# Evaluation of National Centers for Environmental Prediction Global Forecast System at the Atmospheric Radiation Measurement Program Southern Great Plains Site

*F. Yang, H-L. Pan, S. Moorthi, and S. Lord*
*Environmental Modeling Center, National Centers for Environmental Prediction*
*Camp Springs, Maryland*

*S. Krueger*
*Department of Meteorology*
*University of Utah*
*Salt Lake City, Utah*

## Abstract

Since 2001 output from the National Centers for Environmental Prediction global weather forecast system (GFS) has been routinely processed to produce single column profiles at locations corresponding to the Atmospheric Radiation Measurement (ARM) Program sites. In the present study, GFS forecast was examined and compared with ARM observations at the Southern Great Plain Central Facility for the years from 2001 through 2004. The comparison was made for the first 48 hours of forecast and primarily for surface energy fluxes and clouds. For most of the variables examined, the forecast compared well with the observations. The performance of the model has been improving along with model upgrades. For certain variables, persistent biases exist. For instance, the model overestimated latent heat flux by 80 $W/m^2$ over the 1-3 PM time period of the day and, coincidently, overestimated surface downward solar flux by 43 $W/m^2$, and underestimated sensible heat flux by 48 $W/m^2$. As a result, the surface net heat budget in the forecast was balanced due to cancellation of errors. For cloud, the model was able to capture its observed temporal evolution and vertical distribution during major synoptic events. However, on average, the model largely underestimated cloud fraction in the lower and middle troposphere, and slightly overestimated in the upper troposphere. The forecast missed daytime non-precipitating low clouds seen in the observations. Diurnal cycle of clouds in the middle to lower troposphere was much weaker in the forecast than in the observation. The bias in cloud diurnal cycle was found to be responsible for a phase shift in the diurnal cycle of the forecast cloudy-sky surface downward longwave flux (SDLW). Further assessment was made for time before and after an upgrade of the GFS longwave radiative transfer module. After the upgrade, a systematic low bias of cloudy-sky SDLW found in the earlier forecast was eliminated; however, clear-sky SDLW at night became too large. The latter was probably caused by a nighttime warm bias in the lower and middle atmosphere. This investigation demonstrated the unique value of ARM observations in diagnosing and understanding forecast model errors, and for improving the forecast model physics.

# Introduction

The U.S. Department of Energy established the ARM Program in the early 1990's to help resolve uncertainties related to the role of clouds and their influence on radiative feedback processes in the atmosphere (Stokes and Schwartz 1994; Ackerman and Stokes 2003). The ultimate goal of this program is to improve parameterizations of cloud and radiation in global atmospheric general circulation models (GCM), and hence to enhance the capability of these GCMs in simulating present climate and projecting future climate changes. Intensive and long-term measurements of surface and atmospheric quantities have been carried out at surface sites in the U.S. Southern Great Plains (SGP), northern slope of Alaska, and in the tropical western Pacific, representing different climate regimes. Many investigators have used these measurements for evaluating, testing and improving numerical models at different scales, ranging from GCMs to cloud-resolving models (e.g., Randall and Cripe 1999; Xie and Zhang 2000; Xu et al. 2002).

Even though the initial motive of the ARM program was to improve the performance of climate models, in recent years a few investigators have successfully applied ARM observations to the evaluation of numerical weather prediction (NWP) models at operational weather forecast centers (e.g., Mace et al. 1998; Hinkelman et al. 1999; Morcrette 2002). ARM provides certain products that are unique, not available from observations either at conventional meteorological stations or by satellites. Mace et al. (1998) compared the precipitation and vertical distribution of clouds in the European Center for Medium-Range Weather Forecast (ECMWF) forecast in winter 1997 with data from a millimeter-wave radar operated at the ARM SGP site, and found the forecast model tended to predict the onset of deep cloud events too soon. Morcrette (2002) assessed the cloud and radiation fields in the first 36 hours of the operational ECMWF forecast in April-May 1999 using ARM observations at the SGP site. By taking advantage of the comprehensive ARM observations at high temporal and vertical resolution the author was able to identify an overestimate of surface downward solar flux in the forecast and to attribute the source of the bias to insufficient clear-sky gaseous absorptions. Hinkelman et al. (1999) validated the National Centers for Environmental Prediction (NCEP) regional Eta forecast model against ARM observations at the SGP site using observations in the first half of 1997. They also found an overestimate of surface downward solar flux up to 50 $W/m^2$ in the Eta forecast and attributed half of the bias to insufficient extinction of shortwave radiation by water vapor and aerosols. These studies demonstrate that, in addition to their usage in climate models, ARM observations can be as well used for ascertaining the quality of operational weather forecast and for improving forecast model physical parameterizations.

In the past, most GCM and NWP developers resorted to different strategies for developing and evaluating parameterizations of model physical processes. When testing a new parameterization, GCM developers usually relied on multi-year climate simulations and the comparison of model mean statistics with observations. The success of this strategy was sometimes hindered by limitations like insufficient observations, cancellation of simulation errors, and strong feedbacks between different physical processes. At NWP centers, developers relied more on case studies of short to media-range weather forecasts of extreme events for evaluating parameterization schemes. In recent years, a new approach for evaluating GCM parameterization has gained some momentum (e.g., Jakob 2003; Phillips et al. 2004). In this approach, GCMs are initialized with NWP global reanalyses, run in a short-range NWP forecast mode for one or a few years to generate large samples covering different weather and climate

regimes, and then assessed against satellite and ground observations. This approach was built upon the promise that large-scale dynamical state of the GCM remains close to the observed at the early-stage of the forecast mode; therefore, systematic forecast biases can be attributed predominantly to parameterization deficiencies (Phillips et al. 2004).

In the present study, we evaluate the NCEP global weather forecast in the past few years at the ARM SGP site. We take an approach that emphasizes more on the statistical behavior of the forecast model in all the years rather than on the performance of the model during extreme synoptic events. The purpose is to identify systematic forecast biases and to provide recommendations for further model improvement. Since early 2001, NCEP has been processing the operational NWP global forecast to produce column profiles at locations corresponding to all ARM permanent sites. Up to now, the archived data have not been used for model evaluation and development. At NCEP, Hinkelman et al. (1999) ever used a few months of ARM observations and evaluated the regional Eta forecast model. The NCEP global forecast model consists of different physical and dynamical packages from the Eta model. This study makes use the comprehensive and continuous ARM observations to assess the NCEP operational forecast at the ARM SGP site in the years from 2001 through 2004. The focus is primarily on clouds and surface energy fluxes.

In Section 2, we introduce first the current NWP model, describe in certain detail the model's physical parameterizations relevant to this study; and then present the ARM observations and NCEP forecasts used. A strategy for comparing modeled and observed quantities at different temporal and spatial scales is also briefly discussed. Evaluation of surface energy fluxes and surface-air temperature are given in Section 3, and the evaluation of clouds in Section 4. Section 5 investigates further the interaction between clouds and radiation, and how the interaction contributed to the forecast errors identified in Sections 3 and 4. Section 6 summarizes the study.

# Model, Observations, and Evaluation Strategy

## Description of the NCEP GFS

The NCEP GFS is a spectral numerical model with primitive dynamical equations and consists of a suite of complex and advanced parameterizations for atmospheric physics (e.g., Sela 1980; Kanamitsu 1989; Kalnay et al. 1990). The model has been under constant development and evaluation. An online documentation of the model is given at http://wwwt.emc.ncep.noaa.gov/gmb/moorthi/gam.html. Here we describe some of the model features and the major changes made to the model since 2001 that are relevant to the present investigation.

The GFS uses spectral triangular truncation in the horizontal and sigma coordinate in the vertical, extending from the earth's surface to the top of the atmosphere. The model was upgraded on 29 October 2002 from T170L42 to T254L64, roughly from a horizontal grid size of 75 km to 55 km. Time step is 7.5 minutes in T170L42 and 5 minutes in T254L64 for the computation of dynamics and physics, except that the full calculation of longwave radiation is done once per 3 hours and shortwave radiation once per an hour. Corrections are made at every time step to adjust for diurnal variations in shortwave fluxes through the atmosphere and in upward longwave flux at the surface.

For shortwave radiation, the parameterization is based on Chou and Suarez (1999) and modified by Hou et al. (2002) for the NCEP models. It contains eight broad spectral bands in the ultraviolet and visible region and one spectral band in the near infrared region, and resolves absorbing gases of ozone, water vapor, carbon dioxide, and oxygen. A delta-Eddington approximation method is used for multi-scattering calculation. Random cloud overlapping is assumed in the operational GFS. Cloud optical properties depend on liquid/ice water path and cloud effective radius (Slingo 1989; Fu 1996). Cloud optical thickness is calculated from the predicted cloud condensate path. Cloud single-scattering albedo and asymmetry factor are functions of effective radius of the cloud condensate. The effective radius for ice is taken as a linear function of temperature decreasing from a value of 80 microns at 263.16 K to 20 microns at or below 223.16 K. For water droplets with temperatures above 273.16 K, an effective radius of 5 microns is used. For super-cooled water droplets between the melting point and 253.16 K, a value between 5 and 10 microns is used. Effects from rain drops and snow are not included in the operational GFS. Atmospheric aerosol effect is included. Horizontal distribution of surface albedo is a function surface vegetation type (Matthews 1985) in a manner similar to Briegleb et al. (1986). Monthly variation of surface albedo is derived in reference to Staylor and Wilbur (1990).

For longwave radiation, a major change was made on August 28, 2003. The Chou and Suarez (1994) scheme was replace by the Rapid Radiative Transfer Model (RRTM) (Mlawer et al. 1997). The Chou and Suarez scheme computes longwave absorption and emission by water vapor, carbon dioxide, ozone, clouds, aerosols and trace gases ($N_2O$, $CH_4$, CFC–11, CFC–12, and HCFC–22), and contains 8 broad bands. Three different approaches are used to compute the transmission functions for different gaseous absorbers: (1) the k–distribution method with linear pressure and/or temperature scaling, (2) two–parameter scaling with precomputed look–up tables, and (3) one–parameter scaling. For RRTM, it computes longwave absorption and emission by water vapor, carbon dioxide, ozone, clouds and various trace gases including $N_2O$, $CH_4$, $O_2$, and four types of halocarbons (CFCs). Aerosol effect is not included. It uses a correlated-k distribution method and a transmittance look-up table that is linearly scaled by optical depth to achieve high accuracy and efficiency. The algorithm contains 140 unevenly distributed intervals in 16 broad spectral bands. RRTM employs an advanced CKD_2.4 scheme (Clough et al. 1992) to compute absorption by water vapor at the continuum bands. A maximum-random cloud overlapping method is used. Cloud radiative properties external to the RRTM depend on cloud liquid/ice water path and the effective radius of ice particles and water droplets (Hu and Stamnes 1993; Ebert and Curry 1992).

