8.2 NORTH AMERICAN REGIONAL REANLAYSIS: END USER ACCESS TO LARGE DATA SETS

W. Ebisuzaki [1], J. Alpert [2], J. Wang [3], D. Jovic [3], P. Shafran [3]
[1]-NCEP/CPC, [2]-NCEP/EMC, [3]-SAIC,NCEP/EMC

1. Abstract

The North American Regional Reanalysis (NARR) is a reprocessing of the historical meteorological observations using NCEP's regional forecast model (ETA) and associated 3D-VAR data assimilation system (EDAS). The products of NARR will be a new set of meteorological analyses covering the North American domain with a 32 km horizontal resolution, 3 hour temporal resolution and 50 hPa vertical resolution for October 1978 and onwards. However, this much more detailed view the North American weather and climate produces an archive of approximately 80 TB.

We expect to be able to support a large majority of users' data requests with a 5 TB on-line data set. However, even a 5 TB subset can not be downloaded much less stored by most users. In order to reduce the data transport problem, we have adopted multiple approaches as a single method will not satisfy most users.

2. Introduction

NARR builds on NCEP's experiences in producing and distributing two global reanalyses (Kalnay et al, 1996; Kanamitsu et al, 2002). NCEP wanted to improve upon the earlier reanalyses by increasing the horizontal resolution by factor of 8 (32 km vs. 250 km) and by providing better estimates of the near surface variables and soil moisture. These factors should expand the number of applications of this reanalysis. More details about NARR are provided by Mesinger et al. (2004) and Shafran et al (2004).

Corresponding author's address:
Wesley Ebisuzaki, Climate Prediction Center,
W/NP51, Rm 811, WWBG,
5200 Auth Rd,
Camp Springs, MD, USA 20746;
email: Wesley.Ebisuzaki@noaa.gov

3. NARR Output

The complete NARR archive is approximately 80 TB and includes,

- input observations
- input observations with QC marks and differences from analyses and 1st guess
- input analyses, ex. sea surface temperature, snow depth, sea ice
- plots of observations locations, QC
- plots of fits of analyses to observations
- plots of fits to global reanalysis
- plots of analyses
- 3 day forecasts every 2 ½ days
- analyses and 3 hour forecasts (1st guess) in 3 different formats (model restart, GRIB format on model grid, GRIB format on Lambert conformal grid)

The bulk of the 80 TB is taken by the forecasts and the 3 different variations of the analyses and 1st guess fields. Experience from earlier reanalyses suggests that a majority of users only want the analyses and flux quantities (e.g., precipitation, latent heat, OLR) from the 1st guess. Demand for the forecasts is expected to be small. See the Appendix for the analysis and flux variables..

The analyses and fluxes are available in 3 different formats. The formats differ by grid type (model and Lambert conformal) and by format (binary restart, GRIB). The model grid is unsupported by many visualization programs as the wind and mass points are staggered. The binary restart file has a non-standard format and is much larger than the GRIB files because it contains much information for restarting the model. Consequently the Lambert conformal (AWIPS) GRIB data would be the best suited for most of the users. A 'merged' data set based on the analyses+fluxes on the AWIPS grid is approximately 5 TB (60 MB every 3 hours). We expect that this merged dataset (Appendix) along with some of the smaller data sets will satisfy most users.

4. Coping with Large Data Sets

One can distribute large data sets using physical media (tapes, DVDs, CDs). Tapes can be used to transfer large amounts of data but can be expensive for both the sender and receiver of the tapes. We are limiting our use of tapes to transferring data to two archive sites (National Center for Atmospheric Research, and San Diego Supercomputing Center at UCSD). DVDs and CDs have insufficient storage except for small specialized subsets. Since we expect a diverse set of users, it would be difficult to design a subset that would satisfy a sizable fraction of users.

We expect to satisfy most users through on-line servers. Distributing data from large data sets over the Internet is not a problem unique to the reanalyses. To tackle the bandwidth problem, we have to look at how people use the data.

