Evaluation of the NCEP Global Forecast System at the ARM SGP Site

FANGLIN YANG AND HUA-LU PAN

Environmental Modeling Center, National Centers for Environmental Prediction, Camp Springs, Maryland

STEVEN K. KRUEGER

Department of Meteorology, University of Utah, Salt Lake City, Utah

SHRINIVAS MOORTHI AND STEPHEN J. LORD

Environmental Modeling Center, National Centers for Environmental Prediction, Camp Springs, Maryland

(Manuscript received 2 August 2005, in final form 13 March 2006)

ABSTRACT

This study evaluates the performance of the National Centers for Environmental Prediction Global Forecast System (GFS) against observations made by the U.S. Department of Energy Atmospheric Radiation Measurement (ARM) Program at the southern Great Plains site for the years 2001–04. The spatial and temporal scales of the observations are examined to search for an optimum approach for comparing grid-mean model forecasts with single-point observations. A single-column model (SCM) based upon the GFS was also used to aid in understanding certain forecast errors. The investigation is focused on the surface energy fluxes and clouds. Results show that the overall performance of the GFS model has been improving, although certain forecast errors remain. The model overestimated the daily maximum latent heat flux by 76 W m^{-2} and the daily maximum surface downward solar flux by 44 W m^{-2}, and underestimated the daily maximum sensible heat flux by 44 W m^{-2}. The model's surface energy balance was reached by a cancellation of errors. For clouds, the GFS was able to capture the observed evolutions of cloud systems during major synoptic events. However, on average, the model largely underestimated cloud fraction in the lower and midtroposphere, especially for daytime nonprecipitating low clouds because shallow convection in the GFS does not produce clouds. Analyses of surface radiative fluxes revealed that the diurnal cycle of the model's surface downward longwave flux (SDLW) was not in phase with that of the ARM-observed SDLW. SCM experiments showed that this error was caused by an inaccurate scaling factor, which was a function of ground skin temperature and was used to adjust the SDLW at each model time step to that computed by the model's longwave radiative transfer routine once every 3 h. A method has been proposed to correct this error in the operational forecast model. It was also noticed that the SDLW biases changed from mostly negative in 2003 to slightly positive in 2004. This change was traced back to errors in the near-surface air temperature. In addition, the SDLW simulated with the newly implemented Rapid Radiative Transfer Model longwave routine in the GFS is usually 5–10 W m^{-2} larger than that simulated with the previous routine. The forecasts of surface downward shortwave flux (SDSW) were relatively accurate under clear-sky conditions. The errors in SDSW were primarily caused by inaccurate forecasts of cloud properties. Results from this study can be used as guidance for the further development of the GFS.

1. Introduction

The U.S. Department of Energy established the Atmospheric Radiation Measurement (ARM) Program in the early 1990s to help resolve uncertainties related to the role of clouds and their influence on radiative processes in the atmosphere (Stokes and Schwartz 1994; Ackerman and Stokes 2003). The ultimate goal of this program is to improve parameterizations of clouds and radiation in global atmospheric general circulation models (GCMs), and hence to enhance the capability of GCMs to simulate present climate and to predict future climate changes. Intensive and long-term measurements of surface and atmospheric quantities have been

Corresponding author address: Fanglin Yang, EMC/NCEP, 5200 Auth Rd., Camp Springs, MD 20724.
E-mail: fanglin.yang@noaa.gov

carried out at the ARM surface sites in the U.S. southern Great Plains (SGP), the northern slope of Alaska, and in the tropical western Pacific. 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; Luo et al. 2003; Lenderink et al. 2004; Luo et al. 2005).

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; Luo et al. 2005). ARM provides certain unique products not available from observations either at conventional meteorological stations or from satellites. Mace et al. (1998) compared the vertical distribution of clouds and precipitation in the European Centre for Medium-Range Weather Forecasts (ECMWF) forecasts in the winter of 1997 with data from a millimeter-wave radar operated at the ARM SGP site, and found that the forecast model tended to predict the onset of deep cloud events too soon. Morcrette (2002) assessed the cloud and radiation fields for the first 36 h of operational ECMWF forecasts in April–May 1999 using ARM observations at the SGP site. By taking advantage of the comprehensive ARM observations at high temporal and vertical resolutions the author was able to identify an overestimate of surface downward solar flux in the forecasts and to attribute the source of this bias to insufficient clear-sky gaseous absorptions. Hinkelman et al. (1999) evaluated the National Centers for Environmental Prediction (NCEP) regional Eta model forecasts against ARM observations at the SGP site using observations for the first half of 1997. They also found an overestimate of the surface downward solar flux up to 50 W m^{-2} and attributed half of this bias to insufficient extinction of shortwave radiation by water vapor and aerosols. Luo et al. (2005) used ARM observations made at the SGP site in June and July 1997 and cloud-resolving model simulations to evaluate the representation of cirrus clouds in a single-column version of the NCEP Global Forecast System (GFS). These studies demonstrated that, in addition to their usage in climate models, ARM observations can also be used for ascertaining the quality of operational weather forecasts and for improving forecast model physical parameterizations.

In the past, GCM and NWP developers often resorted to different strategies for developing and evaluating parameterizations of model physical processes. When testing a new parameterization, GCM developers usually relied on multiyear climate simulations to allow comparison of model mean statistics with observations. The success of this strategy was sometimes hindered by limitations such as insufficient observations, cancellation of simulation errors, and strong couplings between different physical processes. At NWP centers, developers relied more on case studies of short to medium-range weather forecasts of extreme events to evaluate 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 more years to generate a large number of samples that include different weather regimes. The forecasts are then assessed against satellite and ground observations. This approach was built upon the premise that the large-scale dynamical state of the GCM remains close to the observed state in the short-term forecasts. If this is the case, systematic forecast biases can then be attributed predominantly to parameterization deficiencies (Phillips et al. 2004).

In the present study, we evaluate NCEP GFS short-term forecasts (6–30 h) at the ARM SGP site. We take an approach that emphasizes the overall statistical behavior of the forecast model for all of 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 improvements. Since early 2001, NCEP has been processing the operational GFS forecasts to produce column output at locations corresponding to all ARM permanent sites. Until now, the archived output has not been systematically evaluated or used for model development. As noted earlier, Hinkelman et al. (1999) used a few months of ARM observations to evaluate the regional Eta forecast model. The GFS does not consist of the same physics and dynamics packages as does the Eta model. This study makes use of the comprehensive and continuous ARM observations to assess the GFS forecasts at the ARM SGP site for the years 2001–04. The focus is primarily on clouds, air temperature, and surface energy fluxes.

In section 2, we first introduce the current NCEP GFS model, describe in some detail the model's physical parameterizations relevant to this study, and then present the ARM observations and NCEP forecasts. Strategies for comparing the model with observations at different temporal and spatial scales are explained. The evaluation of surface energy fluxes and surface air temperature is given in section 3 and the evaluation of clouds is in section 4. In section 5, we describe a set of

single-column model (SCM) experiments that were carried out to understand the sources of certain model errors and how some of the different errors were intrinsically linked to each other. Section 6 summarizes this study.

2. Model, observations, and evaluation strategy

a. Description of the NCEP GFS

The NCEP GFS is a global spectral numerical model based on the primitive dynamical equations that includes a suite of parameterizations for atmospheric physics (e.g., Sela 1980; Kanamitsu 1989; Kalnay et al. 1990). The model has been under constant development and evaluation (e.g., Moorthi et al. 2001). Here we describe some of the model features and the major changes made to the model from 2001 to 2004 that are relevant to the present investigation.

The GFS uses spectral triangular truncation in the horizontal and a sigma coordinate in the vertical that extends from the earth's surface to the top of the atmosphere. A prognostic scheme for cloud condensate (Moorthi et al. 2001; Zhao and Carr 1997; Sundqvist et al. 1989) was implemented in the T170L42 version of the model on 15 May 2001. The model was further upgraded on 29 October 2002 from T170L42 to T254L64 (i.e., from a horizontal grid size of about 75 to 55 km). The time steps for computing dynamics and physics are 7.5 min in T170L42 and 5 min in T254L64, except for the full calculation of longwave radiation that is done once every 3 h and shortwave radiation done once every hour. Corrections are made at every time step to adjust for the diurnal variations in the shortwave fluxes through the atmosphere and in the upward longwave fluxes from the surface.

