Chuyên đề I. Download và xử lý ban đầu số liệu tái phân tích

Người thực hiện:

1. Mở đầu

Bộ số liệu tái phân tích khí tượng là bộ số liệu khí tượng được xây dựng lại từ bộ số liệu quan trắc và các bộ số liệu khác được thu thập trong quá khứ như quan trắc tại trạm, đo đạc, ghi chép, vệ tinh ...) bằng một phương pháp đồng hóa đồng nhất cho một giai đoạn nhất định.

Trong hoạt động dự báo thời tiết cũng như nghiên cứu khí hậu bằng phương pháp số, các mô hình dự báo sử dụng thông tin trạng thái điều kiện ban đầu của khí quyển để dự báo/dự tính trạng thái khí quyển trong tương lai. Trạng thái ban đầu cung cấp thông tin đầu vào cho việc dự báo phải bao gồm một số thông tin, đủ để mô hình xác định sự tiến triển của trạng thái khí quyển trong tương lai. Các thông tin này trong quan trắc có thể không đầy đủ hoặc cần tính toán một cách gián tiếp. Ngoài ra bộ số liệu tái phân tích cần xây dựng trên lưới đồng nhất, cần xác định giá trị tại tâm lưới hoặc các nút lưới mà thông tin quan trắc thu thập trên trạm không đồng đều về mặt không gian, thời gian và có thể chứa đựng trong nó một số sai sót giá trị quan trắc. Kỹ thuật đồng hóa dữ liệu được sử dụng để xây dựng lại một bộ số liệu mới dựa trên tất cả những thông tin thu nhận, trong đó số liệu được phân tích đồng nhất về mặt không gian, thời gian và bao gồm cả việc chỉnh lý các sai số.

Ngoài việc được sử dụng để tái tạo lại bộ số liệu trong quá khứ, số liệu tái phân tích còn là công cụ có giá trị để nghiên cứu khí hậu trong tương lai. Bộ số liệu tái phân tích là cơ sở để có thể đánh giá được năng lực của một mô hình dự báo, từ cơ sở các kết quả đánh giá thu được có thể nhận biết được năng lực của mô hình trước khi phục vụ dự báo hoặc dự tính trong tương lai. Độ tin cậy của số liệu tái phân tích phụ thuộc vào nguồn thu thập số liệu, vì vậy có sự khác biệt tùy thuộc vào vị trí, thời gian và biến (trường giá trị)

Các bộ số liệu tái phân tích từ dự án đầu tiên đến nay đã qua 3 mốc cải tiến quan trọng:

Giai đoạn đầu: Từ những ý tưởng đầu tiên, trung tâm NCEP/NCAR đã nỗ lực xây dựng bộ số liệu tái phân tích NCEP/NCAR (version 1) và NCEP-DOE (version 2 – được nâng cấp từ version1)

Giai đoạn 2 với sự phát triển của số liệu vệ tinh và nhiều nguồn số liệu khác, chất lượng của số liệu tái phân tích đã được nâng cao một cách đáng kể với 1 số sản phẩm (ERA40, JRA25) tương ứng. Giai đoạn 2 này, số liệu tái phân tích đã được cải tiến và bổ xung đồng hóa bộ đầy đủ các bức xạ tại đỉnh khí quyển mà vệ tinh thu thập được. Ví dụ, trong giai đoạn đầu, NCEP-NCAR và NCEP-DOE không đồng hóa các số liệu vệ tinh như quá trình vận chuyển ẩm, trên đại dương. Các thông lượng bức xạ tại đỉnh khí quyển. ERA40 hiện nay vẫn là bộ số liệu được sử dụng khá phổ biến trong các nghiên cứu khí hậu.

Giai đoạn thứ 3, với các sản phẩm như ERA-Interim, MERRA, CFSR, sử dụng phương pháp tiếp cận đồng hóa dữ liệu đa chiều hơn (ví dụ như phân tích và đồng hóa 4D-Var), nhằm giải quyết được các nhước điểm còn tồn tại ở giai đoạn 2.

Thách thức cũng như một số bài toán trong tương lai đang được tiếp cận, chẳng hạn như xây dựng bộ số liệu tái phân tích lại thế kỷ 20, đây sẽ tiếp tục là một thách thức bởi, nó không bao gồm phần lớn các thông tin quan sát hiện đại trong thời gian như gần đây.