Specific fields, ex. 10 m winds
Regions, ex Washington state
Time series
Calculated quantities such as 5-day mean
Plots (as compared with raw data)

Each of the previously listed uses allows a reduction in data volume. For example, a time series may consist of 20 numbers extracted from 20 58-MB files. A request for the 2-m temperature rather than the entire set of variables reduces the data need from 58 MB to 143 KB. In short, each of these uses allows a reduction in data size. So we have decided that multiple approaches are needed.

Another factor to remember is that some data needs may be small while others may take weeks to process (experience from global reanalysis). The system has to be usable for both small and large downloads. In addition, users will have varying amounts of experience dealing with meteorological data formats such as GRIB.

4.1 GRIB data

The gridded analyses from NARR are in GRIB format. For the GRIB data, we allow the users to select the desired: fields, levels, times, region, and resolution. We have filters that work on the GRIB files to reduce the data volumes based on the selected options. The subsetting utility is written as a cgi-bin perl script (ftp2u) which is accessed over the web. With a browser, the user clicks on the desired files, desired filtering options and then selects the deposition of the file. The output can be saved on either the server's or another (anonymous) ftp server.

The ftp2u script consists of two parts.

1. Create a page listing user options such as
   files
   fields/levels/times/regions/resolution
2. Read user selections
   create the filter parameters
   filter (subset) the dataset
   send the data to local or remote ftp site

The cgi-bin script is not huge, as it was originally written as a shell script and later ported into perl as the capabilities of the program were expanded. Much of the hard work is done by the filters such as wgrib and ggrib which were coded in C for portability.

The casual user can access the ftp2u utility using a browser. However, some users need to download large amounts of data or download data on a schedule (i.e., forecasts, real-time analyses). In these cases, the point and click of the browser needs to be replaced by a programmable interface. This can be done easily. For example, the ftp2u is a cgi-bin script that doesn't use cookies and uses the POST method of accessing the arguments. Consequently the URL is state-less and all the arguments appear on the URL. For example, suppose I want to down-load the 10 m winds for an analyses for a subdomain of 20N-50N by 90W-60W. After selecting the form and clicking on 'START FTP', I get the page showing the status of my download. The URL for the page is,

http://wesley.wwb.noaa.gov/cgi-bin/ftp2u_rr.sh?file=AWIP3200.1995020100.merged&wildcard=&lev_10_m_above_gnd=on&var_UGRD=on&var_VGRD=on&subregion=&leftlon=270&rightlon=300&toplat=50&bottomlat=20&machine=wesley.wwb.noaa.gov&user=anonymous&passwd=sample&ftpdir=%2Fincoming_1hr%2Fwesley&prefix=&dir=%2Fanalyses%2Fmerged_AWIP32%2F199502

The URL looks complicated but if you look closely, you see it is just an address and a list of arguments separated by ampersands and special characters have been replaced by

their hexadecimal equivalent. Suppose I wanted to the next day's winds, I could enter the above URL into the browser with the slight change from

file=AWIP3200.1995020100
to
file=AWIP3200.1995020200

When I load the new URL, ftp2u would download the next day's winds. Of course, typing URLs into browser windows is tedious, but that same 'downloading' of the URL can be automated in shell scripts with programs such as wget and wwwgrab. In addition, one can use library functions in languages such as perl, java and Vbasic. So if a cgi-bin script follows a few simple rules, it has a programming interface that is easy to use. My experience with other data sets is that most of the downloads are initiated using the programming interface.

4.2 Plots and Time Series

Sending plots can reduce the data transfer significantly. A plot (png) file can be 50 KB whereas the original data file may be 50 MB. In addition, a plot of a time series may be 20 KB which involved a processing large amounts of data.

The plotting program (pdisp) is similar to the ftp2u program. It is a cgi-bin script that has two parts.

1. Create pages listing user options such as
   - file
   - type of plot, variable(s)
   - simple calculations (ex. averaging)
2. Parse options
   - create and execute GrADS script
   - make png file
   - for time series, the GrADS script makes the text data file
   - make html page for viewing

At its simplest, you point pdisp at a directory and all the GrADS compatible data sets in that directory are on the web. Here, pdisp is simply a GUI over GrADS. However, pdisp includes a simple language that allows one to exploit the capabilities of GrADS. For example, one can open secondary files and create anomalies or do comparisons. These mini-programs are embedded in the GrADS control file and give the owner of the control file much control over the production of the plot. In addition, the pdisp program allows extraction of time series for reduction in the data transfer.