The shortwave radiation parameterization is based on Chou and Suarez (1999) and was modified by Hou et al. (2002) for the GFS. It contains eight spectral bands in the ultraviolet and visible region and one spectral band in the near-infrared region. It includes absorptions by ozone, water vapor, carbon dioxide, and oxygen. A random-maximum cloud overlapping is assumed for radiative transfer calculations in the operational GFS. Cloud optical depth is parameterized as a function of the predicted cloud condensate path and the effective radius of cloud particles (r_e). Cloud particle single-scattering albedo and asymmetry factors are functions of r_e. For water droplets, r_e is fixed at 10 μm over the ocean, and specified as $r_e = \min[\max(5 - 0.25T_c, 5), 10]$ μm over land, where T_c is temperature in degrees Celsius. For ice particles, r_e is an empirical function of ice water content and temperature that follows Heymsfield and McFarquhar (1996). The radiative effects of rain and snow are not included in the operational GFS, but the direct radiative effect of atmospheric aerosols is included. The surface albedo over land varies with the surface type, solar spectral band, and season, and is further adjusted by a solar zenith-angle-dependent factor for the direct solar beam. When the ground has snow cover the grid-mean surface albedo is first computed separately for snow-free and snow-covered areas, and then combined using a snow-cover fraction that depends on the surface roughness and snow depth. Snow albedo depends on the solar zenith angle (Briegleb 1992).

A major change was made in longwave radiation on 28 August 2003. The Geophysical Fluid Dynamics Laboratory (GFDL) model (Schwarzkopf and Fels 1991) was replaced by the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997). The RRTM computes longwave absorption and emission by water vapor, carbon dioxide, ozone, cloud particles, and various trace gases including N_2O, CH_4, O_2, and four types of halocarbons [chlorofluorocarbons (CFCs)]. Aerosol effects are not included. For consistency with the earlier GFDL module, no trace gases are included in the RRTM for the GFS forecasts. The RRTM uses a correlated-k distribution method and a transmittance lookup table that is linearly scaled by optical depth to achieve high accuracy and efficiency. The algorithm contains 140 unevenly distributed intervals in 16 spectral bands. It employs the Clough–Kneizys–Davies (CKD_2.4) continuum model (Clough et al. 1992) to compute absorption by water vapor at the continuum band. Longwave 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 parameterized by the Simplified Arakawa–Schubert scheme (SAS; Pan and Wu 1995; Grell 1993; Arakawa and Schubert 1974). SAS considers only the deepest cloud type instead of a spectrum of clouds as in the original Arakawa–Schubert scheme. However, we have made a change in the cloud-top selection algorithm so that the top is now selected randomly to be a layer between the neutral buoyancy level (the maximum cloud top) and the level of the equivalent potential temperature minimum. Convection occurs when the cloud work function exceeds a threshold. Cloud mass flux is determined using a quasi-equilibrium assumption based on this threshold. The cloud model also incorporates a saturated downdraft as well as the evaporation of convective precipitation. Entrainment into the updraft and detrainment from the downdraft below the cloud base are included. Water substance detrained at the cloud top is separated into

condensate and vapor, with the former being treated as a source of prognostic cloud condensate.

The parameterization of shallow convection follows Tiedtke's (1983) diffusion scheme. Shallow convection occurs when convective instability exists without deep convection. The cloud-base level is the lifting condensation level. Enhanced vertical mixing is invoked between the cloud-top and cloud-base levels with a fixed profile of vertical diffusivity. Shallow convection does not produce clouds and has no direct radiative effects.

The GFS scheme for large-scale condensation and precipitation is based on Zhao and Carr (1997). Cloud condensate is a prognostic variable. Convective detrainment and grid-scale condensation are the source terms, and evaporation and grid-scale precipitation are the sink terms for cloud condensate. Evaporation of rain in unsaturated layers below the level of condensation is allowed. The cloud fraction used by the large-scale condensation scheme is computed following Sundqvist et al. (1989). For the radiative transfer calculations, however, a different diagnostic scheme (Moorthi et al. 2001) based on Xu and Randall (1996) is used to compute the cloud fraction. This scheme takes into account cloud condensate from both cumulus convection and large-scale condensation.

In the planetary boundary layer (PBL), a nonlocal scheme (Troen and Mahrt 1986; Hong and Pan 1996) is used to compute turbulent transport. The PBL height is diagnostically determined using a bulk Richardson number approach. The turbulent diffusivity profile is specified as a cubic function of the PBL height and scaled by parameters derived from similarity requirements.

The Monin–Obukhov similarity–profile relationships are used to compute surface wind stress and sensible and latent heat fluxes. The formulation is based on Miyakoda and Sirutis (1986) but has been modified for very stable and very unstable situations. Land surface evaporation includes direct evaporation from soil and canopy, as well as transpiration (Pan and Mahrt 1987). A two-layer soil model computes soil temperature and soil volumetric water content at depths of 0.1 and 2.0 m (Pan and Mahrt 1987). The deep-soil temperature at 4-m depth is specified. The heat capacity, thermal and hydraulic diffusivity, and hydraulic conductivity coefficients are strong functions of soil moisture content. The vegetation canopy is allowed to intercept precipitation, which can then evaporate. Runoff from the surface and drainage from the bottom soil layer are also calculated.

b. GFS model output for ARM sites

Since 2001, the NCEP GFS forecasts have been processed to produce single-column profiles at locations corresponding to ARM ground observation sites, including the Southern Great Plains 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. Standard model variables have been extracted and archived, including atmospheric profiles and surface fluxes, at 3-h forecast intervals out to 48 h. The model was initialized each day at 0000 and 1200 UTC. This investigation is focused on the 0000 UTC runs, and more specifically on the 6–30-h forecasts for the diagnosis of diurnal variations. Analyses of the forecasts beyond 30 h led to the same conclusions. Results from the 1200 UTC runs were also similar to those from the 0000 UTC runs.

Energy fluxes at the earth's surface and top of the atmosphere, including radiation, latent and sensible heat, and wind stresses were saved as 3-h averages. Surface rainfall and snowfall were saved as 3-h accumulations. State variables such as the surface and atmospheric temperatures, cloud fraction, cloud condensate, and specific humidity were saved as instantaneous values at the end of each 3-h forecast interval. The profiles were saved on the model's sigma levels and then projected onto standard isobaric layers (1000–25 hPa, at 25-hPa intervals) and standard atmospheric heights (surface to 20 km, at 250-m intervals) for comparison with observations. The NCEP GFS forecasts for all of the ARM sites have also been converted to the network Common Data Form (netCDF) format and archived at the ARM External Data Center at Brookhaven National Laboratory (more information available online at http://www.arm.gov/xds/static/gfs.stm).

c. ARM observations at the Southern Great Plains site

The observations used for this study (see Table 1) were obtained from the ARM Climate Research Facility Data Archive (see online at http://www.arm.gov/data/). The comparisons between GFS forecasts and ARM observations were made primarily at the Central Facility (CF-1). Observations at the SGP extended facilities (EFs) were also used for a scale-dependence test (see section 2d). Figure 1 depicts the ARM SGP observation network. Note that EF-13 is collocated with the CF. Measurements from both EF-13 and CF-1 were used to represent observations at the CF site for model evaluation.

Because the ARM observations listed in Table 1 have different temporal frequencies, they were processed to match the 3-hourly GFS outputs. For flux variables, all measurements in a 3-h period were used to derive 3-hourly means. For state variables, measure-

TABLE 1. ARM data streams and variables used for forecast evaluation, where EF refers to ARM Extended Facility sites.

