Projet RExHySS

Annexe Volet 1.
Méthode de désagrégation quantile/quantile conditionnelle

Michel Déqué
Météo-France/ CNRM

Septembre 2009

But

Cette méthode est plus une méthode de correction que de descente d'échelle, car elle s'appuie sur des simulations numériques à résolution 50 km, telles qu'obtenues avec ARPEGE étiré (Gibelin et Déqué, 2003), utilisé dans les projets FP6-ENSEMBLES et FP6-CECILIA. On ne cherche pas à reconstituer des structures cohérentes de petite échelle, mais juste à conserver les caractéristiques climatologiques de la petite échelle. La méthode est utilisée dans les projets ANR-CLIMATOR et GICC-REXHYSS, avec un traitement spécial dans ce dernier projet.

Historique

La méthode de correction a été introduite dans le projet GICC-IMFREX à partir de 32 stations du réseau synoptique de Météo-France (Déqué, 2003), puis plus de 100 stations du réseau SQR (Déqué, 2007). La technique a évolué au cours de ce projet. Le principe de départ était que dans une fourniture au grand public, un certain nombre d'indices ne parlent pas si on montre juste la différence climat futur moins climat présent (ou plutôt climat perturbé moins climat de référence) ou leur rapport. Il faut donc montrer deux cartes. Le site web IMFREX étant destiné à un public non nécessairement averti, afficher une carte qui pouvait être interprétée comme le climat présent alors que ce n'était qu'une sortie ARPEGE avec ses qualités et ses défauts, me répugnait un peu. Imaginons le professeur de géographie diffusant cette carte à ses élèves comme étant le climat de la France, ou l'entrepreneur se calant sur ces chiffres pour optimiser son rendement ... tous les bémols n'auraient pas suffit. Il fallait des cartes du climat présent qui ressemblent beaucoup au vrai climat présent. Dans le projet GICC-CARBOFOR, on s'était aussi heurté au problème que la maladie de l'encre du chêne rouge prospérait sur toute la France en « climat présent » parce qu'ARPEGE ne donnait jamais de température inférieure à -10°C. Il ne suffisait donc pas de corriger la moyenne en modifiant de 1°C toutes les températures. Il fallait s'attaquer à la fonction de densité probabiliste (pdf en anglais) de la variable.

Dans IMFREX, nous avons donc travaillé sur chaque variable séparément, à savoir température minimale diurne, température maximale diurne et précipitations. Si $F_m()$ est la fonction de répartition d'une variable X simulée par le modèle (par exemple le cumul quotidien de précipitations) en un point de grille et pour une saison, et si $F_o()$ est la fonction de répartition de la variable observée à une station proche de ce point, $Y=F_o^{-1}(F_m(X))$ est une variable qui a la même pdf que les observations. Cette technique revient à considérer que le modèle ne simule pas une variable physique mais un rang dans sa propre échelle de valeurs. Dans la pratique les deux fonctions de répartition sont obtenues en triant les séries et en gardant les deux extrêmes et les 99 centiles. Au delà des extrema les valeurs sont extrapolées en conservant la même correction additive que pour l'extremum correspondant. Dans certains cas (précipitations, humidité relative) on n'a pas d'extrapolation puisqu'il existe un extremum absolu d'un ou des deux côtés de la fonction.

CLIMATOR

Le projet CLIMATOR, coordonné par l'INRA, a pour but d'élaborer des outils et des références pour analyser la vulnérabilité des agrosystèmes face au changement climatique. Le végétal est sensible au climat et le modèle de végétal encore plus. Il est nécessaire d'alimenter les modèles végétaux avec des séries les plus proches possibles de séries d'observation, qui ont souvent servi à calibrer ces modèles. La correction d'IMFREX garantit cette propriété pour chaque variable individuellement. Dans CLIMATOR nous avons 12 sites (réseau d'observation INRA) pour lesquels 6 variables quotidiennes sont disponibles sur une période de référence. Ces variables sont corrélées et corriger l'une sans tenir compte de l'autre risque de conduire à un sextuplet physiquement irréaliste: par exemple des températures froides et des précipitations abondantes. L'idéal serait d'imposer la pdf du sextuplet en conditionnant chaque fonction de correction par rapport aux 5 autres variables. Mathématiquement la solution s'appelle l'espérance conditionnelle. Pour un vecteur gaussien c'est une fonction linéaire (qu'on appelle aussi régression linéaire multiple). Pratiquement dans un cas non-gaussien comme le nôtre il faudrait des millénaires de données pour une estimation précise.

J'ai donc proposé de tirer un avantage de notre désagrément. Si les variables étaient indépendantes, il suffirait de les traiter une par une comme dans IMFREX. Elles ne le sont pas, mais cette redondance peut nous aider à réduire la dimension du problème et la rendre statistiquement abordable.

