Introduction à NetCDF

École normale supérieure
L3 sciences de la planète
2021/2022

Lionel GUEZ
guez@lmd.ipsl.fr
Laboratoire de météorologie dynamique
Bureau E212

Qu'est-ce que NetCDF ?

- Un format de fichier destiné principalement à stocker des tableaux de nombres multi-dimensionnels (et donc à l'informatique scientifique).
  - Convention : suffixe “.nc”
- Des bibliothèques de procédures dans différents langages pour créer, lire, modifier des fichiers dans ce format. On dit encore : des API, pour *application programming interface*. Une API en Fortran, une en C, etc.

Quelques perspectives (1/2)

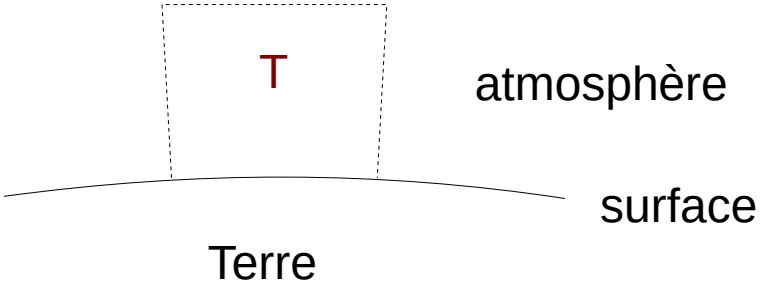
- NetCDF est développé par Unidata (organisme public américain).
- Depuis 1988
- Les bibliothèques NetCDF sont libres et gratuites.

Quelques perspectives (2/2)

- Énormément utilisé en météorologie et en océanographie. Moins hégémonique mais utilisé aussi en géologie.
- Des dizaines de logiciels tiers gratuits pour manipuler (découper, assembler, faire des moyennes…) et visualiser les fichiers NetCDF.

Stocker un champ multi-dimensionnel (1/3)

- Exemple : la température T (mesurée ou calculée) dans un certain domaine de l'atmosphère (tri-dimensionnel)
- On veut stocker non seulement les valeurs de T mais la position associée à chaque valeur.



Stocker un champ multi-dimensionnel (2/3)

- Idée de coordonnées : on stocke des vecteurs de coordonnées. Trois vecteurs pour un champ tri-dimensionnel.
- Exemple :
  longitude(15 valeurs), latitudes(8), altitude(12) pour un tableau de températures de profil (15, 8, 12). T(1, 3, 2) est la valeur de température à la position longitude(1), latitude(3), altitude(2).

Stocker un champ multi-dimensionnel (3/3)

- Nota bene : les vecteurs de coordonnées prennent peu de place par rapport au champ.
- On veut stocker aussi des « métadonnées » : l'unité du champ (dans l'exemple de la température : en K), d'où viennent les valeurs (modèle, observations…) etc.

Format auto-descriptif

- Un fichier au format NetCDF contient toutes ces informations : non seulement des champs multi-dimensionnels mais aussi les coordonnées et les métadonnées.
- On dit donc que le format NetCDF est auto-descriptif.

Format binaire

- Le format NetCDF est « binaire ».
- D’un point de vue pratique : on ne peut voir (intelligiblement) le contenu d’un fichier NetCDF ni avec la commande Unix cat, ni avec un éditeur de texte, ni avec un traitement de texte, ni avec un tableur…

Avantages du format binaire (1/2)

- Taille réduite de fichier
  Exemple : 8.328622e-17
  4 octets en binaire, 12 octets en texte
- Rapidité d'exécution des instructions de lecture ou d’écriture
  (économie du temps de conversion à partir des caractères ou vers des caractères)

Avantages du format binaire (2/2)

- Exactitude : la valeur stockée est exactement la valeur utilisée par le programme.
  - Alors que la représentation textuelle passe par un changement de base, entre bases 2 et 10.
  - Pas forcément de représentation exacte binaire d'un décimal.
  - Exemple :
    $1 / 10 = 0,1 = (0,000110011001\ldots)_2$
    $= 1/16 + 1/32 + 0/64 + 0/128 + 1/256 + 1/512 + 0/1024 + \ldots$

NetCDF : les avantages du binaire sans les inconvénients (1/2)

- L’organisation interne (tableaux multi-dimensionnels, coordonnées, métadonnées) est fixée par NetCDF, gérée automatiquement (et cachée) en écriture comme en lecture.