Data stream and identifier in ARM archive	No. of sites	Measurements used for this study	Frequency
SMOS, sgp30smosExx.b1	16 EF	2-m surface temperature; precipitation	30 min
EBBR, sgp30ebbrExx.b1	14 EF	Sensible and latent heat fluxes; ground soil heat flux	30 min
Best-Estimate Radiative Flux, sgpbeflux1longC1.c1	CF-1	Surface downward and upward solar and longwave fluxes	1 min
SIRS, sgpsirsExx.b1	22 EF and CF-1	Surface downward solar and longwave fluxes	1 min
ARSCL, sgparsclbnd1clothC1.c1	CF-1	Cloud-base and cloud-top heights	10 s
"MWR PROF" VAP, sgpmwrprofC1.c1	CF-1	Vertical profiles of water vapor, temperature	1 h

ments made at the time closest to the end of a 3-h period were chosen to represent the values for the entire 3-h period.

d. Evaluation strategy and scale-dependence tests

One of the purposes of this study is to evaluate the overall performance of the GFS against ARM observations for the 4 yr from 2001 to 2004. We will compare time series, diurnal cycles, and seasonal and annual means of the GFS forecasts with available ARM observations to detect model biases, to find if the model performance has been improving as the model is upgraded, and to give recommendations for model bias corrections and for future model development.

The GFS single-column output represents mean atmospheric conditions over a model grid area, roughly 70×70 km^2 in size for the T170 version and 55×55 km^2 for the T254 version. However, the ARM observations were made at single points. How can we properly compare the model grid values with single-point observations? Ideally, model results should be compared to the means of observations over an area comparable to the model grid size; however, not all observations are available at all the SGP facilities (Fig. 1). Is it meaningful to compare model grid values with observations made at the single CF site? Barnett et al. (1998) investigated the temporal and spatial scales of surface shortwave fluxes measured at the ARM SGP sites and the Oklahoma MESONET. They showed that, for 3-hourly averages, the temporal correlation between the fluxes measured at the SGP CF and the mean over an area 60 km $\times$ 60 km is about 0.7 for cloudy-sky conditions and 0.8 for clear-sky conditions. Long et al. (2002) examined total cloud amount over the SGP network measured during the ARM 1997 and 2000 cloud intensive operational periods. They found that, in general, the representativeness of the SGP CF cloud cover decreases as the distance used for computing average cloud cover increases. On a daily basis, the average distance is 125–150 km for a correlation of 0.5–0.6 between the CF cloud amount and the areal mean, and 75–100 km for a correlation of 0.8–0.9. These studies suggest that for an area of about the current GFS grid size (approximately 60 km $\times$ 60 km at the SGP), the ARM observations at the CF can be used to represent the grid-size-mean values of solar radiation and clouds. We extend the studies of Barnett et al. (1998) and Long et al. (2002) by performing scale-dependence tests for a few more variables relevant to this study and by focusing on the biases instead of the temporal correlations.

The ARM SGP Cloud and Radiation Test bed (CART) observation network covers an area of approximately 300×300 km^2. It 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 within the CART centered at the central facility that has approximately the same size as the T254 model grid 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. The CF site is covered by grazed pasture and wheat. Of the eight sites included in OB2, six are also covered by grazed pasture and wheat and the other two by alfalfa and native prairie. OB3 has 33 sites in total; among them only 13 sites are covered by grazed pasture and/or wheat, while the others are covered by grass, rangeland, alfalfa, forest, native prairie, and ungrazed pasture. The surface type over the domain OB3 is rather inhomogeneous. Next, domain averages over OB2 and OB3 were computed for surface air temperature and precipitation from the Surface Meteorological Observation System (SMOS) instruments, surface downward and upward solar and longwave radiative fluxes from the Solar Infrared Radiation Station (SIRS) instruments, and latent and sensible heat fluxes from the Energy Balance Bowen Ratio (EBBR) instruments (Table 1).

For brevity, we show in Fig. 2 only the forecast biases of surface downward solar and longwave fluxes for the 6–30-h forecasts made each day in 2003 relative to the observation OB1, the differences between observations

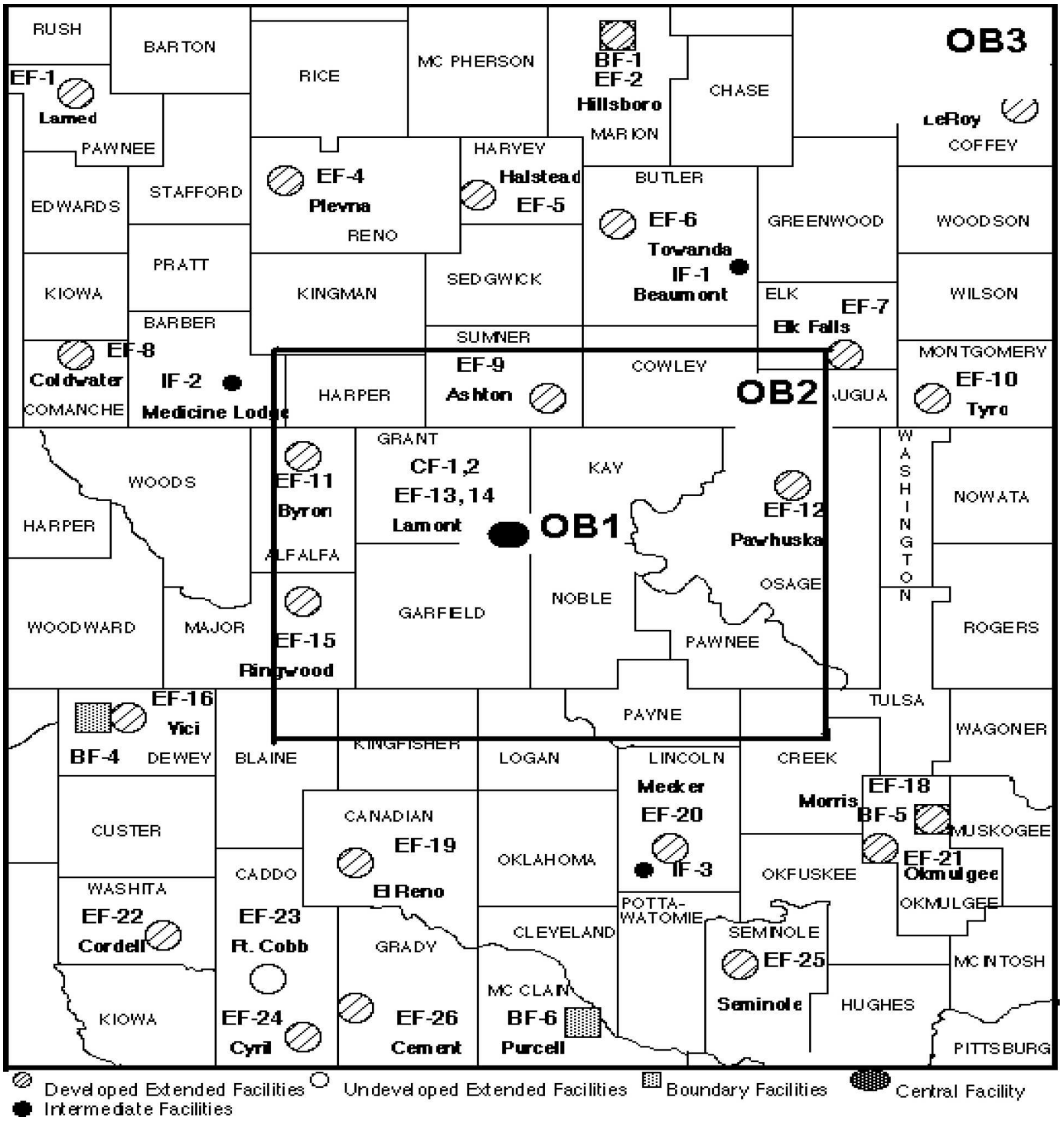

FIG. 1. Facilities at the ARM SGP site (available online at http://www.arm.gov/). The entire domain is about 300 km × 300 km in size (OB3). For the scale-dependence test described in section 2d, a subdomain (OB2) comparable in size to the grid size of the NCEP T254 (~55 km) forecast model is chosen. Observations averaged over OB2 are compared with the single-point measurements made at the central facility (OB1), and with the observations averaged over the entire SGP site (OB3).

OB2 and OB1, and the differences between observations OB3 and OB1. Tests for other variables produce similar results. Figure 2 shows that the forecast biases (FCST − OB1) are always larger than the differences between the observations for OB2 and OB1. However, the differences between the observations for OB3 and OB1 are at least as large as the forecast biases. The different surface conditions over OB1 and OB3 are responsible for the differences in the observed surface fluxes. These scale-dependence tests suggest that comparing GFS forecasts with either the single-point observations at OB1 or with the means over OB2 will lead to similar conclusions. The results are consistent with the findings of Barnett et al. (1998) and Long et al. (2002). In the reminder of this paper, all comparisons are made between model forecasts and observations at the CF site.