Penetrative convection is simulated by the Simplified Arakawa-Schubert scheme (SAS; Pan and Wu 1994; Arakawa and Schubert 1974; Grell 1993). SAS considers only the deepest cloud instead of a spectrum of clouds as in the original Arakawa-Schubert scheme. Convection occurs when cloud work function exceeds a certain threshold. Cloud mass flux is determined using a quasi-equilibrium assumption based on this threshold. The current cloud model also incorporates a downdraft mechanism as well as the evaporation of precipitation. Entrainment of updraft and detrainment of downdraft in sub-cloud layers are included. Water detrained at cloud top is separated into condensate and vapor, with the former being treated as a source of prognostic cloud condensate. In addition, we have made a change in the cloud top selection algorithm, that is, the top is determined by the level where cloud parcel becomes stable with respect to the environment.

The parameterization of shallow convection follows Tiedtke (1983)'s diffusion scheme. Shallow convection occurs when convective instability exists without deep convection. Cloud base is determined by lifting condensation level. Vertical diffusion is invoked between cloud top and cloud bottom with a fixed profile of vertical diffusion coefficients.

Large-scale condensation is based on the Zhao and Carr (1997)'s prognostic scheme. Cloud condensate is produced by convective detrainment and grid-scale condensation, and diminished by evaporation and grid-scale precipitation, which is parameterized following Zhao and Carr (1997) for ice and Sundqvist et al. (1989) for liquid water. Evaporation of rain in unsaturated layers below the level of condensation is allowed. Grid-scale cloud fraction is diagnosed following Xu and Randall (1996).

For surface processes, the Monin-Obukhov similarity-profile relationship is used to compute surface wind stress and sensible and latent heat fluxes. The formulation was based on Miyakoda and Sirutis (1986) and has been modified for very stable and very unstable situations. Land surface evaporation includes direct evaporation from soil and canopy, and transpiration from vegetation (Pan and Mahrt, 1987). A two-layer soil model computes soil temperature and soil volumetric water content at depths of 0.1 and 1.0 meters (Pan and Mahrt, 1987). Deep-soil temperature at 4 meters down is specified. Heat capacity, thermal and hydraulic diffusivity and hydraulic conductivity coefficients are strong functions of soil moisture content. Vegetation canopy is allowed to intercept precipitation and re-evaporate. Runoff from the surface and drainage from the bottom layer are also calculated.

## GFS model output for ARM sites

Since late 2000, the NCEP operational global weather forecast has been processed to produce single column profiles at locations corresponding to the ARM ground observation sites, including the SGP Central Facility at Lamont, Okalahoma, the North Slope of Alaska Barrow and Atqasuk Facilities, and the Tropical Western Pacific Manus Island, Nauru Island and Darwin Facilities. All standard output of the global model were extracted and archived, including atmospheric profiles and surface variables, at a three-hour interval for up to 48 hours of model forecast. The twice-daily forecast was initialized at 00Z and 12Z, respectively. This paper includes results based on the 00Z forecast. Investigation of the 12Z forecast led to the same conclusion.

Fluxes at the earth's surface and top of the atmosphere, including radiation, latent and sensible heat, and wind stresses were saved as three-hour averages. Rainfall and snow were three-hour accumulations. State variables such as surface and atmospheric temperatures, pressure, cloud fraction, cloud condensates, water vapor mixing ratio were saved as instantaneous values at the end of each three-hour interval for forecasts starting from 00Z and ending at 48Z. Since atmospheric profiles were produced on the model sigma layers and the model vertical resolution has been changing, all vertical profiles have been projected onto constant isobaric layers (1000 hPa to 25 hPa, at a 25 hPa interval) and atmospheric heights (surface to 20 kilometer, at a 250 meter interval). The data reprocessed for all ARM sites have been archived at NCEP for use by investigators upon request.

## ARM observations at the Southern Great Plains

We list in Table 1 the various ARM data product for the SGP site we obtained from the ARM Archive (http://www.arm.gov) and used for this study. Although the comparison between GFS forecast and ARM observations was made primarily over the Central Facility (CF-1), observations at the SGP extended facilities (EF) were also used for a scale-dependence test (see Section 2.4). Figure 1 depicts the ARM SGP observation network. The thirteenth extended facility (EF-13) is co-located with the CF. Measurements from both EF-13 and CF-1 were used to represent observations at the CF site for model validation.

Observations we obtained from the ARM Archive are of different resolution in time. Since the GFS forecast was saved at a 3-hour interval, all ARM observations were processed to match the frequency and definition of the GFS output. For flux variables such as radiation, latent and sensible heat all measurements in a 3-hour period were used to derive 3-hourly mean fluxes. For state variables such as atmospheric temperature and humidity, measurements made at a time closest to the end of the 3-hour period were then selected. Most of the ARM data have gone through extensive quality control and are reliable for use. We inspected all datasets to exclude bad and missing values that occurred for instrument malfunctioning (see Section 3).

Table 1. ARM data streams and variables used for forecast validation, where Exx refers to the ARM Extended Facility site number.

| Data Stream and Identifier in ARM Archive | Number of Sites | Measurements Used for this Study | Frequency |
|---|---|---|---|
| Surface Meteorological Observation System Instruments (SMOS), sgp30smosExx.b1 | 16 EF | 2-meter surface temperature and humidity; precipitation | 30 minutes |
| Energy Balance Bowen Ratio (EBBR), sgp30ebbrExx.b1 | 14 EF | Sensible and latent heat fluxes; ground soil heat flux | 30 minutes |
| Best-Estimate Radiative Flux, sgpbeflux1longC1.c1 | CF-1 | Surface downward and upward broadband solar and longwave radiative fluxes | 1 minute |
| Solar Infrared Radiation Station (SIRS), sgpsirsExx.b1 | 22 EF and CF-1 | Surface downward broadband solar and longwave radiative fluxes | 1 minute |
| Active Remotely-Sensed Clouds Locations, sgparsclbnd1clothC1.c1 | CF-1 | Cloud base and top heights from laser ceilometer, microwave radiometer, and micropulse lidar observations | 10 seconds |
| "MWR PROF" Value Added Product, sgpmwrprofC1.c1 | CF-1 | Vertical profiles of water vapor, temperature, and cloud liquid water | 1 hour |

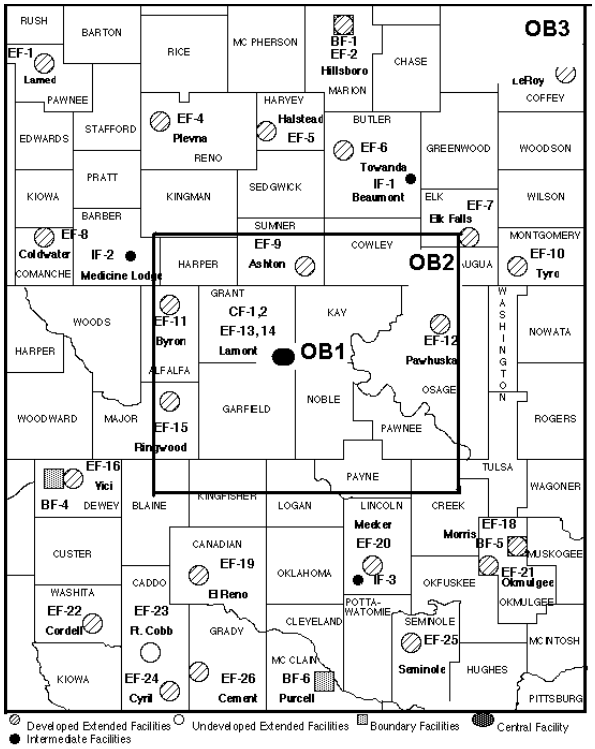


**Figure 1**. Facilities at the ARM SGP site (source http://www.arm.gov/). The entire domain is about 300x300 km in size (OB3). For the scale-dependence test described in Section 2.4, a sub-domain (OB2) comparable in size to the grid of the NCEP T254 (~ 55 km) forecast model is chosen for comparing the mean over the OB2 with the single-point observation at the central facility (OB1), and with the mean over the entire SGP site (OB3).

## Evaluation strategy and a scale-dependence test

In this study, we assess the GFS forecast skill by making use of the rich ARM observations in the four years from 2001 through 2004. We compare time series, diurnal cycles, seasonal and annual means of the GFS forecast with available ARM observations to detect model biases, to find if the model performance has been improving along with model upgrades, and to give recommendations for correcting model bias and developing better parameterizations.

The GFS output represents mean atmospheric conditions over a model grid, roughly 70 x 70 $km^2$ in size for the T170 version and 55 x 55 $km^2$ for the T254 version. However, ARM observations are made at single points. It is a question how to properly compare model grid values with single-point observations. Ideally, model output should be compared to the mean of observations in an area comparable to the model grid size; however, not all observations are available at all sites over the SGP site (Figure 1). Is it meaningful to compare model grid values with observations made at the single CF site? To answer this question we perform a scale-dependence test of the observations.

Figure 1 shows the ARM SGP Cloud and Radiation Testbed (CART) observation network. It covers an area of approximately 300 x 300 $km^2$ and includes three types of facilities, i.e., the central, extended and

boundary facilities. The instruments deployed at each type of facility are often different. We selected a domain in size of about the T254 model grid within the CART site centered at the central facility, and labeled it as OB2. It includes the facilities EF-11, EF-15, EF-9, EF-12, EF-13, EF-14, CF-1, and CF-2. We also defined the entire CART site as domain OB3, and the CF site as OB1. Then, domain averages over OB2 and OB3 were computed for surface air temperature and precipitation from the Surface Meteorological Observation System Instruments (SMOS) instruments, surface downward and upward solar and longwave radiative fluxes from the Solar Infrared Radiation Station instruments, and latent and sensible heat fluxes from the Energy Balance Bowen Ratio (EBBR) instruments (Table 1). Taking the surface downward solar and longwave fluxes as examples, we show in Figure 2 the biases of GFS forecast (FCST) compared to observations at OB1, and the differences between observations over OB2 and OB1, and the differences between observations over OB3 and OB1, for all days in 2003, during which the observations were available for most of the time at all the sites. Tests of other variables give similar results. It can be seen from Figure 2 that the forecast bias (FCST – OB1) was always much large than the difference between observations over OB2 and OB1. However, the difference between observations over OB3 and OB1 was at least as large as the model bias.

