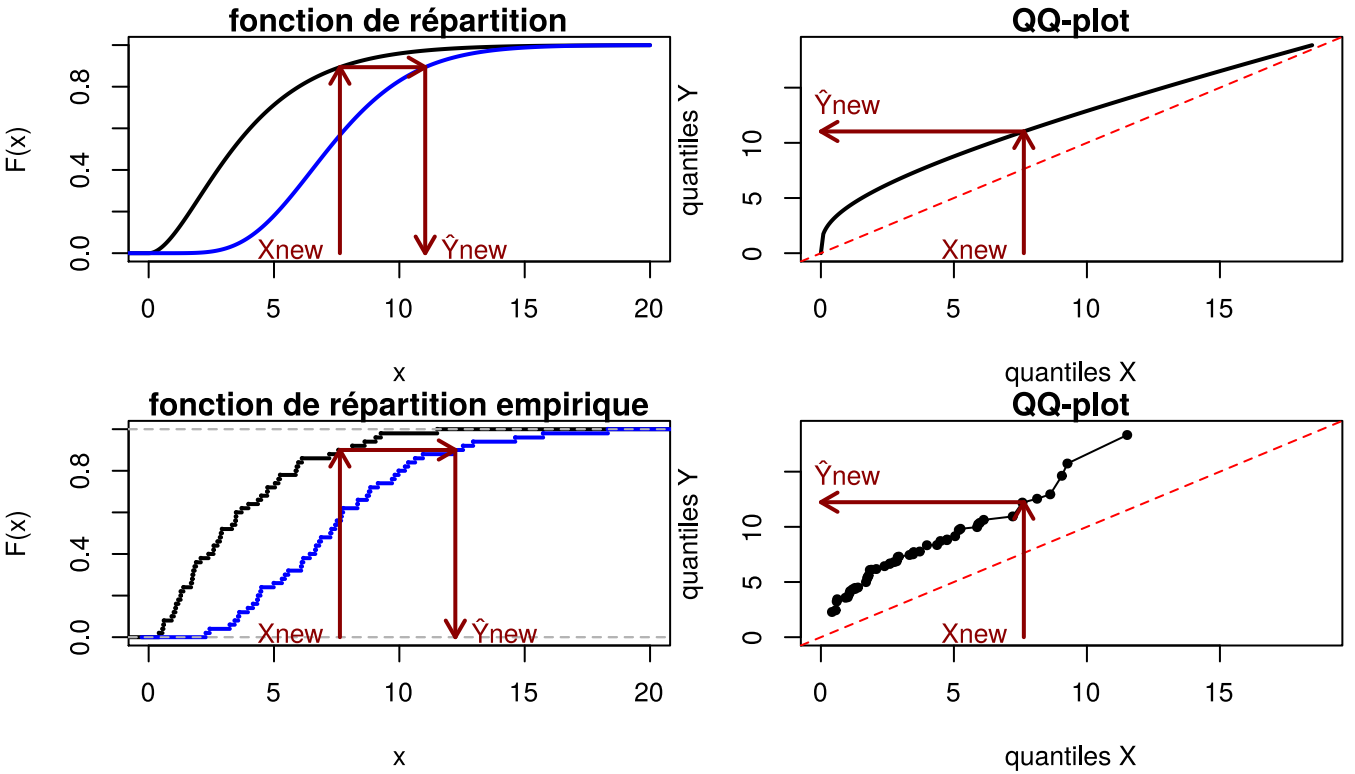
et

$$Y \overset{d}{=} G^{-1}(F(X)) \overset{d}{=} T(X)$$

En pratique, on remplace F et G par leur estimateur empirique

$$\hat{F}(x) = \mathbb{P}(X < x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$$