3. Surface energy fluxes and surface air temperature

In this section, we assess the overall performance of the NCEP GFS at the SGP site for 2001–04. All presentations in this study are based on the forecasts for

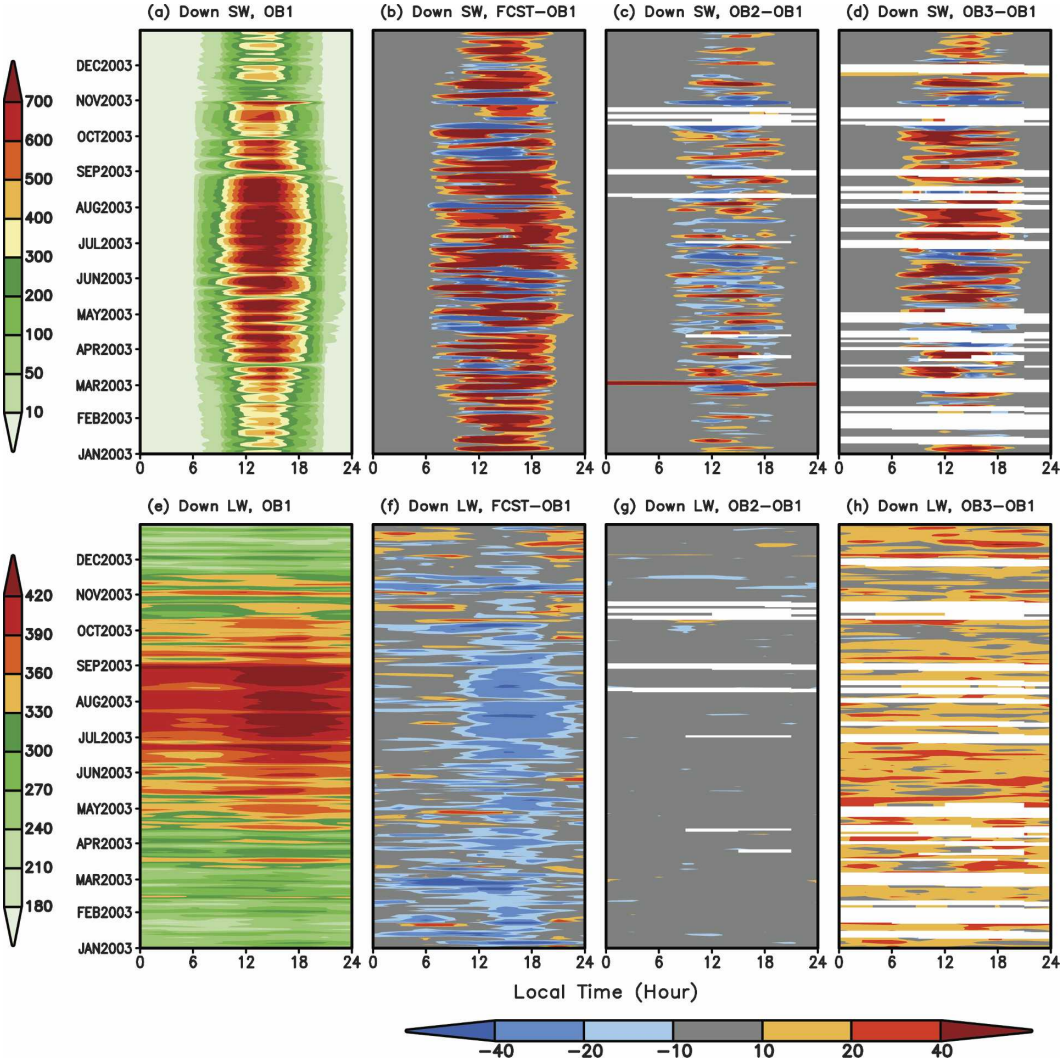

FIG. 2. A scale-dependence test for (a)–(d) surface downward shortwave and (e)–(h) longwave fluxes at the ARM SGP site. (a), (e) ARM observations over OB1, (b), (f) GFS forecast biases relative to the observations over OB1, (c), (g) the differences between the observations over OB2 and OB1, and (d), (h) the differences between the observations over OB3 and OB1. Missing values are masked by white stripes.

the 6–30-h period, which corresponds to a full diurnal cycle (midnight to midnight local time). Shown in Fig. 3 are the ARM observations and the differences between GFS forecasts and observations for the surface downward and upward shortwave fluxes and downward longwave flux. The observations are 3-hourly means and were derived from the Valued-Added Product (VAP) Best-Estimate Radiative Flux (Long 2002; Shi and Long 2002) at 1-min temporal resolution (Table 1). Comparisons between the forecasts and the observations indicate that, qualitatively, the model simulated the observed diurnal and seasonal variations well in these fluxes. However, quantitative differences do exist (Fig. 3). The model overestimated the surface downward shortwave flux in all seasons. The bias reached about 200 W m^{-2} at 1500 LST (local standard time) in the summer. The bias for the surface upward (reflected) shortwave flux was relatively small and sometimes even had the opposite sign to the bias of the downward flux. This discrepancy implies a potential problem in the surface albedo prescribed in the model, which will be further discussed. The model underestimated the surface downward longwave flux during the day and overestimated it at night. On 28 August 2003 the longwave radiative transfer module of the GFS was switched from the GFDL scheme to the Mlawer et al. (1997) RRTM. The bias was noticeably reduced during the day after the switch, however, the

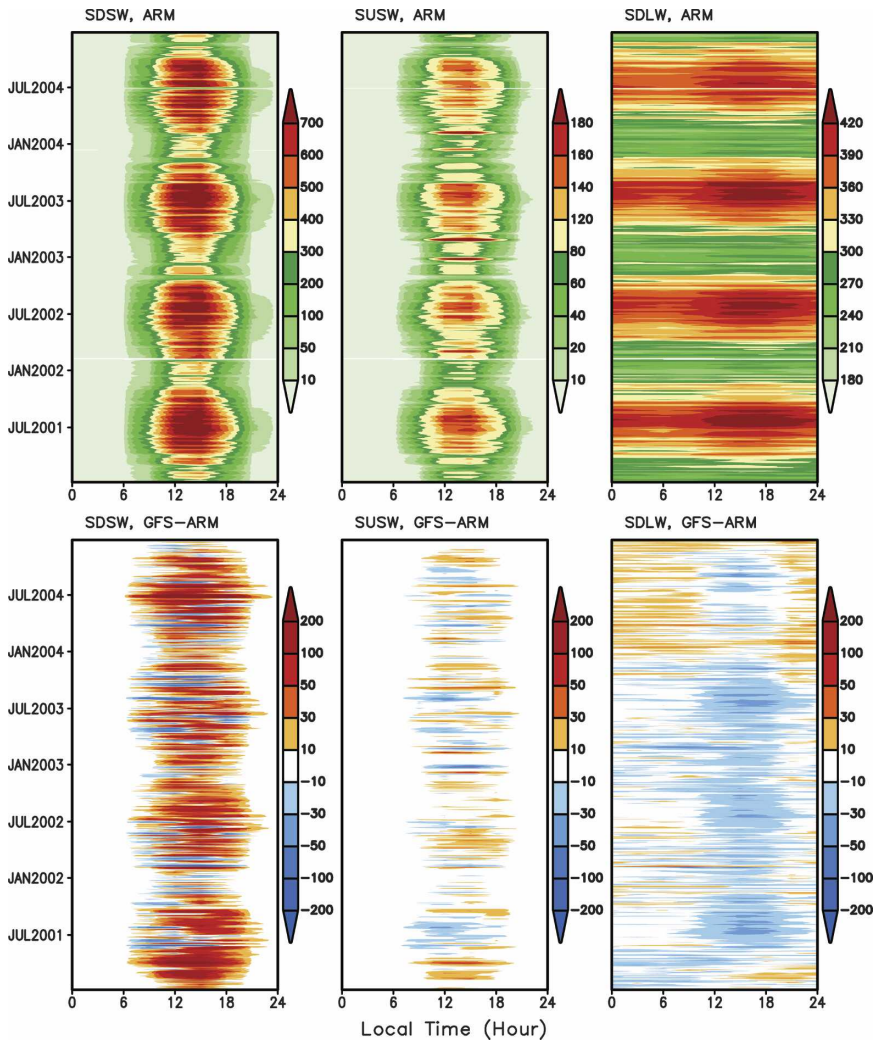

FIG. 3. The 3-hourly mean SDSW, SUSW, and SDLW at the ARM SGP CF/E13 site. (top) ARM observations and (bottom) the differences between the GFS forecasts and ARM observations.

bias became larger at night. In section 5 we describe a set of single-column model experiments that we performed to better understand the source of the longwave flux errors.