Giới thiệu một số trung tâm xây dựng và sản phẩm tái phân tích

Do kĩ thuật xây dựng số liệu tái phân tích hết sức phức tạp và khối lượng công việc đồ sộ nên chỉ một số các trung tâm lớn trên thế giới có những dự án xây dựng các bộ số liệu tái phân tích:

Trung tâm cảnh báo môi trường Hoa Kỳ NCEP (United States National Centers for Environmental Prediction).

Trung tâm nghiên cứu khí quyển quốc gia Hoa Kỳ NCAR (National Center for Atmospheric Research).

Trung tâm dự báo thời tiết hạn vừa Châu Âu ECMWF (European Centre for Medium-Range Weather Forecasts)

Các sản phẩm tái phân tích đầu tiên, ERA-15, tạo ra lại phân tích cho khoảng 15 năm, từ tháng 12 năm 1978 đến tháng Hai năm 1994. Sản phẩm thứ hai, ERA-40 (ban đầu là tái phân tích 40 năm) bắt đầu vào năm 1957 và bao gồm 45 năm đến năm 2002. ERA40 là sản phẩm tái phân tích mở rộng được hoàn thiện hơn thay thế bộ ERA15. Gần đây, ECMWF đã phát hành bộ số liệu ERA-interim, với giai đoạn từ năm 1979 đến nay. Bộ số liệu này đã được đồng hòa toàn bộ số liệu vệ tinh.

Có khá nhiều sản phẩm số liệu tái phân tích khí tượng được cung cấp miễn phí phục vụ nghiên cứu khí hậu, các sản phẩm có thể làm công cụ nghiên cứu khí hậu trong quá khứ. Nó có thể coi là một bộ số liệu "quan trắc hoàn chỉnh" trên không gian lưới. Trong phần 3 này, sẽ trình bầy hướng dẫn chi tiết cách download và sử dụng bộ số liệu tái phân tích ERA40- bộ số liệu hiện đang được khai thác và sử dụng khá phổ biến trên thế giới và bộ số liệu CFSR-bộ số liệu thuộc diện mới nhất và hoàn chỉnh nhất hiện nay.

2. Hướng dẫn cấu trúc và download bộ số liệu tái phân tích

Bộ số liệu ERA40 có thể được download miễn phí trên mạng internet một cách dễ dàng. Tùy thuộc bài toàn cụ thể hoặc mục đích, người sử dụng có thể download một phần hoặc toàn bộ tập số liệu. Ngoài ra một số trang web cũng cung cấp kết quả bộ số liệu hiện thị dưới dạng ảnh trên bản đồ của tập kết quả. Dưới đây là hướng dẫn chi tiết cấu trúc của bộ số liệu tái phân tích ERA40 và cách download và sử dụng bộ số liệu này.

Về cở bản để Cấu trúc bộ số liệu và cách thức download gồm một số bước cơ bản như sau:

1.Truy cập vào trang chủ ECMWF: http://data-portal.ecmwf.int/

2.Trong bảng Datasets có các sản phẩm tái phân tích của ECMWF như: DEMETER, ENACT, ERA interim, ERA40, ERA15 .. như trên hình 3.1. Để chọn bộ số liệu ERA40 chúng ta tiến hành chọn (kích đúp) trong trang Datasets.

2.1. Bóc tách dữ liệu trước khi download

Các file số liệu CFS được lưu trữ dưới định dạng GRIB2 (GRIB edition 2). Mỗi file GRIB thực chất là một tập hợp các bản ghi (mục) dữ liệu riêng và hoàn toàn có thể đứng độc lập. Chính vì vậy một tập các bản ghi GRIB có thể nối với nhau để thành 1 file GRIB và ngược lại, một file GRIB có thể được tách ra thành một tập bản ghi GRIB mà không làm mất thông tin trong các bản ghi.