The pdisp script has a similar programming interface to the ftp2u script. Of course, the variables differ but accessing the plot engine is accomplished in the same manner. The pdisp program is used for distributing the GFS (NCEP's Global Forecast System) forecasts, and most requests are generated with the programming interface.

4.3 GrADS DODS Server (GDS)

A DODS server allows user clients such as GrADS, Ferret, IDL to obtain data from a server's database. For example, one could open a weather forecast on a DODS server and plot the precipitation forecast using your own program. The DODS server would send the precipitation values and your local program would generate the plot. Since only the necessary values are sent from the server, the data transfer is reduced.

The GrADS DODS Server's advantage over generic DODS servers is its compatibility with GRIB and the other formats supported by GrADS. GDS also allows server-side calculations which, if used, could reduce the data transfer. However, server-side analysis could take a long time to evaluate when applied to large datasets. This introduces the problem that some people could slow the server by doing long calculations. The current GDS implementation allows a set CPU limit to prevent overuse by individuals.

5. Summary

The products from the North American Regional Reanalysis will be multiple terrabytes in size. We want the data easily accessible and we plan to do it using web services that minimize the data transferred over the network. We also showed how easy it is to design an interactive http access to the data that also serves as a programming interface for users that need to download larger amounts of data or on a schedule.

6. References

Kalnay, E., et al., 1996: The NCEP/NCAR 40 Year Reanalysis Project, Bull. Amer. Meteor. Soc., **77**, 437-471.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S-K. Yang, J. J. Hnilo, M. Fiorino, and G.L. Potter, 2002: NCEP-DOE AMIP-II Reanalysis (R-2), Bull. Amer. Meteor. Soc., **83**, 1631-1643.

Mesinger, F., et al, 2004: NCEP North American Regional Reanalysis, 15th Symp. On Global Change and Climate Variations, Seattle, WA, 11-15 Jan 2004.

Shafran, P., J. Woollen, W. Ebisuzaki, W. Shi, Y. Fan, R. W. Grumbine, M. Fennessy, 2004: Observational Data Used for Assimilation in the NCEP North American Regional Reanalysis, 20th Intl. Conf. On Interactive Information Processing Systems for Meteor. Ocean. And Hydrology. Seattle, WA, 11-15 Jan 2004.