Figure 4 further compares the forecasts of surface latent and sensible heat fluxes with ARM observations. The observations were derived from the 30-min EBBR measurements (Wesely et al. 1995) at the E13 site (Fig. 1), which is collocated with the C1 facility. We are aware of the fact that a VAP EBBR has been produced by the ARM program to correct some of the spikes found in the measured time series, which were caused by dew, frost, or condensation on the EBBR radiometer domes (more information available online at http://science.arm.gov/vaps/baebbr.stm). However, that product was only available before June 2003 and had many missing values, so we used the original EBBR product. Cross validation indicates that the differences between the original and the VAP are negligibly small compared to the magnitude of the model biases (figures not shown). We screened the original EBBR data to exclude some of the visually detectable bad values. However, as illustrated in Fig. 4, the remaining EBBR measurements of latent and sensible heat fluxes still exhibit some errors before April 2002.

In Fig. 4 upward fluxes are positive. The observed latent heat fluxes were always positive. The sensible heat fluxes were primarily negative at night because of strong surface cooling and positive during the day because of surface heating by solar radiation. The GFS persistently overestimated the latent heat fluxes in both the daytime and nighttime hours in all seasons. The

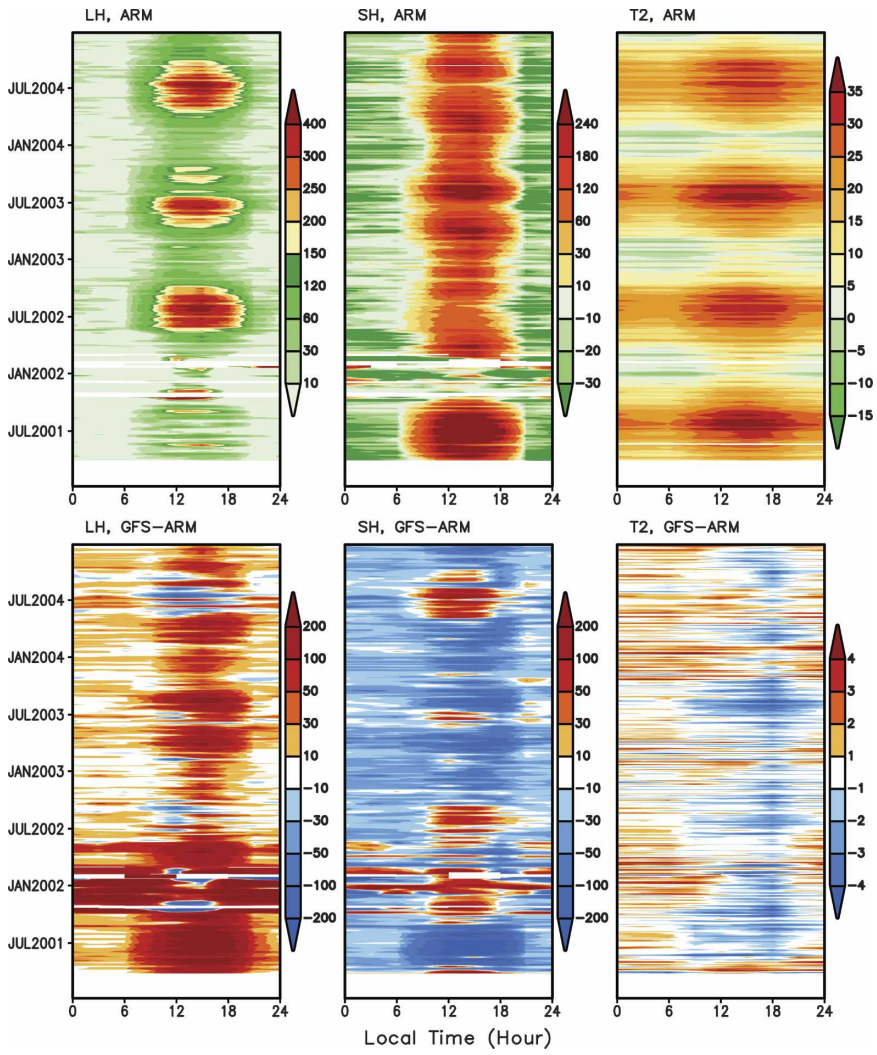

FIG. 4. Same as in Fig. 3, but for latent heat (LH) flux, sensible heat (SH) flux, and the surface 2-m air temperature.

daytime bias reached up to 200 W m^{-2}. The model tended to underestimate the positive sensible heat flux during the day and overestimate the negative flux at night. Figure 4 also includes the 2-m air temperatures. The observations were obtained from the ARM SMOS (Table 1). The model 2-m air temperatures showed predominantly cold biases during the day and warm biases at night.

Up to now we have focused on qualitative measures of the GFS. The forecast biases presented in Figs. 3 and 4 display strong diurnal variations. For quantitative measures, in Fig. 5 we present the mean diurnal cycles of fluxes and surface air temperature averaged over the two years from 1 December 2002 to 30 November 2004, during which all observations were available and were of good quality. The model captured the phases of the observed diurnal variations in all surface energy fluxes and air temperatures, except for the downward longwave radiation. The observed longwave fluxes peaked at 1500 to 1800 LST, while the forecasts peaked later at about 2100 LST. Single-column model experiments (see section 5) indicate that this discrepancy was caused by an error in a scaling factor that depends on the ground skin temperature.

For all other quantities, both the observations and the forecasts reached their maxima at about 1500 LST and minima at about 0600 LST. However, for certain variables large differences in magnitudes exist. At 1500 LST, the model overestimated the surface downward solar flux by 44 W m^{-2} and the latent heat flux by 76 W m^{-2} and underestimated the surface downward longwave flux by 14 W m^{-2}. The model underestimated