Thông tin bản ghi của 1 file GRIB2 có thể lấy được bằng công cụ wgrib2. [Bảng 1] trích đoạn thông tin bản ghi của file số liệu pgbf2012120100.01.2012051400.grb2 khi chạy với wgrib2. Thông tin này được gọi là index, mỗi file GRIB sẽ có một index khác nhau. Mỗi bản ghi bao gồm 6 trường, ngăn cách với nhau bằng một dấu : (xem [Bảng 2]).

```
1:0:d=2012051400:PRES:mean sea level:4824 hour fcst:
2:67691:d=2012051400:HGT:1 mb:4824 hour fcst:
3:104278:d=2012051400:TMP:1 mb:4824 hour fcst:
4:121396:d=2012051400:RH:1 mb:4824 hour fcst:
5:125244:d=2012051400:SPFH:1 mb:4824 hour fcst:
6:144113:d=2012051400:VVEL:1 mb:4824 hour fcst:
7.1:214111:d=2012051400:UGRD:1 mb:4824 hour fcst:
7.2:214111:d=2012051400:VGRD:1 mb:4824 hour fcst:
8:257145:d=2012051400:ABSV:1 mb:4824 hour fcst:
9:291477:d=2012051400:O3MR:1 mb:4824 hour fcst:
10:343977:d=2012051400:HGT:2 mb:4824 hour fcst:
11:386650:d=2012051400:TMP:2 mb:4824 hour fcst:
12:403521:d=2012051400:RH:2 mb:4824 hour fcst:
13:406060:d=2012051400:SPFH:2 mb:4824 hour fcst:
14:440273:d=2012051400:VVEL:2 mb:4824 hour fcst:
15.1:510005:d=2012051400:UGRD:2 mb:4824 hour fcst:
...
```

Bảng 1. Thông tin bản ghi GRIB2

TT	Địa chỉ Byte bắt đầu (tính từ đầu file)	Thời điểm bắt đầu dự báo	Trường Khí tượng	ực	Thời điểm dự báo (tính từ thời điểm bắt đầu)

Bảng 2. Cấu trúc bản ghi GRIB2

Như vậy, mỗi bản ghi trong file số liệu CFS là dữ liệu của một trường khí tượng trên một mực nhất định và ở một thời điểm dự báo nhất định. Từ index của file CFS, bản ghi tương ứng với trường khí tượng cần thiết có thể được trích ra một cách dễ dàng. Ví dụ, bản ghi chứa dữ liệu của trường PRES tại mực "mean sea level" có thể trích dễ dàng từ file pgbf2012120100.01.2012051400.grb2 với index trong [Bảng 1] bằng cách đọc từ byte số 0 tới byte số 67690. Như trên đã trình bày, bản ghi này hoàn toàn có thể đứng độc lập hoặc ghép với các bản ghi khác để tạo thành file GRIB lớn hơn.

Dải địa chỉ byte ứng với một (hoặc một tập) các bản ghi được gọi là byte-range. Trong ví dụ trên, byte range của PRES trên mực "mean sea level" tại thời điểm dự báo 4824h là 0-67690. Khi biết thông tin này, dữ liệu của bản ghi có thể được tải trực tiếp từ internet với lệnh `curl'. Ví dụ:

```
curl -r 0-67690 \
        http://nomads.ncep.noaa.gov/pub/data/nccf/com/cfs/\
        prod/cfs/cfs.20120514/00/6hrly_grib_01/\
        pgbf2012120100.01.2012051400.grb2 > pres.grb2
```

Sau khi thực thi lệnh trên, file pres.grb2 sẽ chứa dữ liệu của PRES trên mực "mean sea level".

Như vậy, khi biết được index của 1 file số liệu CFS (có thể download được từ cùng thư mục với số liệu CFS trên website nomads.ncep.noaa.gov với dung lượng rất nhỏ), và danh sách các mục dữ liệu cần sử dụng, ta có thể trích được đúng dữ liệu cần thiết từ trước khi download, từ đó giảm băng thông kết nối internet và dung lượng lưu trữ.