NetCDF : les avantages du binaire sans les inconvénients (2/2)

- Portable : on peut créer un fichier sur une machine, avec un logiciel ou un langage de programmation donné (avec un compilateur donné si c'est un langage compilé) et le relire sans plus de difficulté sur une autre machine, avec un autre logiciel ou langage (un autre compilateur).
- Fichier NetCDF immédiatement lisible avec des utilitaires gratuits. Notamment : ncdump.

Voir le contenu d'un fichier NetCDF

- Utilitaire ncdump (toujours inclus dans l'installation de NetCDF)
- `ncdump file`
  affiche tout le contenu du fichier (la commande convertit le binaire en texte pour l'afficher).
  En général, pas très lisible parce que :
  - trop long
  - les tableaux multi-dimensionnels sont écrits comme des tableaux à une dimension

L'en-tête d'un fichier

- `ncdump -h file`
  affiche seulement “l'en-tête” (“-h” pour “header”) du fichier, c'est-à-dire les méta-données.

Les trois parties de l'en-tête (1/3)

- Dimensions : elles ont un nom et une valeur entière

Les trois parties de l'en-tête (2/3)

- Variables :
  - type, nom, dimensions entre parenthèses
  - Les dimensions des variables sont celles nommées dans la partie dimensions de l’en-tête du fichier
  - 0 (scalaire), 1 ou plusieurs dimensions
  - Pour chaque variable, des “attributs”. Quelques attributs importants : `units`, `long_name`, `missing_value`

Les trois parties de l'en-tête (3/3)

- Attributs globaux : méta-données pour l'ensemble du fichier, et non pour une variable

Coordonnées (1/2)

- Coordonnée NetCDF : variable NetCDF à une dimension, dont le nom est identique au nom de la dimension.
- En général : une coordonnée par dimension
- Association de valeurs d'une coordonnée aux indices dans une dimension d'un tableau

Coordonnées (2/2)

- Les coordonnées sont reconnues et traitées spécialement par de nombreux utilitaires et programmes qui lisent les fichiers NetCDF.
- Convention : une coordonnée doit être strictement monotone.

Valeurs des coordonnées

- `ncdump -ct file`
- Fait apparaître l'en-tête déjà affiché par `ncdump -h`, plus une nouvelle partie “data”, qui contient les valeurs des coordonnées.

Variables primaires

- Ce sont les variables du fichier NetCDF autres que les coordonnées.
- Pour voir les valeurs d'une variable primaire :
  `ncdump -v var1,… file`
  Utile pour une variable scalaire ou à une dimension, mais peu lisible pour une variable multi-dimensionnelle

Valeurs d'une variable multi-dimensionnelle (1/2)

- Plus puissant que ncdump : ncks (“ks” pour “kitchen sink”)
  ncks est un des opérateurs NCO
  http://nco.sourceforge.net
  NCO : NetCDF operators, à installer en plus de NetCDF
- ncks permet de n'afficher qu'une partie d'une variable : la valeur en un point particulier, ou une “hyper-tranche” quelconque de la variable

Valeurs d'une variable multi-dimensionnelle (2/2)

- `ncks -v var[,...] \`
  `-d dim,[min][,[max]] input-file`
- Valeurs min et max dans l'argument `-d` :
  - sans point décimal : indice
  - avec point décimal : valeur de coordonnée

Logiciels graphiques (1/2)

- Très nombreux logiciels permettant de faire des graphiques à 1, 2 ou 3 dimensions, à partir de variables lues dans un fichier NetCDF

Logiciels graphiques (2/2)

- Une toute petite sélection, du plus simple aux plus complets :
  - ncview
  - Panoply
  - Grads (orienté analyse de données atmosphériques), Ferret (orienté analyse de données océaniques et atmosphériques), GMT (orienté géophysique)
  - Python avec netCDF4 ou xarray et matplotlib, Matlab, IDL

Création ou modification d'un fichier NetCDF

Interfaces dans de nombreux langages :

- développées par Unidata : interfaces en C, Fortran, Java
- développées par des tiers : interfaces en Python (modules netCDF4 et xarray), R, Ruby, C++, Matlab…

Conclusion

- Pour stocker des scalaires ou des vecteurs (en nombre fixé au moment de l'écriture du programme), écrivez des fichiers séquentiels au format texte.
- Pour stocker des tableaux de nombres multi-dimensionnels, écrivez des fichiers NetCDF.