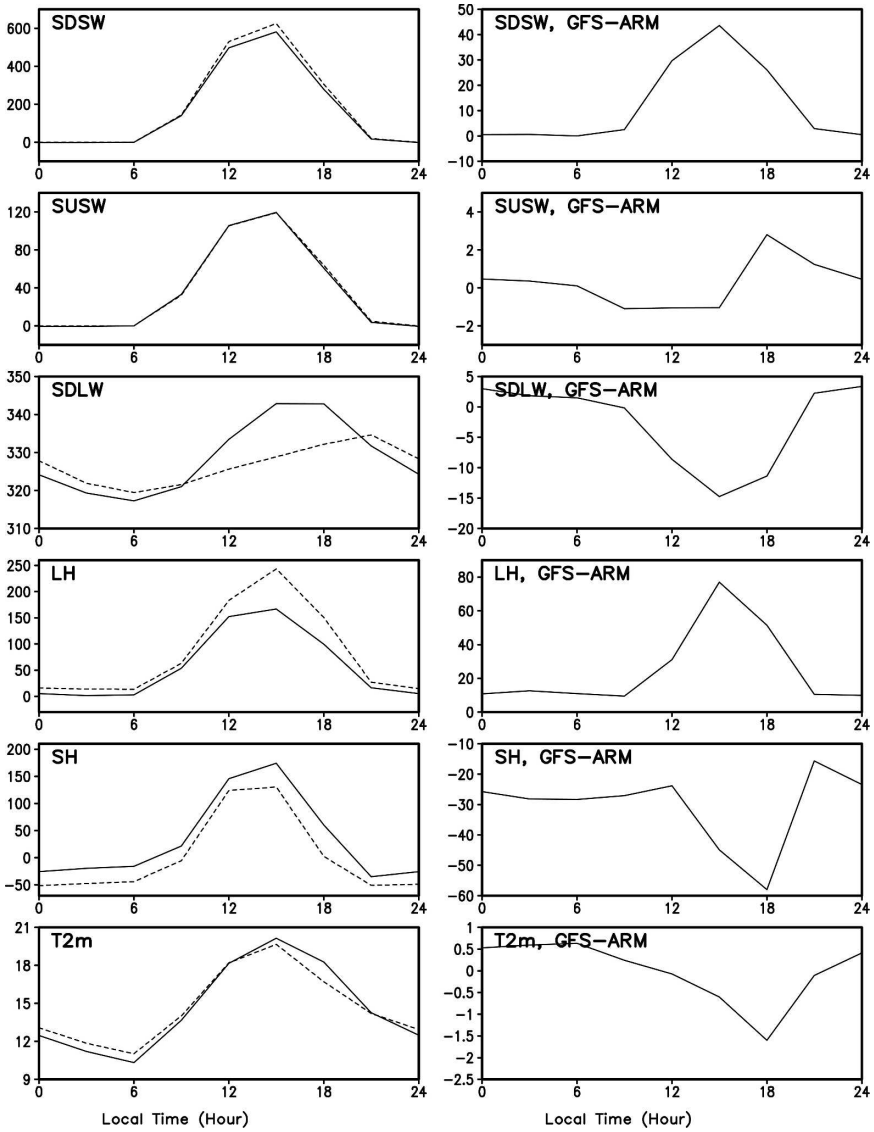

FIG. 5. Mean diurnal cycles of surface energy fluxes (W m^{-2}) and surface air temperature (°C) over the period 1 Dec 2002–30 Nov 2004. (left) Solid lines are for ARM observations and dashed lines are for the GFS forecasts. (right) Differences between the GFS forecasts and ARM observations.

the daytime positive sensible heat flux by 44 W m^{-2} at 1500 LST, and overestimated the nighttime negative flux by about 29 W m^{-2} at 0600 LST. The forecasts of 2-m air temperature were slightly warmer than those observed at night, and colder during the day, with a maximum error of −1.6°C at 1800 LST. It is interesting to note that even though the forecast bias in surface downward solar flux was large the bias in surface reflected solar flux was very small, less than 3 W m^{-2} at all hours of the day. Further analysis found that the surface albedo prescribed in the GFS was smaller than the observed albedo (see Fig. 6).

The above analysis suggests that there is a cancellation of errors among the forecast surface energy terms. The forecasts at 1500 LST are an example of this. Table 2 lists the average surface energy fluxes at 1500 LST shown in Fig. 5 for both the ARM observations and GFS forecasts. To complete the surface energy budget, we also included in Table 2 the surface upward longwave flux and the ground soil heat flux. The observational resources for these fluxes are given in Table 1. The net heat fluxes in Table 2 are very small for both the observations and forecasts. This indicates balanced surface energy fluxes in both systems. However, the

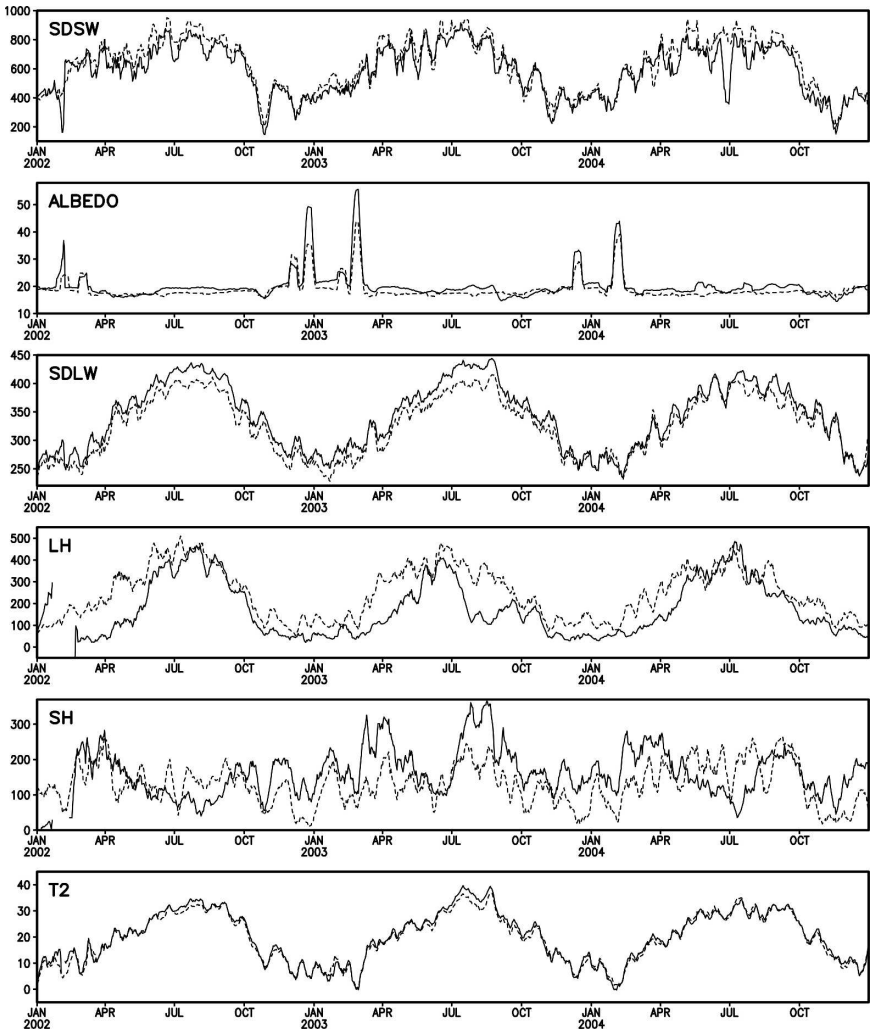

FIG. 6. Time series at 1500 LST extracted from Figs. 3 and 4, except for surface albedo. Solid lines are for ARM observations, and dashed lines for the GFS forecasts. A 10-day running-mean filter was applied to all time series.

balance in the GFS forecasts was achieved by a cancellation of errors. The forecast model suffered from excessive surface cooling, up to 76 W m^{-2}, caused by evaporation (latent heat flux). This large bias was almost compensated for by a 44 W m^{-2} overestimate of downward solar flux and a 44 W m^{-2} underestimate of upward sensible heat flux. The model also underestimated the surface upward and downward longwave fluxes by 14 and 13 W m^{-2}, respectively.

To further examine the seasonal and annual varia-

TABLE 2. Surface energy fluxes (W m^{-2}) at the ARM SGP CF site at 2100 UTC (1500 LST) averaged from 1 Dec 2002 to 30 Nov 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 surface energy fluxes. Downward fluxes are positive and upward fluxes are negative.

	SDSW	SUSW	SDLW	SULW	LH	SH	GH	NET
ARM	581	−120	343	−437	−167	−174	−28	−2
GFS	625	−119	329	−424	−243	−130	−34	4
GFS − ARM	44	1	−14	13	−76	44	−6	6

tions in the energy fluxes, in Fig. 6 we plotted the time series in Figs. 3 and 4 at 1500 LST from 1 January 2002 to 31 December 2004. Instead of showing the surface upward solar flux as in Fig. 3, we included the observed and forecast surface albedo in Fig. 6. The observed albedo was derived from the 1-min VAP Best-Estimate Radiative Flux (Table 1). The SGP CF is covered by pasture and wheat most of the year. Surface albedo is usually in the vicinity of 0.2 if there is no snow on the ground. For the forecasts, surface albedo was determined by the model's surface albedo parameterization at each time step and averaged over a 3-h period. The model tended to underestimate the surface albedo by about 0.02 in absolute values for most of the year. As a result, the forecast surface upward solar fluxes matched the observations, even though the model largely overestimated the surface downward solar fluxes (see Fig. 3 and Table 2).