7. Appendix: Contents of the Merged DataSet

4LFTX:180-0 mb above gnd:anl:Best (4-layer) lifted index [K]
ACPCP:sfc:0-3hr acc:Convective precipitation [kg/m^2]
ALBDO:sfc:anl:Albedo [%]
APCP:sfc:0-3hr acc:Total precipitation [kg/m^2]
APCPN:sfc:0-3hr acc:Total precipitation (nearest grid point) [kg/m^2]
BGRUN:sfc:0-3hr acc:Subsurface runoff (baseflow) [kg/m^2]
BMIXL:hybrid lev 1:anl:Blackadars mixing length scale [m]
CAPE:180-0 mb above gnd:anl:Convective available potential energy [J/kg]
CAPE:sfc:anl:Convective available potential energy [J/kg]
CCOND:sfc:anl:Canopy conductance [m/s]
CD:sfc:anl:Surface drag coefficient [non-dim]
CDCON:atmos col:0-3hr ave:Convective cloud cover [%]
CDLYR:atmos col:0-3hr ave:Non-convective cloud [%]
CFRZR:sfc:3hr fcst:Categorical freezing rain [yes=1;no=0]
CICEP:sfc:3hr fcst:Categorical ice pellets [yes=1;no=0]
CIN:180-0 mb above gnd:anl:Convective inhibition [J/kg]
CIN:sfc:anl:Convective inhibition [J/kg]
CLWMR:P-STACK:Cloud water [kg/kg]
CNWAT:sfc:anl:Plant canopy surface water [kg/m^2]
CRAIN:sfc:3hr fcst:Categorical rain [yes=1;no=0]
CSNOW:sfc:3hr fcst:Categorical snow [yes=1;no=0]
DLWRF:sfc:0-3hr ave:Downward longwave radiation flux [W/m^2]
DPT:2 m above gnd:anl:Dew point temp. [K]
DSWRF:sfc:0-3hr ave:Downward shortwave radiation flux [W/m^2]
EVP:sfc:0-3hr acc:Evaporation [kg/m^2]
FRICV:sfc:anl:Surface friction velocity [m/s]
GFLUX:sfc:0-3hr ave:Ground Heat Flux [W/m^2]
HCDC:high cld lay:3hr fcst:High level cloud cover [%]
HGT:P-STACK:anl:Geopotential height [gpm]
HGT:0C isotherm:anl:Geopotential height [gpm]
HGT:cld base:anl:Geopotential height [gpm]
HGT:cld top:anl:Geopotential height [gpm]
HGT:hybrid lev 1:anl:Geopotential height [gpm]
HGT:max wind lev:anl:Geopotential height [gpm]
HGT:tropopause:anl:Geopotential height [gpm]
HLCY:3000-0 m above gnd:anl:Storm relative helicity [m^2/s^2]
HPBL:sfc:anl:Planetary boundary layer height [m]
ICMR:P-STACK:anl:Ice mixing ratio [kg/kg]
LCDC:low cld lay:3hr fcst:Low level cloud cover [%]
LFTX:500-1000 mb:anl:Surface lifted index [K]
LHTFL:sfc:0-3hr ave:Latent heat flux [W/m^2]
MCDC:mid cld lay:3hr fcst:Mid level cloud cover [%]
MCONV:850 mb:anl:Horizontal moisture divergence [kg/kg/s]
MCONV:BL-STACK:anl:Horizontal moisture divergence [kg/kg/s]
MSLET:MSL:anl:Mean sea level pressure (ETA model) [Pa]
MSTAV:0-100 cm down:anl:Moisture availability [%]
PEVAP:sfc:0-3hr acc:Potential evaporation [kg/m^2]