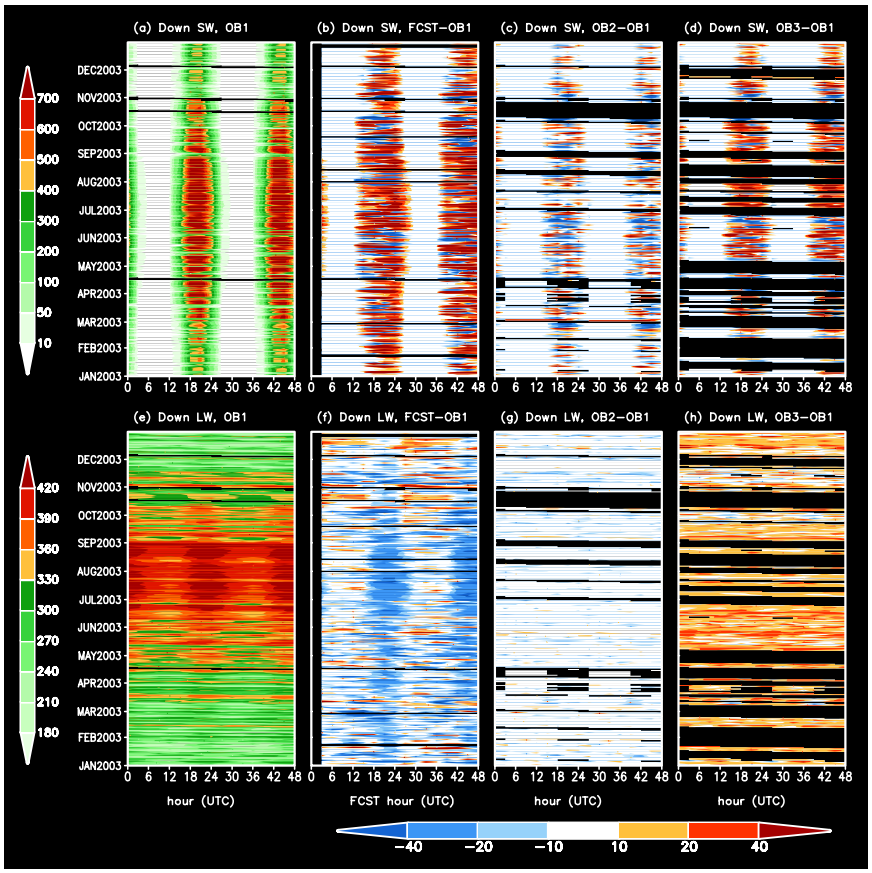


**Figure 2**. A scale-dependence test for surface downward shortwave (upper panels) and longwave (lower panels) fluxes over the ARM SGP site. Shown in the first column are ARM observations over OB1, the second column GFS forecast biases respective to the observations over OB1, the third column the differences between observations over OB2 and OB1, and the fourth column the differences between observations over OB3 and OB1.

This scale-dependence test suggests that comparing the GFS forecast with either the observations at the single point OB1 or with the means over the OB2 domain would lead to qualitatively the same conclusion. From now on, all comparisons will be made between model forecast and the observations over the CF site. It should be point out that the scale-dependence test was not performed for quantities that were only measured at the CF. Nevertheless, we assume that all measurements at CF were of representativeness for the mean conditions over the OB2 domain.

## Surface Energy Fluxes and Surface Air Temperature

The ARM program provides continuous measurements of surface energy fluxes and atmospheric state variables. In this section, we examine these variables in both the observations and the GFS forecast over the SGP CF site. The first goal is to access qualitatively the performance of the model in the four years from 2001 through 2004 for all forecast up to 48 hours. Shown in Figure 3 are ARM observations, and the differences between the GFS forecast and observations, of the surface downward and upward shortwave fluxes and downward longwave flux. All fluxes are 3-hour averages. For instance, fluxes at 6Z represent the means from 3Z to 6Z. For the forecast, fluxes at 0Z were set to missing value since the model was initialized at 00Z and no flux accumulations were available. The observed 3-hourly mean fluxes were derived from the ARM Valued-Added Best-Estimate Radiative Flux (Long 2002; Shi and Long 2002), which has a one-minute resolution (Table 1). In Figure 3, observations in two consecutive days were aligned along the x-axis to match the actual forecast time. This layout allows us to examine qualitatively the diurnal, seasonal and annual variations of these quantities.

In general, the forecast model simulated well the observed diurnal and seasonal variations of the variables shown in Figure 3. Certain biases in magnitude exist. The model overestimated the surface downward shortwave flux in all daytime hours and in all seasons. The largest bias reached about 200 $W/m^2$ at about 21Z (3 PM local time) in the summer. For surface upward (reflected) shortwave flux, the bias was relatively small and sometimes even had an opposite sign to the bias of the downward flux. This discrepancy implies a potential problem in the surface albedo prescribed to the model, which will be discussed later. For the surface downward longwave flux, the model underestimated the flux in the day and overestimated at night. Keep in mind that on 28 August 2003 the longwave radiative transfer module of the GFS was switched from the Chou and Suarez (1994)'s scheme to the Mlawer et al. (1997)'s RRTM. Noticeably, the bias was reduced in the daytime after the switch; however, the bias became larger at night. In Section 5 we will analyze the relationship between radiative fluxes and clouds to better understand the source of errors.

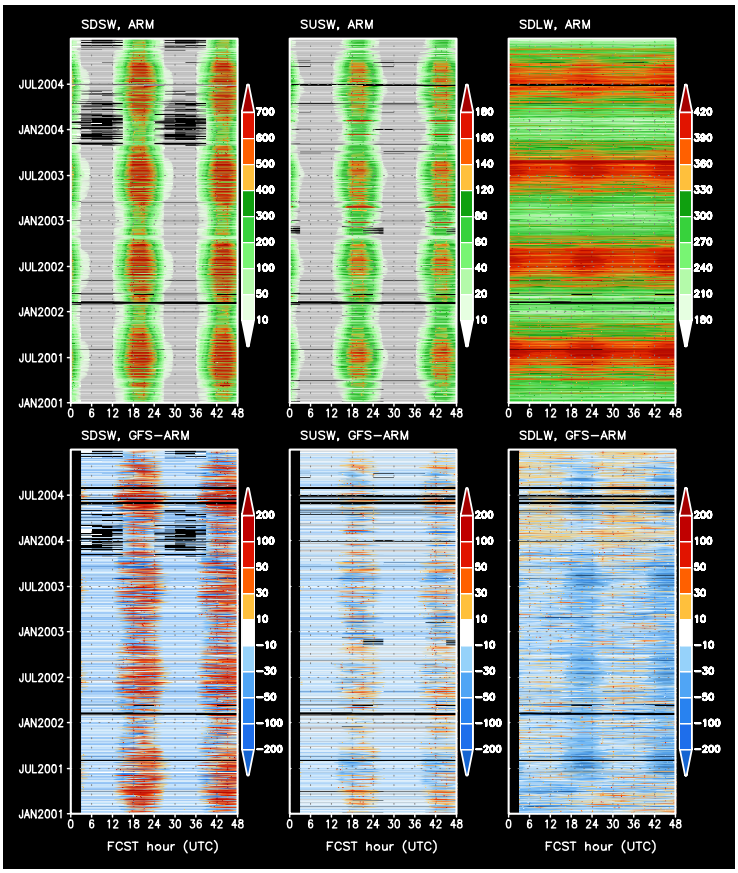


**Figure 3**. 3-hourly mean surface downward shortwave flux (SDSW), upward shortwave flux (SUSW) and downward longwave flux (SDLW) over the ARM SGP CF/E13 site. Upper panels are ARM observations; lower panels are the differences between the GFS forecast and ARM observations.

In Figure 4 we compare further the forecast of surface latent and sensible heat fluxes with ARM observations at the SGP CF site, in a format similar to that of Figure 3. The observations were derived from the 30-minute EBBR measurements (Wesely et al. 1995) at the E13 site (Figure 1), which is collated with the C1 facility. We were aware of the fact that a value-added EBBR product had been produced by the ARM program to correct some of the spikes found in the measurement time series, which were caused by interferences of dew, frost, or condensation on the EBBR radiometer domes (http://science.arm.gov/vaps/baebbr.stm). However, that product was only available before June 2004, and had many missing values. We opted to use the original EBBR product. Cross-validation indicates that the difference between the original and the value-added product is negligibly small compared to the magnitude of model biases (figures not shown). We have screened the original EBBR data before use to exclude some of the visually detectable bad values. Still, as illustrated in Figure 4 the EBBR measurements of latent and sensible heat fluxes had some problem before April 2002.

In Figure 4, latent and sensible heat fluxes are defined upward positive. For the observations, latent heat flux was always from the surface to the atmosphere. Sensible heat flux was primarily from the atmosphere to the surface at night because of strong surface cooling and from the surface to the atmosphere in the day because of surface heating by the sun. The GFS persistently overestimated latent

heat flux in both the daytime and nighttime hours in all seasons. The daytime bias reached up to 200 W/m$^2$. For sensible heat flux, the model tended to overestimate the surface-to-atmosphere flux in the day and underestimate the atmosphere-to-surface flux at night.

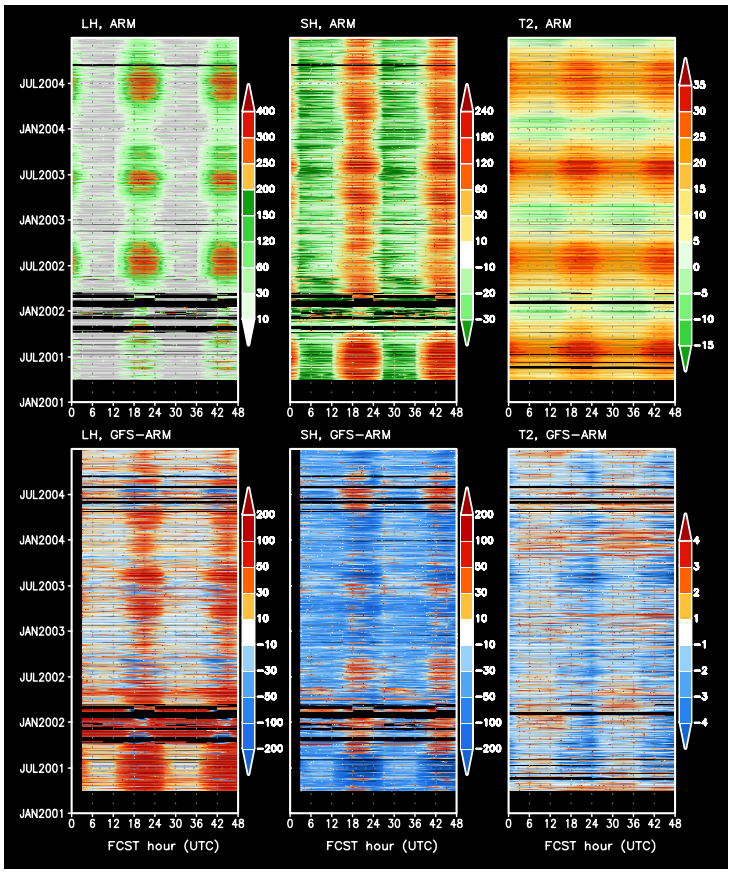


**Figure 4**. Same as in Figure 3, except for latent heat (LH), sensible heat (SH) and surface 2-meter air temperature.

Included in Figure 4 is also the 2-meter air temperature. The observation was derived from the ARM SMOS data stream (Table 1), which had a 30-minute resolution. The model showed predominantly cold biases in the day and warm biases at night.