There are several spikes in the time series of surface albedo in Fig. 6 in the winter and early spring in both the observations and forecasts. These spikes occurred on the days with snow on the ground. Not surprisingly, the GFS was able to portray the occurrence of snow events (figures not shown) since the observed snow cover was assimilated into the initial conditions of the forecast system. However, the surface albedo derived from the model's surface parameterization was always smaller than the albedo observed over a snow-covered surface. In the GFS, the surface albedo over a snow-covered surface depends on snow depth, surface temperature, solar zenith angle, and surface roughness (Hou et al. 2002). Further investigations are required to determine the exact cause of the albedo bias.

In Fig. 6, the forecast surface downward longwave flux matched the observed flux much better after August 2003 than before. In August 2003, the longwave radiative transfer scheme in the GFS was switched from the GFDL module to the RRTM module (Mlawer et al. 1997). In section 5 we describe the SCM experiments we made to assess how the two modules behave with identical sky conditions. The forecasts overestimated latent heat flux during the day (Fig. 5), and the biases were larger in the spring and fall seasons than in the summer (Fig. 6). Additional model experiments and observational analyses are required to determine the cause of the biases. The observed sensible heat fluxes showed large seasonal and annual variations. These fluxes were much larger in the spring and fall than in the winter and summer, and were larger in 2003 than in the other years. These fluxes in the GFS forecasts did not vary much with the season and year. Overall, the model underestimated the surface-to-atmosphere sensible heat flux in the afternoon.

4. Cloud fraction

a. Definition of cloud fraction

Clothiaux et al. (2000, 2001) created the Active Remotely Sensed Clouds Locations (ARSCL) VAP. This product combines measurements from the millimeter cloud radar, 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 cloud-base and cloud-top heights for each group of contiguous clouds in the vertical column at a 10-s temporal resolution. Jakob et al. (2004) proposed a method that compares model clouds with pointwise observations based on probabilistic distributions. They used the ARSCL product to evaluate the total cloud cover in the summer of 1997 simulated by a cloud-resolving model and the ECMWF operational model. Here we take the conventional time-averaging approach to examine the scale dependence of clouds at all layers of the troposphere.

From the GFS forecasts, cloud fractions were saved for the model's sigma layers as instantaneous values at the end of each of the 3-h forecast intervals. However, the time step for the GFS physics was 7.5 min in the T170 version before 29 October 2002, and 5 min in the T254 version thereafter, so the forecasts were undersampled.

There is an issue of incompatibility in space and time between the observations and the model forecasts. Cloud fraction from the model represents the percentage of a model grid area that is covered by clouds within each physical time step. For single-point observations, cloud fraction is usually defined as cloud occurrence frequency within a given time period (e.g., Lazarus et al. 2000). The problem is how to choose the period. One approach is to set it as the duration of the last call of the GFS physics routines within each 3-h postprocessing period. However, the so-defined cloud fraction (occurrence frequency) is undersampled in space compared to the model output. To compensate for the undersampling in space, one can define a cloud occurrence frequency over an "impact" period of time, which, to a certain degree, represents the dynamical advection of cloud condensate from neighboring points surrounding the observation column. This impact period varies in space and time since the temporal scale of

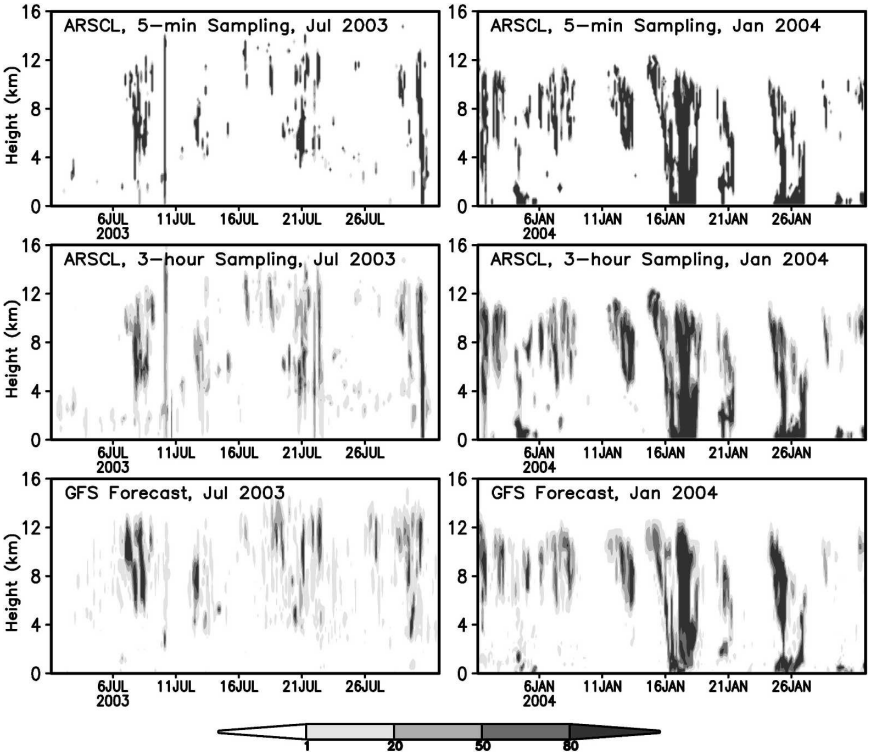

FIG. 7. Vertical distributions of cloud occurrence frequency for ARSCL observations and fractional cloud cover for the GFS forecasts in (left) July 2003 and (right) January 2004. The ARSCL cloud occurrence frequencies were calculated using the 5-min and 3-h sampling methods (see text for details).

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 computed the corresponding cloud fractions (occurrence frequency) for the observations. The purpose is to test the sensitivity of cloud occurrence frequency to the different impact periods.

Before using the observed ARSCL cloud-base and cloud-top heights, which had a 10-s resolution, they were mapped to layers that have a 250-m vertical resolution from the surface to 20 km. Within each layer, a cloud occurrence of one was recorded if clouds were detected and zero if no clouds were detected. Then we divided the 4-yr archive into 3-h subsets. Cloud occurrence frequencies were computed by averaging the 10-s cloud occurrence records over the last 5 min, 7.5 min, 15 min, 1 h, and 3 h of each 3-h subset. In such a way, cloud occurrence frequency derived with an impact period shorter than 3 h used only part of the observations within each 3-h subset.

As a demonstration, in Fig. 7 we show the vertical distributions of the observed cloud occurrence frequency in July 2003 and January 2004 for the two extreme cases, the 5-min and 3-h impact periods. The overall structures of the cloud systems computed from the two methods are similar, except that there is more fractional cloud cover (less than 100%) for the 3-h case than for the 5-min case. We computed and plotted in Figs. 8a,b the total (column integrated) cloud amount from 1 December 2002 to 30 November 2004 obtained from the cloud fraction profiles using a random-overlap assumption. The temporal correlation between the two ARSCL time series is 0.97; however, at times the total cloud amount for the 5-min case, which is undersampled in time, can be quite different from that for the 3-h case. The 2-yr mean total cloud amount is 56.4% for the 5-min case and 61.7% for the 3-h case.

The analysis indicates that cloud occurrence frequency does vary in magnitude with the impact period, although its temporal evolution does not. In practice, we do not know which impact period is the most realistic. 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. In this investigation, for each of the 3-h processing periods the GFS clouds were undersampled in time; the ARM clouds were undersampled in space, and also in time if all data within the 3-h