POT:M-STACK:anl:Potential temp. [K]
POT:hybrid lev 1:anl:Potential temp. [K]
POT:sfc:anl:Potential temp. [K]
PRATE:sfc:3hr fcst:Precipitation rate [kg/m^2/s]
PRES:M-STACK:anl:Pressure [Pa]
PRES:cld base:anl:Pressure [Pa]
PRES:cld top:anl:Pressure [Pa]
PRES:cond lev:anl:Pressure [Pa]
PRES:hybrid lev 1:anl:Pressure [Pa]
PRES:max wind lev:anl:Pressure [Pa]
PRES:sfc:anl:Pressure [Pa]
PRES:tropopause:anl:Pressure [Pa]
PRESN:sfc:anl:Pressure (nearest grid point) [Pa]
PRMSL:MSL:anl:Pressure reduced to MSL [Pa]
PWAT:atmos col:anl:Precipitable water [kg/m^2]
RCQ:sfc:anl:Humidity parameter in canopy conductance [fraction]
RCS:sfc:anl:Solar parameter in canopy conductance [fraction]
RCSOL:sfc:anl:Soil moisture parameter in canopy conductance [fraction]
RCT:sfc:anl:Temperature parameter in canopy conductance [fraction]
RH:2 m above gnd:anl:Relative humidity [%]
RH:0C isotherm:anl:Relative humidity [%]
RH:hybrid lev 1:anl:Relative humidity [%]
SFEXC:sfc:anl:Exchange coefficient [(kg/m^3)(m/s)]
SHTFL:sfc:0-3hr ave:Sensible heat flux [W/m^2]
SNOD:sfc:anl:Snow depth [m]
SNOHF:sfc:0-3hr ave:Snow phase-change heat flux [W/m^2]
SNOM:sfc:0-3hr acc:Snow melt [kg/m^2]
SNOWC:sfc:anl:Snow cover [%]
SOILL:SOIL-LAYERS:anl:Liquid volumetric soil moisture (non-frozen) [fraction]
SOILM:0-200 cm down:anl:Soil moisture content [kg/m^2]
SOILW:SOIL-LAYERS:anl:Volumetric soil moisture (frozen + liquid) [fraction]
SPFH:P-STACK:anl:Specific humidity [kg/kg]
SPFH:BL-STACK:anl:Specific humidity [kg/kg]
SPFH:M-STACK:anl:Specific humidity [kg/kg]
SSRUN:sfc:0-3hr acc:Surface runoff (non-infiltrating) [kg/m^2]
TCDC:atmos col:3hr fcst:Total cloud cover [%]
TKE:P-STACK to 600 mb:anl:Turbulent Kinetic Energy [J/kg]
TKE:hybrid lev 1:anl:Turbulent Kinetic Energy [J/kg]
TMP:P-STACK:anl:Temp. [K]
TMP:BL-STACK:anl:Temp. [K]
TMP:M-STACK:anl:Temp. [K]
TMP:cld top:anl:Temp. [K]
TMP:sfc:anl:Temp. [K]
TMP:tropopause:anl:Temp. [K]
TSOIL:800 cm down:anl:Soil temp. [K]
TSOIL:SOIL-LAYERS:anl:Soil temp. [K]
UFLX,VFLX:sfc:anl:Zonal momentum flux [N/m^2]
UGRD,VGRD:P-STACK:anl:u wind [m/s]
UGRD,VGRD:BL-STACK:anl:u wind [m/s]
UGRD,VGRD:M-STACK:anl:u wind [m/s]
UGRD,VGRD:max wind lev:anl:u wind [m/s]
UGRD,VGRD:tropopause:anl:u wind [m/s]
ULWRF:nom. top:0-3hr ave:Upward long wave radiation flux [W/m^2]
ULWRF:sfc:0-3hr ave:Upward long wave radiation flux [W/m^2]
USTM,VSTM:6000-0 m above gnd:anl:u-component of storm motion [m/s]
USWRF:nom. top:0-3hr ave:Upward short wave radiation flux [W/m^2]
USWRF:sfc:0-3hr ave:Upward short wave radiation flux [W/m^2]
VEG:sfc:anl:Vegetation [%]
VIS:sfc:anl:Visibility [m]
VVEL:P-STACK:anl:Pressure vertical velocity [Pa/s]
VVEL:BL-STACK:anl:Pressure vertical velocity [Pa/s]

VWSH:tropopause:anl:Vertical speed shear [1/s]
WCCONV:2-LAYER:0-3hr acc:Water condensate flux convergence (vertical int) [kg/m^2/s]
WCINC:2-LAYER:0-3hr acc:water condensate added by precip assimilaition [kg/m^2/s]
WCUFLX:2-LAYER:0-3hr acc:Water condensate zonal flux (vertical int) [kg/m/s]
WCVFLX:2-LAYER:0-3hr acc:Water condensate meridional flux (vertical int) [kg/m/s]
WEASD:sfc:anl:Accum. snow [kg/m^2]
WVCONV:2-LAYER:0-3hr acc:Water vapor flux convergence (vertical int) [kg/m^2/s]
WVIWVUFLX:2-LAYER:0-3hr acc:Water vapor zonal flux (vertical int)[kg/m/s]
WVVFLX:2-LAYER:0-3hr acc:Water vapor meridional flux (vertical int) [kg/m/s]

The following is in a separate file (increment)

PWAT:atmos col:3hr fcst:Precipitable water [kg/m^2]
WEASD:sfc:3hr fcst:Accum. snow [kg/m^2]

P-STACK=25 mb from 1000-700, 50 mb from 700-300, 25 mb from 300-100
BL-STACK=0-30, 30-60, .. 150-180 mb above ground, hybrid level 1
M-STACK: 2, 10 and 30 m above ground except for UGRD, VGRD, POT which don't have 2 m
2-LAYER:0-700 mb, 0-Top of Atmosphere
SOIL-LAYERS=4 soil layers, 0-10, 10-40, 40-100, 100-200 cm below groundNC:2-LAYER:0-3hr acc:water vapor added by precip assimilation [kg/m^2/s]