So far the assessment has focused on the qualitative behavior of the forecast system. The forecast biases shown in Figures 3 and 4 varied strongly with the diurnal cycle. To obtain a quantitative view of the diurnal variations, we present in Figure 5 the mean fluxes and surface air temperature for up to 48 hours of forecast averaged for the two years from 1 July 2002 through 30 June 2004, during which all observations were available and of good quality. The model captured the basic features of the observed diurnal variations of all surface energy fluxes and air temperature, except for the downward longwave radiation. The observed longwave flux peaked in late afternoon at 3 PM to 6 PM (21Z to 24Z), while the forecast peaked in the evening at about 9 PM (27Z). Further analysis indicates that this discrepancy was caused by a bias in the diurnal cycle of forecast clouds in the middle and lower atmosphere (see Section 5). For all other quantities, both the observations and forecast reached maxima at about 21Z

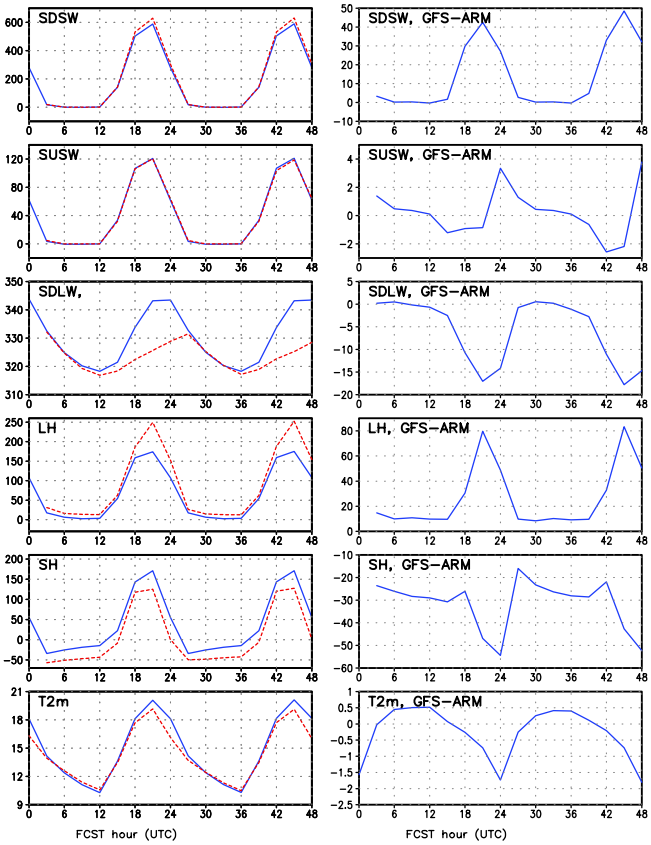


**Figure 5**. Diurnal cycles of surface energy fluxes and surface air temperature averaged for the period from 1 July 2002 through 30 June 2004. Left panels: blue lines are for ARM observations, red dotted lines are for GFS forecast. Right panels: differences between the GFS forecast and ARM observations. Labels are defined in Table 2.

(3 PM local time) and minima at about 12Z (6 AM local time). However, large biases in magnitude for certain variables still exist. At 21Z, the forecast overestimated surface download solar flux by about 42 W/m$^2$ and latent heat flux by 80 W/m$^2$, and underestimated surface downward longwave flux by 17 W/m$^2$. For sensible heat flux, the model underestimated the daytime surface-to-atmosphere flux by about 50 W/m$^2$ at 21Z, and overestimated the nighttime atmosphere-to-surface flux by about 30 W/m$^2$. The forecast was slightly warmer than the observed at night, and colder in the day, reaching up to –1.7°C at 24Z (6 PM local time). It is interesting to note that even though the forecast surface downward solar flux had a large bias, the forecast surface reflected solar flux had only a very small bias, less than 3 W/m$^2$ in all hours of the day. The reason is that surface albedo prescribed in the GFS was smaller than the observed. We will revisit the issue in Figure 6.

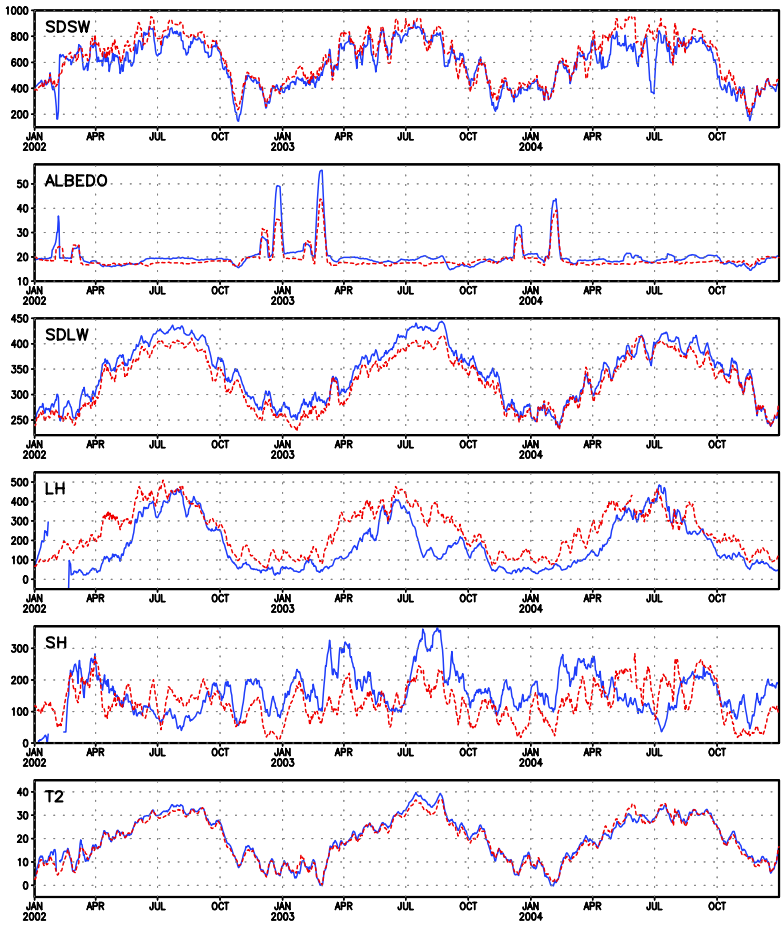


**Figure 6**. Time series at 21Z (3 PM local time) from 1 January 2002 through 31 December 2004 extracted from Figures 3 and 4, except for surface albedo. Blue lines are for ARM observations, and red dotted lines for the GFS forecast. A 10-day running-mean filter was applied to all time series.

The result suggests that there is a cancellation of errors among the forecast surface energy terms. Taking the 21Z (3 PM local) forecast as an example, we list in Table 2 the numbers of surface energy fluxes shown in Figure 5 for both ARM observations and the GFS forecast. For energy budget consideration, included in Table 2 are also surface upward longwave flux and ground soil heat flux, for which the observation resources are given in Table 1. The small net heat flux numbers in Table 2 indicate that surface energy was almost balanced for both the ARM observations and GFS forecast; however, the balance for GFS forecast was a result of cancellation of errors. The forecast model suffered from too strong surface cooling, up to 80 W/m$^2$, by evaporation (latent heat). This large bias was almost compensated by a 43 W/m$^2$ overestimate of downward solar flux and a 48 W/m$^2$ underestimate of upward sensible heat flux. The biases in the surface upward and downward longwave fluxes were not negligible either.

**Table 2**. Surface energy fluxes (W/m$^2$) over the ARM SGP CF site at 21Z (3 PM local time) averaged over the period from 1 July 2002 through 30 June 2004, where SDSW is downward shortwave, SUSW upward shortwave, SDLW downward longwave, SULW upward longwave, SH sensible heat, LH latent heat, GH ground soil heat flux, and NET the summation of all aforesaid surface energy fluxes. For clarity, fluxes that contribute to the gain of surface energy budget are defined positive.

| | SDSW | SUSW | SDLW | SULW | LH | SH | GH | NET |
|---|---|---|---|---|---|---|---|---|
| ARM | 586 | – 121 | 343 | – 435 | – 169 | – 173 | – 32 | – 1 |
| GFS | 629 | – 120 | 325 | – 421 | – 249 | – 125 | – 36 | 3 |
| GFS – ARM | 43 | 1 | – 18 | 14 | – 80 | 48 | – 4 | |

To examine further the seasonal and annual variations of these energy terms, we selected the time series at 21Z (3 PM local time) in Figures 3 and 4 from 1 January 2002 through 31 December 2004, and plotted them in Figure 6. Instead of showing the surface upward solar radiation as in Figure 3, we presented in Figure 6 the observed and forecast surface albedo. The observed albedo was derived from the 1-minute Valued-Added Best-Estimate Radiative Flux (Table 1). The SGP CF is covered by pasture and wheat for most time of the year. Surface albedo is usually in the vicinity of 0.2 if there is no snow on the ground. For the forecast, surface albedo is determined by the model's surface albedo parameterization at each time step and averaged over a 3-hour period. The model tended to underestimate the observed surface albedo by about 0.02 in absolute value in most time of the year. As a result, the forecast upward solar radiation coincidently matched with the observation, even though the forecast largely overestimated the surface downward solar radiation (see Figure 3 and Table 2).

For the time series of surface albedo shown in Figure 6, there were several spikes in the winter and early spring seasons in both the observations and the forecast. These spikes with high surface albedo corresponded to days with snow on the ground. Not surprisingly, the GFS was able to predict the occurrence of snow events well (figures not shown) since observed snow was assimilated into the forecast system's initial conditions. However, surface albedo derived from the model's surface parameterization was always smaller than the observed over snow-covered surface. For the GFS, surface albedo over snow-covered surface depends on snow depth, surface temperature, solar zenith angle and surface roughness (Yang et al. 2001). More experiments and detailed observations are required to investigate the exact cause of the forecast bias in surface albedo.

In Figure 6, the forecast surface downward longwave flux matched with the observed much better after August 2003 than before. In August 2003, the longwave radiative transfer scheme in the GFS was switched from Chou and Suarez (1994) to the RRTM module (Mlawer et al. 1997). The improvement in all-sky flux is encouraging, but may not necessarily point to a problem in the Chou and Suarez module. Possible changes in cloud simulation and the feedbacks between clouds and radiation make the interpretation complicated. In Section 5 we will compare further how the two longwave radiative transfer modules behaved under clear and cloudy-sky conditions.