2.2. Tách và ghép

Trong phần trên, chúng tôi đã trình bày phương pháp tách dữ liệu trước khi download để tiết kiệm băng thông và dung lượng lưu trữ. Tuy nhiên, khi số lượng trường dữ liệu (tương ứng với số bản ghi GRIB) quá nhiều thì thời gian cần thiết để thiết lập kết nối (handshake) với server sẽ đóng một tỉ trọng lớn trong tổng thời gian download số liệu. Ví dụ, nếu download từng bản ghi trong tổng số 194 bản ghi cần thiết từ file pgbf, chương trình sẽ phải kết nối với server 194 lần. Với dung lượng trung bình của mỗi bản ghi chỉ vào khoảng vài chục KB thì thời gian handshake với server còn lớn hơn cả thời gian download dữ liệu thực. Có thể nhận thấy rằng, một số bản ghi có địa chỉ liền kề nhau trong file số liệu CFS gốc. Vì vậy, để giảm thời gian thiết lập kết nối, byte-range của các bản ghi này sẽ được kết hợp với nhau. Trở lại ví dụ trên, thay vì download từ byte 0-67690 và 67691-104277 cho "PRES mean sea level" và "HGT 1mb", ta sẽ download liền từ byte 0 đến hết byte104277.

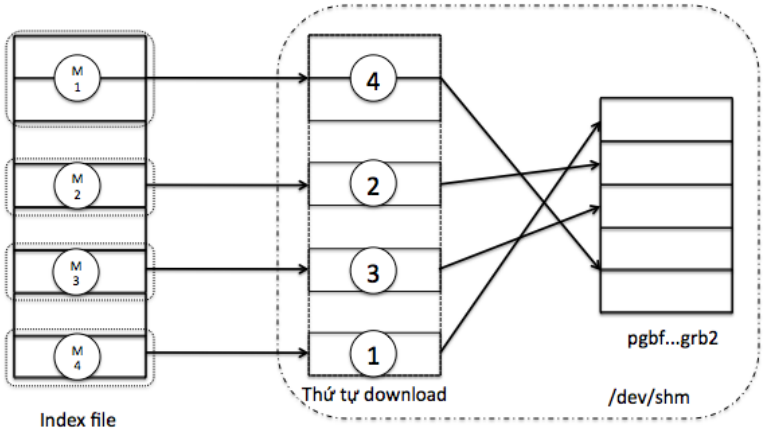
2.3. Tăng tốc download với giải thuật đa luồng và bộ đệm RAM

Sau ghi ghép các khoảng byte-range, số lần kết nối với server để download các bản ghi cho các file flxf, ipvf và pgbf sẽ giảm xuống còn từ vài lần đến vài chục lần. Nếu download tuần tự, số lượng kết nối này vẫn còn khá lớn và tuỳ thuộc vào chất lượng kết nối internet, thời gian download có thể không giảm được nhiều. Để tăng tốc download, nhiều byte-range có thể được download cùng một lúc, giống như phần mềm IDM hay Flashget vẫn làm trên Windows. Việc download một byte-range có thể được điều khiển bởi một luồng (thread).

Lựa chọn số lượng thread tối đa được phép chạy cùng lúc (Nthread) cũng rất quan trọng. Khi số lượng thread quá nhiều, kết nối internet có thể bị nghẽn khiến cho lưu lượng dữ liệu download giảm xuống. Thêm nữa, khi có quá nhiều thread chạy, tương ứng với số kết nối cùng lúc tới server tăng, tường lửa trên server có thể chặn tất cả kết nối từ địa chỉ IP của máy chạy chương trình download. Thực nghiệm trên đường kết nối internet của nhóm nghiên cứu (VNPT FTTH 40 Mbps) cho thấy Nthread=16 sẽ cho kết quả tối ưu. Để khống chế số lượng thread tối đa được chạy cùng một lúc, chúng tôi sử dụng 1 file tạm trong RAM (thread_num) chứa số thread hiện chạy. Chương trình điều khiển chính (cfs_dl) tạo ra thread_num và ban đầu nó ghi duy nhất một giá trị là Nthread. Khi cfs_dl cần khởi tạo 1 thread download, nó sẽ check xem giá trị trong thread_num có > 0 hay không, nếu đúng, thread sẽ được chạy và giá trị trong thread_num giảm đi 1. Khi 1 thread chạy xong, giá trị trong thread_num sẽ được tăng thêm 1. Bằng cách này, số lượng thread được chạy cùng 1 lúc sẽ không thể vượt quá Nthread.