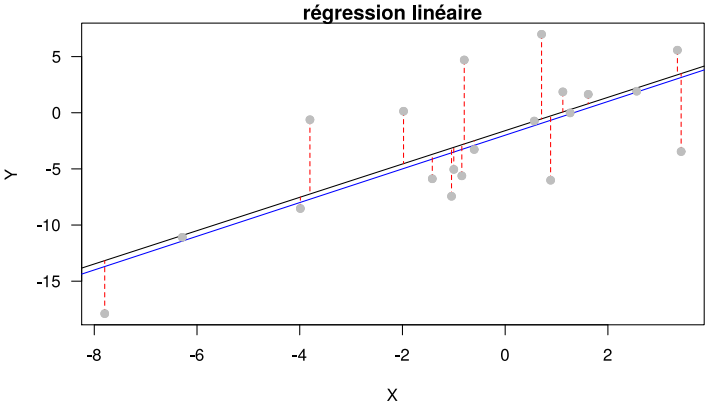
Exemple synthetique



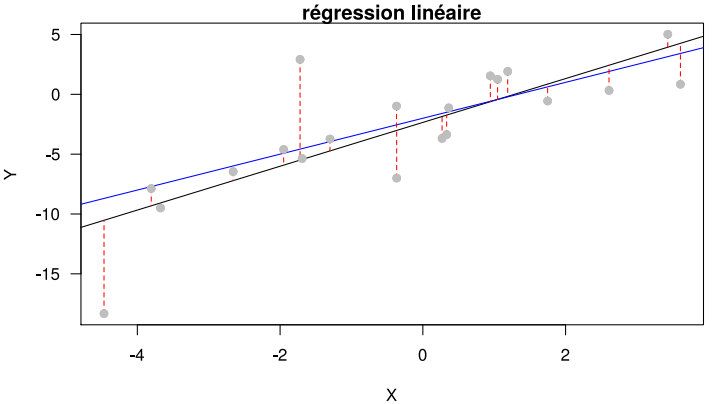
Erreur d'estimation

```
n <- 20
X <- rnorm(n, 0, 3)
Y <- 1.5 * X - 2 +  rnorm(n, 0, 4)
```

Echantillon 1



Echantillon 2



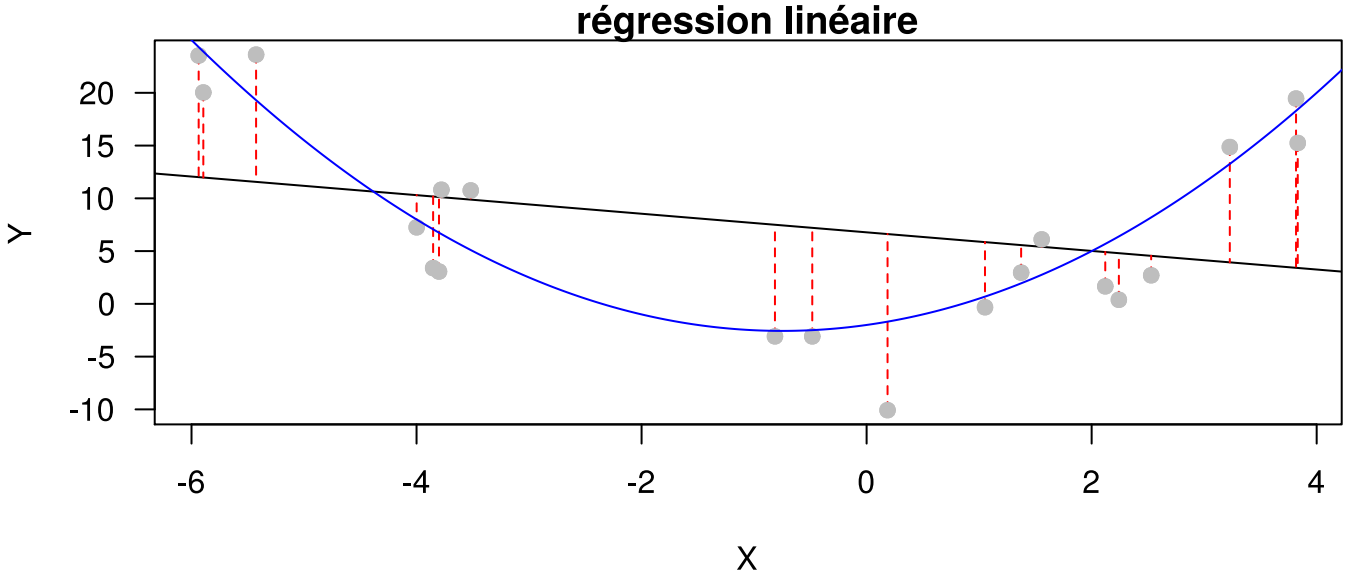
Le caractère aléatoire des données conduit à des estimations imparfaites.

Model misspecification

Wikipedia: *"In statistics, model specification is part of the process of building a statistical model: specification consists of selecting an appropriate functional form for the model and choosing which variables to include."*

