

La préparation des données pour Scampeï

Les données brutes

Le projet Scampeï a utilisé quatre sources de données:

1. les analyses Safran qui servent d'observations de référence pour la période 1958-2011 (seule la période 1961-1990 est utilisée dans le projet). Ces données sont disponibles sous deux formes : les zones Symposium et la grille Safran
2. les simulations Aladin (CNRM/GAME) sur une grille à 12 km couvrant la France. Une simulation couvre 1960-2000 avec des conditions aux limites latérales venant du modèle Arpege global à grille étirée, de résolution 50 km sur l'Europe. Trois scénarios sont disponibles : le scénario A1B couvre 2001-2100, les scénarios A2 et B1 couvrent chacun deux périodes 2020-2050 et 2070-2100. Dans le projet, on n'utilise que les trois intervalles 1961-1990 (référence), 2021-2050 (futur proche) et 2071-2100 (futur lointain). On extrait 6341 points du modèle sur le territoire métropolitain.
3. Les simulations LMDZ (CNRS/LMD) sur une grille à 20 km couvrant la France. Ces simulations couvrent les trois intervalles de 30 ans ci-dessus et deux scénarios A1B pilotés par un modèle LMDZ à 50 km sur l'Europe piloté lui-même par un troisième modèle LMDZ global à résolution 300 km. Deux scénarios A1B sont utilisés, l'un avec les températures de la mer de l'IPSL, l'autre avec celles du CNRM (les mêmes que pour la simulation Aladin A1B ci-dessus). On extrait 2565 points du modèle sur le territoire métropolitain.
4. Les simulations MAR (CNRS/LGGE) sur une grille à 20 km couvrant la France. Ces simulations couvrent les trois intervalles de 30 ans ci-dessus pour le scénario A1B. Les conditions aux limites latérales sont les mêmes que pour le deuxième scénario LMDZ. On extrait 2279 points du modèle sur le territoire métropolitain.

Les analyses Symposium

Il s'agit d'observations ou de pseudo-observations horaires sur 615 zones climatologiquement homogènes à des tranches d'altitude espacées de 300m. Le nombre de tranches dépend de la région.

On a au total 2452 séries d'analyses pour les champs suivants:

- température à 2m (K)
- vent moyen à 10m (m/s)
- humidité relative de l'air à 2m (%)
- précipitations (mm)
- part des chutes de neige (0=pluie,1=neige)
- rayonnement infrarouge descendant (W/m²)
- rayonnement solaire direct (W/m²)
- rayonnement solaire diffus descendant (W/m²)
- nébulosité (de 0 à 1)

Un des buts de Scampeï est de fournir des séries de même type (même format, mêmes unités, même climatologie pour la période de référence) à partir des trois modèles de climat pour forcer le modèle de sol/neige/végétation ISBA/ES

La grille Safran

A partir des analyses précédentes, on a interpolé les données sur une grille de 8km, pour des besoins d'hydrologie. Il y a 9892 points sur cette grille (les coins, les points sur mer sont exclus) dont seulement 8602 couvrent le territoire métropolitain. Les points éliminés sont des extrapolations sur le Valais Suisse destinées à couvrir le bassin du Rhône dans son intégralité. Les variables sont:

- température 2m
- précipitations
- chutes de neige
- vitesse du vent

- humidité spécifique
- rayonnement solaire descendant
- rayonnement infra-rouge descendant

Un autre but de Scampeï est de cartographier des indices climatiques à partir des sorties des trois modèles, en se servant de ces données pour corriger et régionaliser les variables.

Le traitement des analyses Safran

On constitue d'abord des séries quotidiennes à partir des séries horaires

- température minimale à 2m
- température maximale à 2m
- précipitations liquides totales
- chutes de neige totales
- vitesse du vent moyenne
- vitesse du vent maximale
- humidité spécifique moyenne
- rayonnement solaire descendant moyen
- rayonnement infra-rouge descendant moyen

On effectue une classification automatique du géopotential quotidien à 500 hPa (données ERA40) pour déterminer pour chacune des quatre saisons (DJF, MAM, JJA, SON) quatre régimes de temps sur le domaine Europe-Atlantique. Pour chaque saison, chaque régime, chaque point de la grille Safran et chaque variable météorologique quotidienne on calcule 103 bornes:

1=0.

2=Q(.5%)

3=Q(1%)

....

101=Q(99%)

102=Q(99.5%)

103=Q(99.5%)+1000.

Où Q(p) représente la valeur pour une variable V telle que:

$$\text{Proba}(V < Q(p)) = p$$

Comme la température est en K, le choix des bornes 1 et 103 assure une correction quasi-constante pour les valeurs extrêmes en dehors de [Q(.5%),Q(99.5%)]. Le vent maximum diurne est corrigé à partir des valeurs de rafales observées de la climathèque.

Pour les zones Symposium, on constitue les 12 champs quotidiens suivants :

- température minimale à 2m
- température maximale à 2m
- précipitations liquides totales
- chutes de neige totales
- vitesse du vent minimale à 10m
- vitesse du vent maximale à 10m
- humidité relative minimale à 2m
- humidité relative maximale à 2m
- rayonnement solaire descendant moyen au sol
- rayonnement infra-rouge descendant minimal au sol
- rayonnement infra-rouge descendant maximal au sol
- nébulosité moyenne

Le vent maximum n'est pas corrigé par rapport aux observations, car il s'agit ici de fournir des forçages cohérents à un modèle se sol, pas d'étudier des fréquences d'extrêmes. Puis les 103 bornes sont calculées par régime et par saison pour les 2452 séries.