Pour chacune des 12 stations et pour chaque saison, une analyse en composante principale (ACP) des observations est effectuée sur les vecteurs de taille 6. Naturellement on m'a appris à l'école qu'on n'additionne pas des carottes avec des patates et les données sont divisées par leur écart-type; on appelle cette technique l'ACP normée et la matrice à diagonaliser est la matrice de corrélation. Les deux premières composantes principales expliquent 70% du signal et elles vont être utilisées pour définir quatre « classes de temps » équiprobables. On commence par couper l'échantillon en deux par la médiane de la première CP (environ 40% du signal). Puis on calcule les médianes conditionnelles de la deuxième CP dans les deux sous-populations. On peut donc affecter à chaque jour de la série d'observations un numéro de classe compris entre 1 et 4. Pour le modèle, il faut que les classes signifient physiquement quelque chose de similaire aux classes de l'observation, afin de les apparier. Il faut en outre que pour chaque classe ait la même probabilité d'occurrence pour le modèle et pour l'observation, afin que la pdf du modèle corrigé soit la même que la pdf observée. Pour cela, on ne fait pas de nouvelle ACP, mais le calcul de la moyenne et de l'écart type normalisateur se fait avec les données du modèle (sur la période de référence 1971-2006). Les données du modèle sont alors projetées sur les mêmes axes que l'observation (deux premières CP). Les seuils d'attribution aux 4 classes (médianes) sont recalculées avec les données du modèle (toujours sur la période de référence 1971-2006) afin d'obtenir 4 classes équiprobables. On a donc pour chaque jour de chaque simulation un numéro de classe compris entre 1 et 4. Naturellement à la fin du scénario A2, les classes ne sont plus équiprobables.

La fonction de correction est ensuite construite pour chaque station, chaque saison, et chaque couple de classe. Il suffit pour cela de calculer les fonctions de répartition $F_m(X)$ et $F_o(X)$ pour chaque classe. Diviser par 4 la taille de l'échantillon, par rapport à la méthode IMFREX, permet de calculer encore des centiles, même si la précision des extrêmes est moins bonne. Mais il vaut mieux avoir des sextuplets cohérents dans 90% des cas que des extrêmes plus précis. On constate en particulier que la correction des précipitations n'est pas identique dans une classe « chaude, pluvieuse à forte humidité et à faible rayonnement global » que dans une classe « froide, humide et peu pluvieuse », de même que dans IMFREX on avait constaté que la correction n'était pas la même l'hiver et l'été.

REXHYSS

Une nouvelle difficulté apparaît dans ce projet, du fait qu'on a beaucoup plus de points d'observation (1662) que de points du modèle (58) sur la région Seine-Somme du projet. Les observations proviennent ici des analyses SAFRAN. Il n'est pas question de faire 1662 ACP.

La solution adoptée consiste à réduire la dimension du problème pour le rendre statistiquement abordable. Pour chacune des 1662 mailles et pour chaque saison, une analyse en composante principale (ACP) des observations (les données sont normalisées par leur écart-type et la matrice à diagonaliser est la matrice de corrélation) est effectuée sur des vecteurs de taille 7. Les deux premières composantes principales (CP) expliquent 70% du signal et elles vont être utilisées pour définir quatre « classes de temps » équiprobables. On commence par couper l'échantillon en deux par la médiane de la première CP (environ 40% du signal). Puis on calcule les médianes conditionnelles de la deuxième CP dans les deux sous-populations. On peut donc affecter à chaque jour de la série d'observations un numéro de classe compris entre 1 et 4. Pour le modèle, il faut que les classes signifient physiquement quelque chose de similaire aux classes de l'observation, afin de les apparier. Il faut en outre que chaque classe ait la même probabilité d'occurrence pour le modèle et pour l'observation, afin que la pdf du modèle corrigé soit la même que la pdf observée. Pour cela, on ne fait pas de nouvelle ACP, mais le calcul de la moyenne et de l'écart type normalisateur se fait avec les données du modèle (sur la période de référence 1970-2005). Les données du modèle sont alors projetées sur les mêmes axes que l'observation (deux premières CP). Les seuils d'attribution aux 4 classes (médianes) sont recalculées avec les données du modèle (toujours sur la période de référence 1970-2005) afin d'obtenir 4 classes équiprobables. On a donc pour chaque jour de chaque simulation un numéro de classe compris entre 1 et 4. Naturellement, à la fin du scénario climatique choisi (en 2100), les classes ne sont plus équiprobables.

La fonction de correction est ensuite construite pour chaque maille, chaque saison, et chaque couple de classe. Il suffit pour cela de calculer les fonctions de répartition Fm(X) et Fo(X) pour chaque classe. Diviser par 4 la taille de l'échantillon, par rapport à la méthode IMFREX, permet de calculer encore des centiles, même si la précision des extrêmes est moins bonne. Mais il vaut mieux avoir des septuplets cohérents dans 90% des cas que des extrêmes plus précis. On constate en particulier que la correction des précipitations n'est pas identique dans une classe « chaude, pluvieuse à forte humidité et à faible rayonnement global » que dans une classe « froide, humide et peu pluvieuse », de même que dans IMFREX on avait constaté que la correction n'était pas la même l'hiver et l'été. Soulignons également l’importance de l’échantillonnage temporel : effectuer la correction au pas de temps journalier plutôt que toutes les 6 heures améliore la représentation de la persistance des épisodes secs par rapport à SAFRAN (en revanche pas d’effet sur les épisodes pluvieux, voir annexe V1b).

Références

Déqué, M. (2003). Température et précipitations extrêmes sur la France dans un scenario de changement climatique. Actes du Colloque de l'Association Internationale de Climatologie, Varsovie, 2003, 4 pp.

Déqué, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. Global and Planetary Change, 57, 16-26 .

Gibelin A. L. and Déqué M. (2003). Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution model. Clim. Dyn., 20, 327-339.