From Figure 5 we knew that the forecast greatly overestimated daytime surface latent heat flux. The comparison between observations and the forecast in Figure 6 indicates that the overestimate was much larger in the spring and fall transition seasons than in the summer.

In Figure 6, the observed surface sensible heat flux showed significant seasonal and annual variations in the three years of comparison. Usually, the flux was much larger in spring and fall than in winter and

summer. The fluxes in the spring and fall of 2003 were much larger than that in the same seasons of the other years. For the GFS forecast, the flux did not vary much with the season and year. Overall, the model underestimated the early-afternoon surface-to-atmosphere sensible heat flux.

In this section, the comparison indicates that the balance of surface energy in the forecast was a result of cancellation of errors among the individual flux terms. The forecast largely overestimated evaporation, surface downward solar radiation and underestimated sensible heat flux in the day. The forecast gave a surface albedo lower than the observed, especially over snow-covered surfaces. The largest bias in latent heat flux occurred in the spring and fall instead of in the summer. The forecast did not capture well the observed seasonal and interannual variations of sensible heat flux. The GFS simulated better surface downward longwave flux after the change of the GFS radiative transfer module in August 2003. In Section 5 we will demonstrate how clouds may have played a role in modulating the forecast biases identified in this section. The following section is dedicated to the evaluation of clouds.

# Cloud Fraction

Clouds remain a major source of uncertainty in both climate and numerical weather forecast models. Improvement in cloud simulation has been slow, in part because satellite and conventional ground observations can not provide detailed information about cloud vertical structure for model valuation. One of the major objectives of the ARM program is to obtain detailed cloud information that can be use to improve the understanding of cloud and radiation interaction (Stokes and Schwartz 1994). Although the ARM observations are limited to a few sites over the globe, they have proved to be invaluable for diagnosing and improving climate model physics when used in combination with ground and satellite observations (e.g., Jakob 2003; Philips et al. 2004). Here we compare GFS forecast with ARM observations for clouds over the SGP CF site from 2001 through 2004.

## Definition of cloud fraction

For ARM observations, Clothiaux et al. (2000) and Clothiaux et al. (2001) created an active remotely-sensed clouds locations (ARSCL) value-added product. This product combines laser ceilometers, microwave radiometers, and micropulse lidars to produce an objective determination of hydrometeor height distributions and estimates of their radar reflectivities, vertical velocities, and Doppler spectral widths. A subset of this product (data stream sgparsclbnd1clothC1.c1 in Table 1) gives information of cloud bases and top heights for each group of contiguous clouds in the vertical column at a 10-second temporal resolution. At the time of this investigation, this subset of product for the SGP site is available up to June 2004.

For the GFS forecast, vertical distributions of cloud fraction were saved on the model sigma layers as instantaneous values, instead of 3-hour means, from the last call of the model physical subroutines during each of the 3-hour post-processing period. The output is under-sampled in time. The time step for the GFS physics was 7.5 minutes in the T170 version before 29 October 2002, and was 5 minutes in the T254 version after.

There is an issue of comparability in space and time between the observations and model forecast. Cloud fraction from the model represents the percentage of a model grid area that is covered by clouds

within each model physical time step. For single-point observations, cloud fraction has been conventionally defined as cloud occurrence frequency in a given period of time (e.g., Lazarus et al. 2000). The question is how to choose the averaging period. One approach is to match the period exactly to the duration of the GFS last physics time step within each of the 3-hour output period. However, the so-defined cloud fraction (occurrence frequency) is under-sampled in space compared to its peer from the model. To compensate the insufficient sampling in space, one can define alternatively cloud occurrence frequency over an "impact" period of time, which, in some degree, represents the dynamical advection of cloud condensates from neighboring points within a model grid that surrounds the observation column. This "impact" period varies with space and time since the time scale of dynamical transport by advection depends on wind speed. We selected a few constant "impact" periods, ranging from a few minutes up to three hours, and compute the corresponding cloud fractions (occurrence frequency). The purpose is to test the sensitivity of cloud occurrence frequency to different definitions of "impact" period.

Before use, the observed ARSCL cloud base and top heights, which had a 10-second resolution and were archived from 1 January 2001 through 30 June 2004, were mapped to standard heights with a 250-meter vertical resolution from the surface up to 20 km. Cloud occurrence was recorded as one if clouds were detected, and zero if no clouds were detected on each of the standard height. Then we divided the entire archive into 3-hour subsets. Cloud occurrence frequencies were computed by averaging the 10-second cloud occurrence records in the last 5 minutes, 7.5 minutes, 15 minutes, one hour and three hours, respectively, over each of the 3-hour subset. As a result, cloud occurrence frequency derived with the "impact" period shorter than 3 hours used only part of the observations within each 3-hour subset.

For demonstration, we show in Figure 7 the vertical distributions of the observed cloud occurrence frequency in July 2003 and January 2004 for two extreme cases, the 5-minute and 3-hour "impact" periods. Not surprisingly, the overall structures of the cloud system computed with the two methods are similar, only that there are more clouds with fractional cover less than 100% for the 3-hour case than for the 5-minute case. For each case, we computed further the total (column-integrated) cloud amount, using a random-overlap assumption, for the two years from 1 July 2002 through 30 June 2004, and plotted in Figure 8 the time series of total cloud amount and the two-year-mean cloud amount. One can see that at times the total cloud amount for the 5-miniute case, which is under-sampled in time, can be very different from that for the 3-hour case. The two-year mean total cloud amount is 53.7% for the 5-minute case, and 57.8% for the 3-hour case. For other cases, the mean total cloud amount falls in between.

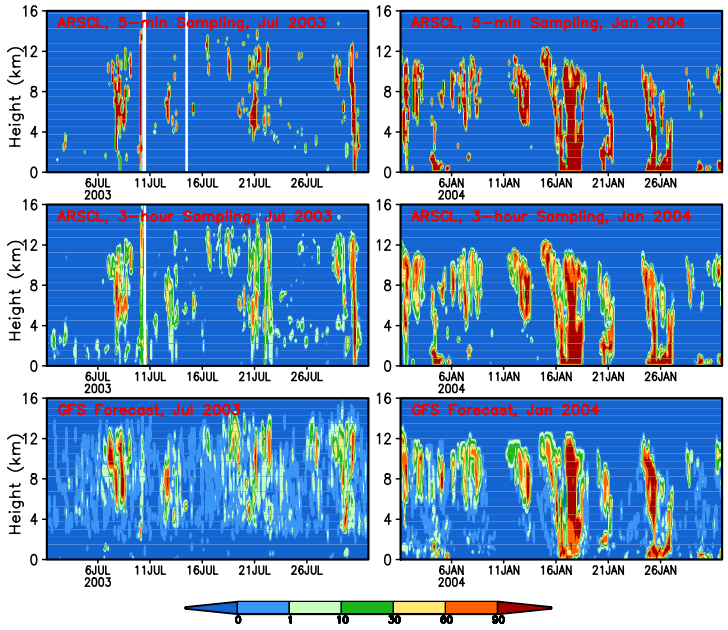


**Figure 7**. Vertical distributions of cloud occurrence frequency for ARM ARSCL observations, and fractional cloud cover for the GFS forecast in July 2003 (left panels) and January 2004 (right panels). For observations, the 5-minute and 3-hour sampling methods were used to compute cloud occurrence frequency (see text for details).

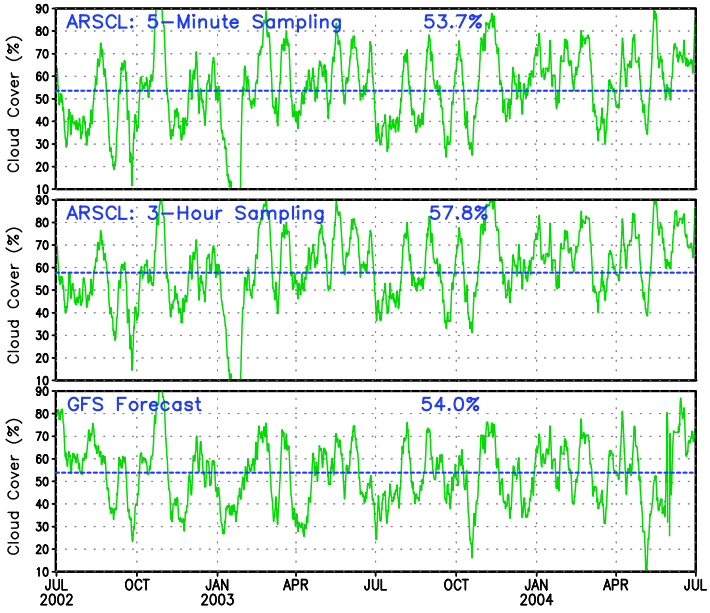


**Figure 8**. Total cloud amount integrated from the surface to 16 km using a random-overlap assumption. All data used have a 3-hour temporal and 250-meter vertical resolution. The time series were smoothed by a 10-day running-mean filter to exclude high-frequency variations. The thick dotted line and the number on the top of each panel represent mean cloud amount averaged over the two years from July 2002 through June 2003.

The investigation indicates that cloud occurrence frequency does vary with the definition of "impact" period used for time averaging. In practice, we do not know which "impact" period is the most realistic one. When comparing model clouds with observations, one has to keep in mind the uncertainty of "observations" arising from the different definitions of cloud occurrence frequency. For this investigation, in each of the 3-hour processing period, the GFS clouds were under-sampled in time; the ARM clouds were under-sampled in space, and in time if not all data within the 3-hour period were used for deriving cloud occurrence frequency.

## GFS Forecast and ARM observations

To compare GFS clouds with the observed ARSCL clouds, the archived GFS instantaneous cloud fractions on sigma layers were projected onto the same standard heights of 250-metere resolution as for the observations. For illustration, the vertical distributions of GFS clouds in the first 24 hours of forecast at the ARM CF site are shown in Figure 7 for July 2003 and January 2004, together with the corresponding ARSCL cloud occurrence frequency distributions. The model captured the observed cloud structure for all major synoptic events in both the warm and clod months, such as the strong convections in late July 2003 and in the middle of January 2004. The forecast compared more favorably with the observation from the 3-hour sampling than that from the 5-minute sampling. The model underestimated clouds in the boundary layer and lower troposphere. There were many clouds with fractional cloud cover less than 1% in the forecast. The model was capable of producing such clouds whenever the relative humidity slightly exceeds a critical value. For observations, the instruments might not be able to detect these clouds.

We also computed and plotted in Figure 8 the total cloud amount for the forecast from July 2002 through June 2004. The forecast captured generally well the ups and downs in the observed time series. The two-year-mean total cloud amount for the forecast is 54%, which falls to the lower end of the observed ones based on different sampling methods but is within the uncertainty of observations.