Các tập bản ghi GRIB (mảnh dữ liệu GRIB) sau khi được download thì có thể được nối với nhau theo bất cứ thứ tự nào. Khi số lượng mảnh dữ liệu nhiều, đặc biệt với file pgbf, việc nối dữ liệu trên ổ cứng có thể mất nhiều thời gian. Do lượng dữ liệu của 1 file là tương đối nhỏ (lớn nhất chỉ khoảng 8.2MB), tất cả các mảnh dữ liệu có thể được download về và ghép lại thành file GRIB trong RAM trước khi được chuyển tới nơi lưu trữ lâu dài trên ổ cứng. Trên Linux, việc này rất đơn giản là thực hiện các thao tác trên trong /dev/shm.

Quá trình tách file index thành các byte-range (Mx), download mảnh dữ liệu và ghép thành file GRIB được minh hoạ trong [Hình 1]. Mảnh dữ liệu nào download về trước sẽ được nối vào trước trong file GRIB cuối cùng.



Hình 1. Tách index file, download các mảnh dữ liệu và ghép thành file GRIB

2.4. Thử nghiệm

Các chương trình get_inv và cfs_dl được đặt trong cron để tự động khởi chạy vào 0h UTC thứ 2 hàng tuần. Server đặt chạy chương trình có 2 CPU QuadCore 2.5GHz, 8 GB RAM. Nthread được đặt bằng 16. Kết nối internet là đường VNPT FTTH 40 Mbps. Thử nghiệm chạy thực tế cho thấy tốc độ download số liệu tăng khá nhanh. Nếu download toàn bộ file số liệu bằng wget/curl thì thời gian cần thiết để download các file flxf, ipvf là khoảng 10-15 giây và thời gian cần để download file pgbf là từ 30 đến 60 giây. Trong khi đó thời gian tương ứng để download bằng chương trình của chúng tôi là 3-5 giây và 10-15 giây. Tổng thời gian download toàn bộ trên 8000 file số liệu giảm tới hơn 1/2. Thêm nữa, trong suốt quá trình download số liệu, tài nguyên của server không bị chương trình chiếm dụng nhiều.

Không chỉ tăng tốc quá trình download, giải pháp của chúng tôi cũng giảm dung lượng lưu trữ dữ liệu tới 3 lần như đã phân tích ở mục

2.5. Hướng dẫn download tự động

Để chạy chương trình, người sử dụng cần thiết lập các tham số về đường dẫn tới nơi chứa các script, nơi chứa index file và nơi lưu trữ số liệu CFS trong file .ProductionConfig. Mặc định các thư mục này nằm trong thư mục home của người sử dụng. Crontab cần đặt theo minh hoạ sau để download số liệu vào thứ 2 hàng tuần:

```
# Download CFS every Monday

*/15 * * * 1 /home/cfs/scheduler/cfs_dl

*/5 * * * 1 /home/cfs/scheduler/get_inv
```

3. Kết luận

Để có thể mô phỏng được các trường khí tượng, các mô hình khí hậu toàn cầu và khu vực cần được cung cấp các điều kiện ban đầu và điều kiện biên. Các mô hình GCM và RCM đang được sử dụng tại nhóm nghiên cứu của chúng tôi lấy đầu vào là kết quả dự báo của mô hình CFS. Để dự báo khí hậu hạn mùa cho 6 tháng, dự báo cho khoảng thời gian tương ứng của CFS cũng cần được download về. Tổng số file cần download cho một lần dự báo là trên 8000 file với dung lượng lên tới hơn 100 GB, mặc dù số liệu này không cần dùng hết.

Trong chuyên đề này, chúng tôi trình bày giải pháp để tăng tốc quá trình download đồng thời giảm dung lượng cần dùng để lưu trữ số liệu. Các phương pháp sử dụng bao gồm bóc tách dữ liệu trước khi download, download nhiều luồng và dùng RAM làm bộ đệm.

Các kết quả thực nghiệm cho thấy giải pháp của chúng tôi đã tăng tới trên 2 lần tốc độ download và giảm tới 2/3 dung lượng ổ cứng cần sử dụng để lưu trữ số liệu. Phương pháp này hoàn toàn có thể ứng dụng để download các số liệu khí tượng khác để phục vụ cho các mục đích khác nhau, ví dụ download số liệu GFS để phục vụ bài toán dự báo thời tiết.