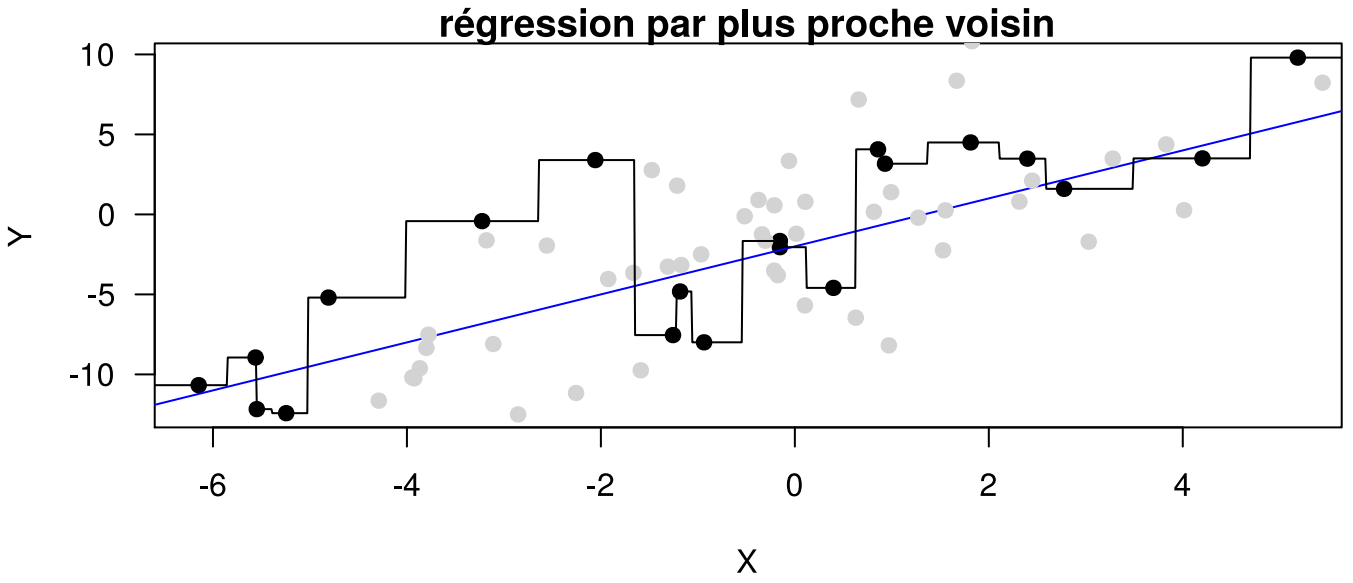
Exemple : modèle linéaire vs non-linéaire

```
n <- 20
X <- rnorm(n, 0, 3)
Y <- X^2 + 1.5 * X - 2 +  rnorm(n, 0, 4)
```



Sur-apprentissage

```
n <- 20
X <- rnorm(n, 0, 3)
Y <- 1.5 * X - 2 +  rnorm(n, 0, 4)
```



Le sur-apprentissage survient lorsque le modéle statistique essaie de reproduire des variations dans les données qui n'ont pas de signification, e.g. ici du bruit.

Plus un modéle est flexible, plus il est susceptible de sur-apprendre.

"All models are wrong, but some are useful"

George Box (1976):

> **2.3 Parsimony**
>
> Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.
>
> **2.4 Worrying Selectively**
>
> Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.

Pour les projections climatiques futures ?

Capacité de généralisation

> Capacité d'un modèle, une fois entraîné, à effectuer des prédictions sur des données qui n'ont pas servies à l'entrainement.

En science du climat, lorsque que l'on applique les méthodes de downscalling ou correction de bias sur les projections, on fait les hypothéses suivantes

A2: Les prédicteurs sont pertinents et modélisés de façon réaliste par le modèle climatique projections climatiques.

L'évolution de ces prédicteurs est également simulée de manière plausible.

A3: La relation établie par le modèle de downscaling statistique reste valide pour des conditions climatiques altérées.

Attention à l'extrapolation de la relation apprise. Cela génère parfois des valeurs non-physiques.

Exemple: la regression par plus proche voisin et la correction quantile-quantile sont des fonctions constante hors de la plage de valeurs observées.

Pour aller plus loin

Ici, nous n'avons parlé que de la correction de biais pour des variables univariées.

Pour le cas univarié, il existe un grand nombre de méthodes et de variantes.

Le système climatique est représenté par un ensemble de variables qui varient à la fois dans le temps et dans l'espace.

Pour corriger plusieurs variable à la fois, il faut faire appel à de la statistique multivariée.

Décider des aspects à corriger: marginales, dépendances spatiales, temporelles et inter-variables.

Certaines méthodes de correction de biais modifient les tendances du changement climatiques. Est-ce une bonne chose ?

Plus généralement, que souhaite-t-on préserver ou corriger du modèle de climat ?

A garder en tête

Maraun (2016)

> Bias correction is a mere statistical post-processing and cannot overcome fundamental mis-specifications of a climate model.

IPCC Workshop on Regional Climate Projections and their Use in Impacts and Risk Analysis Studies

> Bias correction (alternatively: bias adjustment or bias reduction) is a computationally inexpensive and pragmatic tool which, however, is also prone to misuse due to its mathematical simplicity
>
> We strongly discourage the application of BC without prior understanding of the underlying causes of model error and bias. In particular, it is important that users of bias-corrected data understand the source model's representation of physical processes (given that BC cannot compensate for incorrect representation of physical processes in the model). We recommend that BC is ideally carried out in collaboration with experts aware of the limitations of that particular model for the considered region (e.g., the developers of that model).