Le traitement des données Aladin sur la grille Safran

Ce qui suit concerne aussi les données MAR et LMDZ à quelques détails près : les données MAR sont horaires et il manque l'humidité relative qui est calculée par la formule de Tetens, les données LMDZ sont

bihoraires et il manque humidité relative et nébulosité, dans les deux cas les années font 360 jours. On commence par calculer les régimes de temps à partir des modèles à résolution 50 km sur l'Europe (Arpege ou LMDZ) en utilisant les quatre étalons d'ERA40. On affecte chaque point de la grille Aladin sur le point de la grille Safran le plus proche, et on calcule les 103 bornes par saison et par régime pour la période 1961-1990 comme ci-dessus. Puis on associe chaque borne du modèle à la borne correspondante dans Safran pour créer une table d'étalonnage. Les sorties 6-horaires d'Aladin sont agrégées en données quotidiennes, puis corrigées par interpolation linéaire par ces tables en tenant compte du régime du jour. On appelle cette opération la correction quantile-quantile. Le choix des bornes 1 et 103 assure qu'une valeur modèle est toujours comprise entre deux bornes (pas d'extrapolation à envisager). On applique la même correction aux sorties du modèle du futur.

Le traitement des données Aladin sur les zones Symposium

Ce traitement est identique pour les trois modèles Aladin, MAR et LMDZ. La première opération est de ramener les données Aladin sur les 615 zones Symposium par affectation au plus proche voisin. On ne tient pas compte de l'altitude du point Aladin, la correction quantile-quantile s'en chargera. On calcule les mêmes 12 champs quotidiens que pour les analyses. On calcule les 103 quantiles sur la période de référence. Puis on effectue une correction quantile-quantile pour ces champs. Dans le cas de la nébulosité, de l'humidité relative minimale et maximale, on a un souci particulier, car les valeurs modèle 0% ou 100%, qui résultent d'une troncature des données hors-plage dues à la compression au format Grib, correspondent à plusieurs valeurs Symposium dans la table de correction. On effectue un tirage aléatoire parmi les valeurs corrigées possibles, afin d'assurer qu'après correction, la distribution probabiliste des données modèles et celle des analyses soient identiques. Cette identité n'est possible que si la fréquence des régimes du modèle dans la période de référence est identique à celle d'ERA40. C'est à peu près le cas pour Arpege comme pour LMDZ.

La correction quantile-quantile associe la même série Aladin aux diverses séries Safran de la même zone à des altitudes différentes. Ainsi, à partir de 615 séries brutes, on obtient 2452 séries corrigées. Une fois corrigées, la pluie et la neige sont additionnées pour ne considérer que les précipitations totales.

Il reste à reconstituer les valeurs horaires, et calculer la fraction neige et la répartition direct/diffus du solaire. On choisit aléatoirement pour chaque jour Aladin un jour Safran de même mois, même régime de temps et même type de précipitation (< 1 mm/j ou > 1 mm/j). Tous les points d'une même zone Symposium ont le même jour (on utilise le premier point rencontré pour déterminer le type de précipitation). Il faut entre 100 et 300 tirages aléatoires pour trouver le bon jour Safran pour chaque jour Aladin et chaque zone Symposium. Dans quelques cas, le nombre d'itérations dépasse 1000. On décide de faire une recherche chronologique (ce qui privilégie les premières années de Safran) si 450 tirages échouent, car le tirage porte sur environ 900 dates possibles dans Safran. On crée des cycles diurnes artificiels pour 9 variables à partir des 11 champs Aladin corrigés. Si $A(h)$ est la valeur horaire d'Aladin corrigé et $S(h)$ la valeur horaire de Safran pour la journée choisie aléatoirement, on utilise des formules différentes suivant les variables à traiter. Pour température, vent, humidité relative et infrarouge descendant, on écrit:

$$A(h)=a*S(h)+b$$

a et b sont calculés pour que les maxima et minima de la journée soient ceux d'Aladin

Pour les précipitations et le rayonnement solaire descendant, on écrit:

$$A(h)=a*S(h)$$

a est calculé pour que la moyenne quotidienne de la journée soit celle d'Aladin.

Pour le rayonnement solaire descendant, on utilise le même coefficient pour direct et diffus, puisque on ne connaît que le flux total dans Aladin.

Pour la fraction de neige des précipitations on met 0 ou 1 suivant un seuil de température déterminé préalablement à partir des données Safran et qui dépend du lieu et de la saison. Ce seuil varie entre 273 K et 275 K avec quelques cas (0.4%) à 276 K et dépend de l'heure de la journée, de la saison, de l'altitude et de la zone Symposium (pas du régime de temps).

Pour la nébulosité, on écrit:

$$A(h)=a*S(h)+b$$

a et b sont calculés pour que la moyenne quotidienne de la journée soit celle d'Aladin et que la nébulosité corrigée reste entre 0 et 1 en essayant d'atteindre une des deux bornes.

Si la correction quantile-quantile conduit un jour donné à des minimum et maximum diurnes inversés (pour la température, le vent, l'humidité relative ou le rayonnement infrarouge descendant) on ré-ordonne les deux extrêmes.

Cette stratégie suppose que la cohérence spatiale et temporelle vient d'Aladin, tandis que l'échelle de valeurs

et la structure diurne viennent de Safran. On n'a aucune garantie de continuité temporelle à 6h, puisque une nouvelle chronologie aléatoire est tirée, ni de garantie de continuité spatiale entre les zones Symposium à une heure donnée. La cohérence n'est garantie (par Aladin) que sur les paramètres diurnes agrégés, avec de possibles discontinuités aux changements de régime ou de saison. Il ne faut donc pas considérer les séries horaires comme des simulations chronologiques (au sens des observations ou des sorties de modèle), mais comme des outils pour forcer le modèle de sol/neige/végétation Isba/ES de la moins mauvaise façon:

- en respectant le plus possible la climatologie actuelle
- en bridant le moins possible la climatologie future