Judging from the column-integrated cloud amount, the forecast model performed reasonably well. However, this does not guarantee the model was capable of simulating well the vertical structure of clouds, which are of more importance to model physics since they strongly influence radiative heating in the atmosphere and radiative fluxes intercepted by the surface. To examine the diurnal cycles of clouds in all layers of the troposphere, we compare in Figure 9 the vertical distributions of cloud fractions between GFS forecast and ARM observations for up to 48 hours of the forecast averaged for the two years from 1 July 2002 through 30 June 2004.

For observations (Figure 9a), the diurnal cycles were very different in different layers of the troposphere. In the boundary layer and lower troposphere, cloud cover reached its maximum in the day around noon (18Z) and its minimum at midnight (6Z). In the middle and upper troposphere, the diurnal cycle was much weaker and had an opposite phase to the one for lower-layer clouds. More clouds were observed at night than in the day. Zooming into seasons, the diurnal cycle for low clouds was much stronger in the spring and fall than in other seasons, the diurnal cycle for high clouds was the strongest in the summer (figures not shown). The analysis indicates that the SGP CF site is dominated by

stratiform clouds in the lower troposphere in the day, and circus clouds in upper troposphere at night. For in-depth analyses of the cloud characteristics over the SGP CF site readers are referred to the work of Lazarus et al. (2000).

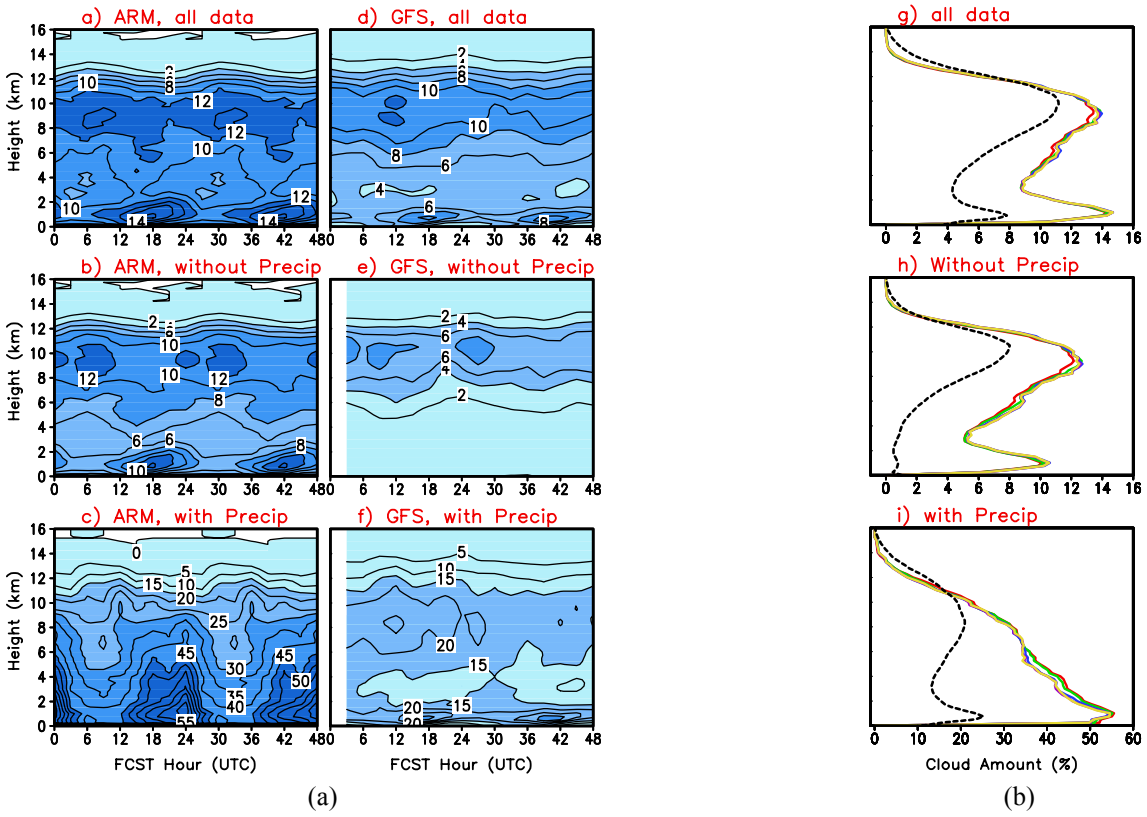


**Figure 9**. Vertical distributions of cloud diurnal cycle and annual mean cloud fraction. Middle column: diurnal cycles of cloud fraction for the first 48 hours of GFS forest averaged for the two years from 1 July 2004 through 30 June 2004. Left Column: same as the middle column except for the ARM ARSCL clouds computed using the 3-hour sampling method (see text for details). Right Column: mean cloud fractions averaged over the 48 hours in the left two columns. The black dotted lines are for GFS forecast, and the clusters of colored lines are for ARM ARSCL observations, computed using five different sampling methods as described in the text, including the 3-hour sampling method. In the middle and lower panels, could fractions were computed for hours with and without precipitation, respectively.

For the GFS forecasts (Figure 9b), the model captured the diurnal cycle of clouds in the boundary layer and lower troposphere, but with a much weaker amplitude and less cloud amount. This can be seen more clearly in Figure 9g, which shows the vertical distributions of clouds for both observations and forecasts averaged for all the hours shown in Figures 9a and 9b. For observations, mean clouds derived with five different sampling methods as described in Section 4.1 were included. The purpose is to quantify model bias against the uncertainty in observations. One can see that the forecast bias is much larger than the spread of observations deduced from different sampling methods. The forecast cloud fraction was always a few percent smaller than the observed in the troposphere below 10 kilometer,

about 7% less for low clouds and 3% less for high clouds. The forecast slightly overestimated cloud fraction near the tropopause. The GFS model was able to simulate a bi-modal vertical distribution of clouds similar to the observed one, with the peaks located at about one and ten kilometers, respectively.

We further analyzed cloud distributions for conditions with and without precipitation. The results are presented in Figure 9 as cloud diurnal cycles. For ARM observations, there were many non-precipitating low clouds in the day and high clouds at night. The GFS captured the nighttime non-precipitating high clouds, but totally missed the daytime non-precipitating low clouds (see Figure 9b and 9e there isn't a 9c-e). The forecast cloud amount with the presence of precipitation was also much less than the observed in the middle and lower troposphere. The lack of low clouds has been found to be a common problem over the globe in the GFS forecast. NCEP is actively working on improving the shallow convection parameterization in the forecast model.

## Cloud and Radiation Interaction

In previous sections we compared separately the GFS forecast clouds and radiative fluxes with observations. It is not clear to what extent the biases in radiative fluxes are attributable to the biases in clouds. In this section, the forecast radiative fluxes under clear-sky and cloudy-sky conditions are separately examined to understand further the source of errors.

In Figure 5 we showed that, on average, the observed all-sky SDLW peaked at 3 to 6 PM while the forecast flux peaked at 6 to 9 PM, and the magnitude of the later was about 15 W/m$^2$ smaller than the former. The longwave radiative transfer routine in the GFS was called once per three hours. There is the possibility that the infrequent calling led to a shift in the forecast longwave flux. If this is the case, this shift should exist under both clear-sky and cloudy-sky conditions. Figure 10 shows the diurnal cycles of SDLW, averaged for the two years from 1 July 2002 through 30 June 2004, for ARM observations and GFS forecast under clear-sky and cloudy-sky conditions, respectively. Under clear sky, the forecast flux matched well with the observation in both magnitude and phase. The flux reached its maximum of about 310 W/m$^2$ in the evening to early night (9 to 12 PM) and its minimum of about 285 W/m$^2$ in the early morning (3 to 6 AM). However, the distribution of cloudy-sky flux is similar to that of all-sky flux. The forecast flux reached its peak at about 3 hours later than did the observed, and was more than 10 W/m$^2$ smaller in magnitude. This analysis leads us to conclude that the error in the forecast SDLW was primarily caused by the errors in the forecast clouds rather than the infrequent calling of the longwave radiative transfer routine. Low clouds radiate almost like black body. The diurnal cycle of observed SDLW followed the diurnal cycle of low clouds (see Figure 9a) with a few hours delay. The GFS forecast had a very weak diurnal cycle of low clouds (Figure 9d), and missed almost all non-precipitating low clouds (Figure 9e). On the other hand, for both observations and forecast, high clouds have a diurnal cycle almost opposite to low clouds (Figures 9a and 9d). There were more high clouds at night than in the day. The lack of low clouds in the forecast and the opposite diurnal cycle of high clouds to that of low clouds caused the forecast all-sky SDLW reached its maximum a few hours later than the observed. The forecast all-sky flux peaked at the same time as the forecast clear-sky flux, at about 9 PM.

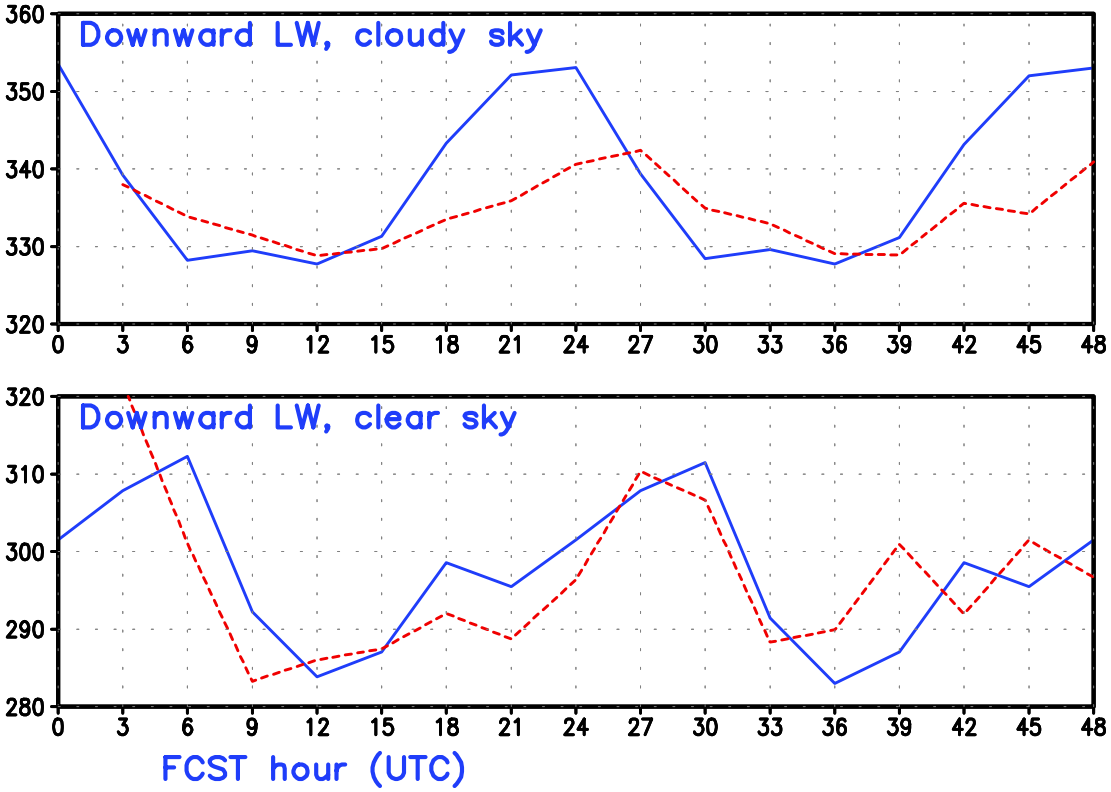


**Figure 10**. Diurnal cycle of surface downward longwave flux averaged for the period from 1 July 2002 through 30 June 2004. Upper panel is for cloudy-sky conditions, and lower panel for clear-sky conditions. Blue lines are for ARM observations, and red dotted dashed lines for GFS forecast.

The longwave radiative transfer module in the GFS was switched from the GFDL Schwarzkopf and Fels (1991) scheme to the AER RRTM (Mlawer et al. 1997) on 28 August 2003. In Figures 3 and 6 the forecast SDLW had a distinct change before and after this switch. The forecast daytime flux was underestimated before the switch with the GFDL scheme, and matched rather closely with the observed after the switch with the AER scheme (Figure 6). However, the forecast nighttime flux became worse after the switch in August 2003, about 10 to 30 W/m$^2$ larger than the observed (Figure 3). We compare in Figure 11 the forecast SDLW with the observed for time before and after the switch, further categorized by daytime (9 AM to 6 PM) and nighttime (9 PM to 6 AM) and by clear-sky and cloudy-sky conditions. Under cloudy-sky conditions, the GFS forecast systematically underestimated the flux in both daytime and nighttime (Figures 11d and 11b) with the GFDL module before 28 August 2003. The forecast flux was greatly improved in both daytime and nighttime with the AER module after the switch (Figures 11f and 11h). The systematical low biases were removed, although there were still considerable day-to-day errors. Under clear-sky conditions, the result is complicated. By comparing Figure 11g with Figure 11c one can see that the daytime clear-sky flux was improved with the AER module. The systematic low bias in the forecast with the GFDL module (Figure 11c) was reduced. For nighttime clear-sky flux, however, a systematic high bias occurred with the AER module (Figure 11e). The analysis indicates that the large positive bias of SDLW at night after August 2003 identified in Figure 3 was mostly due to clear-sky flux instead of cloudy-sky flux.

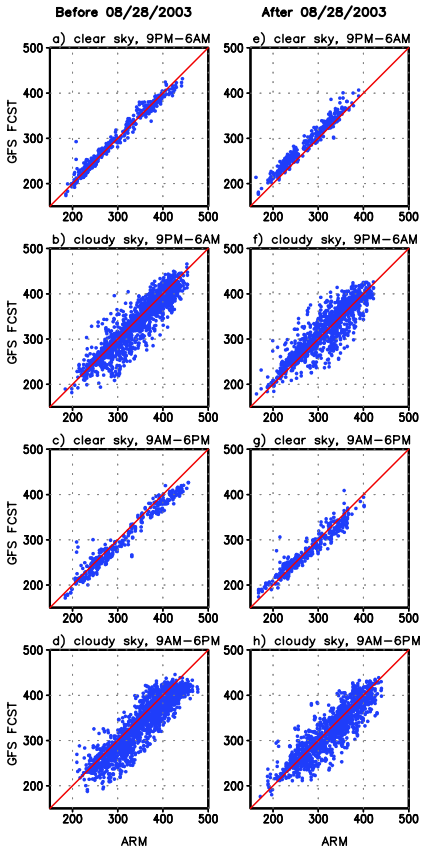


**Figure 11**. Scatter plot of 3-hourly mean surface downward longwave flux, GFS forecast against ARM observations. Left (right) panels are for forecast with the GFDL (AER) longwave radiative transfer routine before (after) 28 August 2003. Upper (lower) four panels are for nighttime (daytime) fluxes from 9 PM to 6 AM (9 AM to 6 PM). Fluxes under clear and cloudy-sky conditions are separated.

We further analyzed the GFS forecast and found that there were no distinct changes in the forecast of clouds before and after the use of the AER module. The improvement in cloudy-sky SDLW after the use of the AER module was most likely attributable to a better treatment of clouds for longwave (LW) radiative transfer by the RRTM. For clear-sky SDLW, the reason for the deterioration at night after the use of AER module (Figure 11e) is not all clear. Our results may not necessarily imply that the AER module itself has problem. We compared the forecast and observed atmospheric temperature at night in the middle and lower troposphere for time before and after the implementation of the AER module on 28 August 2003 (Figure 12) and for clear-sky and cloudy-sky conditions, respectively. (Upper tropospheric temperature was excluded because of its much weaker contribution to SDLW than the lower tropospheric temperature.) The forecast nighttime temperature had a slight warm bias under both clear and cloudy-sky conditions before 28 August 2003 (Figures 12a and 12b). After that time, the warm bias became lager, especially for the clear-sky case (Figure 12c). Similar analysis was performed for atmospheric water vapor (figures not shown). No obvious changes were found before and after August 2003 under either cloudy or clear-sky conditions. Therefore, the deterioration of SDLW was probably caused by the large warm bias in atmospheric temperature at night under clear-sky conditions after August 2003. At this moment, we were not able to explain why the forecast nighttime temperature in the troposphere became warmer after August 2003. This could be coincidental, or actually related to

higher atmospheric longwave heating rates from the AER module itself. Feedbacks intrinsic to the GFS forecasts prevented us from isolating the cause from effect.

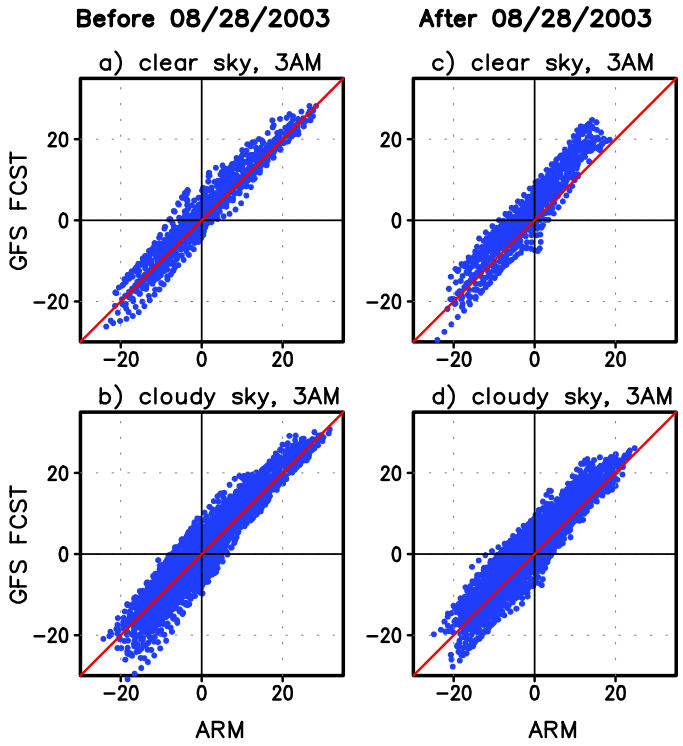


**Figure 12**. Scatter plot of temperature in the lower and middle atmosphere below 5 km at night (3 AM), GFS forecast against ARM observations. Upper (lower) panels are for clear-sky (cloudy-sky) condition, and left (right) panels for forecast before (after) 28 August 2003.

At last, we compare the surface downward solar flux (SDSW) under clear and cloudy-sky conditions in the two years from 1 July 2002 through 30 June 2004 (Figure 13). The forecast clear-sky flux was rather accurate compared to the ARM observations. The SDSW bias identified in Figure 3 and 5 was due to cloudy-sky fluxes. The forecast systematically overestimated SDSW (Figure 13b), and the spread of day-to-day errors is much larger than that for longwave flux. The forecast of solar flux depends strongly on the accuracy of clouds.

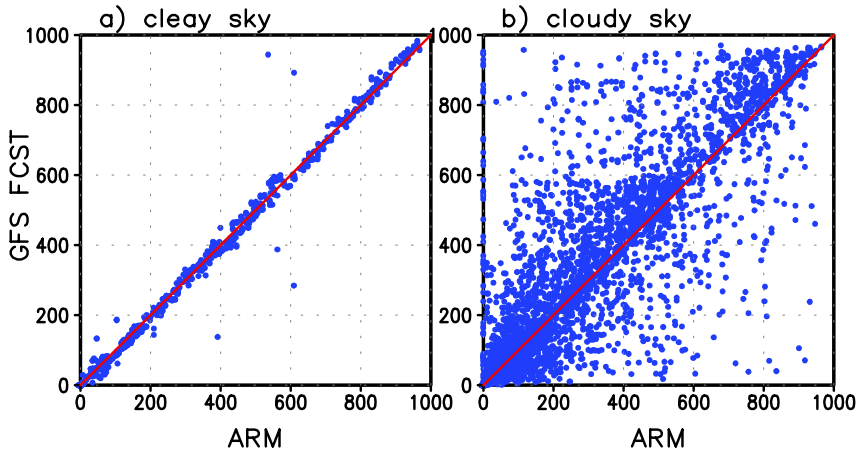


**Figure 13**. Scatter plot of 3-hourly surface downward shortwave flux, GFS forecast against ARM observations, for a) clear-sky and b) cloudy-sky conditions. Time period: 1 July 2002 through 30 June 2004.

# Conclusion

In this study, we evaluated the NCEP operational NWP forecast against ARM observations in the years from 2001 through 2004 at the ARM SGP site, focusing primarily on surface energy fluxes and clouds in the first 48 hours of forecast. Unlike previous studies (e.g. Hinkelman et al. 1999; Morcrette 2002) that targeted on specific synoptic events and utilized ARM observations in a particular season or year to evaluate operational NWP models, we examined the statistical means of the NCEP forecast. The purposes are to diagnose systematic model biases, to explore the source of errors, and to recommend changes for future model development.

For the forecast period we evaluated, the GFS experienced two major changes in configuration, one in spectral and vertical resolutions and the other in longwave radiative transfer algorithm. The forecast showed certain persistent biases, some of which depended strongly on the season and time of the day. For instance, the model tended to overestimate latent heat flux at all time, by up to 80 W/m$^2$ at 3 PM. From the surface energy budget point of view, the forecast bias in latent heat flux was primarily compensated by an overestimate in downward solar flux and an underestimate of sensible heat flux. For clouds, the model was able to capture the observed cloud structures during major synoptic events. However, on average, the model underestimated cloud fractions in the lower and middle troposphere, and slightly overestimated high clouds in the upper troposphere in all seasons. The forecast did not simulate the observed strong diurnal cycle of clouds in the lower troposphere, and missed all the observed daytime non-precipitating clouds in the lower troposphere. This bias was responsible for a phase shift in the forecast of diurnal cycle of surface downward longwave flux in comparison with the ARM observations. Shortwave flux in the forecast was rather accurate under clear-sky conditions, and was overestimated under cloudy-sky conditions.

We compared separately the forecast surface downward longwave flux with ARM observations for clear and cloudy-sky conditions, and for the time before and after the change of the longwave radiative

transfer algorithm on 28 August 2003. The forecast cloudy-sky flux was greatly improved in both daytime and nighttime after the use of the AER RRTM algorithm. A systematical low bias found before the change was eliminated. The forecast daytime clear-sky flux was also improved with the AER module. However, the nighttime flux was largely overestimated with the AER module. This bias was probably caused by a nighttime larger warm bias of atmospheric temperature under clear-sky conditions after August 2003.

This investigation focused on the surface fluxes and clouds. The intensive ARM observations at high temporal and vertical resolutions allowed us to examine in detail the diurnal cycles of the surface fluxes and the vertical structure of clouds. We were able to link some of the forecast biases to particular physical processes; however, for most of the forecast biases there are no simple explanations. Many physical processes and feedbacks might have contributed to the forecast biases. For example, the underestimate of forecast clouds caused excessive solar radiation reaching to the surface, which, in turn, might be partially responsible for the excessive surface latent heat flux. A better forecast of cloud vertical structure, in addition to the vertically integrated total cloud amount, is crucial for improving the overall performance of the forecast model.

Acknowledgments: Fanglin Yang was support by the Department of Energy Atmospheric Radiation Measurement Program. The authors wish to thank Steven Lazarus for his contribution to producing the GFS column output, and to thank scientists at the ARM Archive Center for providing ARM observations.

# References

Ackerman, T, and GM Stokes. 2003. "The Atmospheric Radiation Measurement Program." *Physics Today*, 56:38-45.

Arakawa, A, and WH Shubert. 1974. "Interaction of a cumulus ensemble with the large-scale environment, Part I." *Journal of Atmospheric Science* 31:674-704.

Briegleb, BP, P Minnus, V Ramanathan, and E Harrison. 1986. "Comparison of regional clear-sky albedo inferred from satellite observations and model computations." *Journal of Climate and Applied Meteorology* 25:214-226.

Chou, MD, and MJ Suarez. 1994. An efficient thermal infrared radiation parameterization for use in general circulation models. Technical Memorandum. 104606, Vol. 3.

Chou, MD, and MJ Suarez. 1999. A solar radiation parameterization for atmospheric studies. NASA Tech. Memo. 104606, Vol. 11, 40 pp.

Clothiaux, EE, TP Ackerman, GG Mace, KP Moran, RT Marchand, M Miller, and BE Martner. 2000. "Objective determination of cloud heights and radar reflectivities using a combination of active remote sensors at the ARM CART Sites." *J Applied Meteorology* 39:645-665.

Clothiaux, EE, MA Miller, RC Perez, DD Turner, KP Moran, BE Martner, TP Ackerman, GG Mace, RT Marchand, KB Widener, DJ Rodriguez, T Uttal, JH Mather, CJ Flynn, KL Gaustad, and B Ermold. 2001. "The ARM Millimeter Wave Cloud Radars (MMCRs) and the Active Remote Sensing of Clouds (ARSCL) Value Added Product (VAP)." Department of Energy Technical Memo. ARM VAP-002.1, U.S. Department of Energy, Washington D.C., 56 pp.

Clough, SA, MJ Iacono, and J-L Moncet. 1992. "Line-by-line calculations of atmospheric fluxes and cooling rates: Application to water vapor." *Journal of Geophysical Research* 97:15761-15785.

Ebert, EE, and JA Curry. 1992. "A parameterization of ice cloud optical properties for climate models." *Journal of Geophysical Research* 97:3831-3836.

Fu, Q. 1996. "An accurate parameterization of the solar radiative properties of cirrus clouds for climate models." *Journal of Climate* 9:2058-2082.

Grell, GA. 1993. "Prognostic evaluation of assumptions used by cumulus parameterizations." *Monthly Weather Review* 121:764-787.

Hinkelman, LM, TP Ackerman, and RT Marchand. 1999. "An evaluation of NCEP Eta model predictions of surface energy budget and cloud properties by comparison to measured ARM data." *Journal of Geophysical Research* 104:19535-19549.

Hou, Y-T, S Moorthi, and KA Campana. 2002. Parameterization of solar radiation transfer in the NCEP models. NCEP Office Note, No. 441, 14pp.

Hu, YX, and K Stamnes. 1993. "An accurate parameterization of the radiative properties of water clouds suitable for use in climate models." *Journal of Climate* 6:728-742.

Jakob, C. 2003. "An improved strategy for the evaluation of cloud parameterizations in GCMs." *Bulletin of the American Meteorological Society* 84, 1387–1401.

Kalnay, E, M Kanamitsu, and WE Baker. 1990. "Global numerical weather prediction at the National Meteorological Center." *Bulletin of the American Meteorological Society* 71:1410-1428.

Kanamitsu, M. 1989. "Description of the NMC global data assimilation and forecast system." *Weather and Forecasting*, 4:335-342.

Lazarus, SM, SK Krueger, and GG Mace. 2000. "A cloud climatology of the Southern Great Plains ARM CART." *Journal of Climate* 13:1762-1775.

Long, CN. 2002. The ARM Southern Great Plains Central Facility Best Estimate Radiative Flux CD. DOE Tech. Memo., ARM-TR-007, U.S. Department of Energy, Washington D.C., 19pp.

Mace, GG, C Jakob, and KP Moran. 1998. "Validation of hydrometeor occurrence predicted by the ECMWF model using millimeter wave radar data." *Geophysical Research Letter* 25:1645-1648.

Matthews, E. 1985. "Atlas of Archived Vegetation, Land Use, and Seasonal Albedo Data Sets." NASA Technical Memorandum 86199, Goddard Institute for Space Studies, New York.

Miyakoda, K, and J Sirutis. 1986. Manual of the E-physics. Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542, 57pp.

Mlawer, EJ, SJ Taubman, PD Brown, MJ Iacono, and SA Clough. 1997. "Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave." *Journal of Geophysical Research* 102:16663-16682.

Morcrette, J-J. 2002. "Assessment of the ECMWF model cloudiness and surface radiation fields at the ARM SGP site." *Monthly Weather Review* 130:257-277.

Pan, H-L, and L Mahrt. 1987. "Interaction between soil hydrology and boundary layer developments." *Boundary Layer Meteorology* 38:185-202.

Pan, H-L, and W-S Wu. 1995. (a 1994 is cited in the text on page 4) Implementing a Mass Flux Convection Parameterization Package for the NMC Medium-Range Forecast Model. NMC Office Note, No. 409, 40pp. [Available from NCEP, 5200 Auth Road, Washington, D.C. 20233]

Phillips, TJ, GL Potter, DL Williamson, RT Cederwall, JS Boyle, M Fiorino, JJ Hnilo, JG Olson, S Xie, and JJ Yio. 2004. "Evaluating parameterizations in general circulation models: climate simulation meets weather prediction." *Bulletin of the American Meteorological Society* 85:1903–1915.

Randall, DA, and DG Cripe. 1999. "Alternative methods for specification of observed forcing in single-column models and cloud system models." *Journal of Geophysical Research* 104:24527-24545.

Schwarzkopf, MD, and SB Fels. 1991. "The simplified exchange method revisited: An accurate rapid method for computation of infrared cooling rates and fluxes." *Journal of Geophysical Research* 96:9075-9096.

Sela, J. 1980. "Spectral modeling at the National Meteorological Center." *Monthly Weather Review* 108:1279-1292.

Shi, Y, and CN Long. 2002. Best Estimate Radiation Flux Value Added Procedure: Algorithm Operational Details and Explanations. Department of Energy Technical Memorandum, ARM-TR-008, U.S. Department of Energy, Washington D.C., 55pp.

Slingo, A. 1989. "A GCM parameterization for the shortwave radiative properties of water clouds." Journal of Atmospheric Science 46:1419-1427.

Staylor, WF and AC Wilbur. 1990. Global surface albedoes estimated from ERBE data. Preprints of the Seventh Conference on Atmospheric Radiation, San Francisco CA, American Meteorological Society, 231-236.

Stokes, GM, and SE Schwartz. 1994. "The Atmospheric Radiation Measurement (ARM) Program: Programmatic Background and Design of the Cloud and Radiation Test Bed." *Bulletin of the American Meteorological Society* 75:1201-1221.

Sundqvist, H, E Berge, and JE Kristjansson. 1989. "Condensation and cloud studies with mesoscale numerical weather prediction model." *Monthly Weather Review* 117:1641- 1757.

Tiedtke, M. 1983. The sensitivity of the time-mean large-scale flow to cumulus convection in the ECMWF model. ECMWF Workshop on Convection in Large-Scale Models, 28 November-1 December 1983, Reading, England, pp. 297-316.

Wesely, MW, DR Cook, and RL Coulter. 1995. Surface Heat Flux Data from Energy Balance Bowen Ratio Systems. Preprints of the Ninth Symposium on Meteorological Observations and Instrumentation, Charlotte, NC, 27-31 March 1995, American Meteorological Society, Boston, MA, pp. 486-489.

Xie, SC, and MH Zhang. 2000. "Analysis of the convection triggering condition in the NCAR CCM using ARM measurements." *Journal of Geophysical Research* 105:14983-14996.

Xu, KM, RT Cederwall, LJ Donner, WW Grabowski, F Guichard, DE Johnson, M Khairoutdinov, SK Krueger, JC Petch, DA Randall, CJ Seman, WK Tao, DH Wang, SC Xie, JJ Yio, MH Zhang. 2002. "An intercomparison of cloud-resolving models with the atmospheric radiation measurement summer 1997 intensive observation period data." *Quarterly Journal of the Royal Meteorological Society* 128:593-624.

Xu, KM, and DA Randall. 1996. "A semiempirical cloudiness parameterization for use in climate models." *Journal of Atmospheric Science* 53:3084-3102.

Yang, F, A Kumar, W Wang, H-MH Juang, and M Kanamitsu. 2001. "Snow-albedo feedback and seasonal climate variability over North America." *Journal of Climate (Letter)* 14:4245-4248.

Zhao, QY, and FH Carr. 1997. "A prognostic cloud scheme for operational NWP models." *Monthly Weather Review* 125:1931-1953.