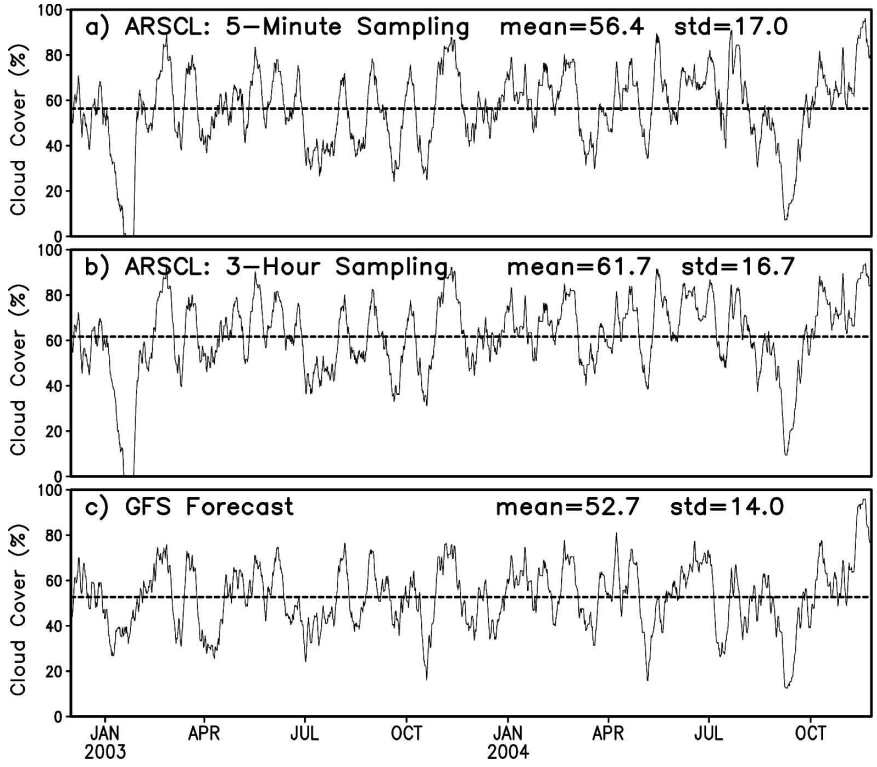

FIG. 8. Total cloud amount from the surface to 16 km derived from the cloud vertical profiles at 250-m intervals based on a random-overlap assumption. All time series were smoothed by a 10-day running-mean filter to exclude high-frequency variations. (a)–(c) The dotted straight line represents mean cloud amount averaged December 2002–November 2004, and the numbers are the mean and standard deviation of the cloud amount. Anomalous correlation between ARSCL cloud amounts from the 5-min and 3-h sampling is 0.97, and the correlations between the GFS forecasts and ARSCL clouds from the 5-min and 3-h sampling are 0.77 and 0.78, respectively.

period were not used for calculating the cloud occurrence frequency.

b. GFS forecasts and ARM observations

To compare with observations (i.e., ARSCL), the archived GFS instantaneous cloud fractions for the sigma layers were projected onto the same 250-m resolution layers as for the observations. In Fig. 7, the forecast cloud fractions at the ARM CF site are compared with the ARSCL cloud occurrence frequency distributions for July 2003 and January 2004. The model captured the observed cloud distributions for the major synoptic events in both the warm and cold month. The forecasts compared more favorably with observations from the 3-h sampling than those from the 5-min sampling. Figure 8c displays the GFS total cloud amount for December 2002–November 2004. The cloud amount was obtained from the forecast cloud fraction profiles by assuming a random overlap. The correlations between the forecasts and the observations sampled at the 5-min and 3-h intervals both reached 0.77; however, the 2-yr-mean total cloud amount from the forecasts is 52.7%, which is about 10% less than the observed sampled at the 3-h intervals.

In the following, we examine diurnal variations in clouds in the vertical between the forecasts and observations. Data in the period from December 2002 to November 2004 were used to derive the seasonal mean diurnal variations shown in Fig. 9. In the ARM observations (Figs. 9a–d), the diurnal cycles were very different in different layers of the troposphere and in different seasons. In the mid- and upper troposphere the diurnal cycle was weak in all seasons except the summer when more clouds were observed at night than during the day. In the lower troposphere and near the boundary layer there were less clouds in summer than other seasons, especially at night. For all seasons, cloud fractions in the boundary layer and lower troposphere were the biggest at about noon and the smallest at about midnight.

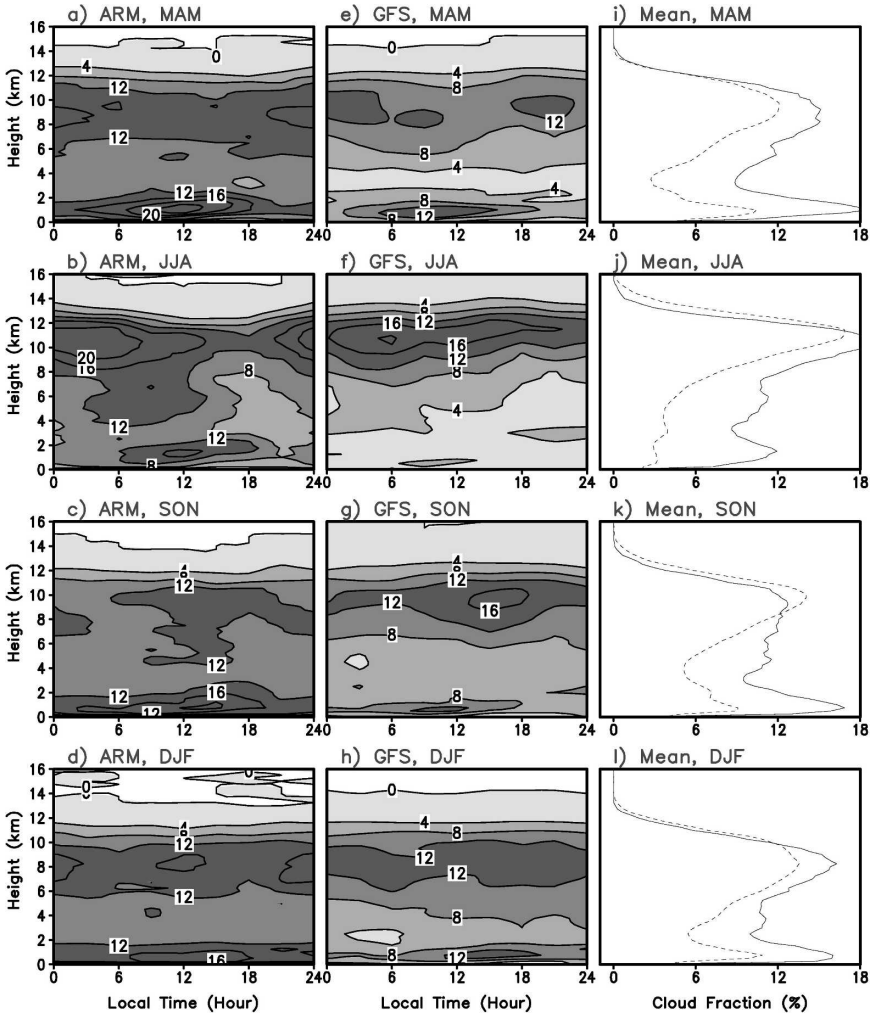

FIG. 9. (a)–(h) Diurnal and vertical distributions of cloud fraction averaged for each season based on 3-hourly mean data December 2002–November 2004. (i)–(l) Mean vertical distributions of cloud fraction averaged over the 24 h in (a)–(h). Dashed lines are for GFS forecasts and solid lines are for ARM observations. The contour intervals (CIs) are 4% in all panels.

Lazarus et al. (2000) found that at the SGP site the major types of the observed low clouds (below 2 km) are stratus, stratocumulus, and cumulus. Stratiform clouds are found in all seasons, but are most prevalent during the winter and most infrequent in the summer. In general, there are slightly more stratiform clouds during the day than at night. In the summer, cumulus dominates and has a strong diurnal cycle, with the peak occurring at about noon. Overall, the distributions of the ARSCL clouds we presented in Fig. 9 are in agreement with Lazarus et al.'s (2000) analyses based on different cloud types.

The GFS (Figs. 9e–h) was able to capture the observed bimodal distribution of clouds in the vertical. There were more clouds in the lower and upper troposphere than in the midtroposphere. The diurnal variation of clouds in the upper troposphere was better simulated than it was in the boundary layer and lower troposphere. The model largely underestimated daytime low clouds in the summer and fall. In the daily averages, the model underestimated clouds in the lower and midtroposphere during all seasons by a few percent (Figs. 9i–l). It is known to us that the GFS has difficulties in simulating shallow convection. Even when shallow convection does occur, the model does not allow the convection to produce clouds. The lack of a strong diurnal cycle of low clouds in the summer and fall at the SGP sites (Figs. 9f,g) was in part caused by the model's deficiency in simulating shallow convection.

To better understand the diurnal and seasonal variations in clouds, we separated the cloud distributions shown in Fig. 9 into conditions with and without pre-

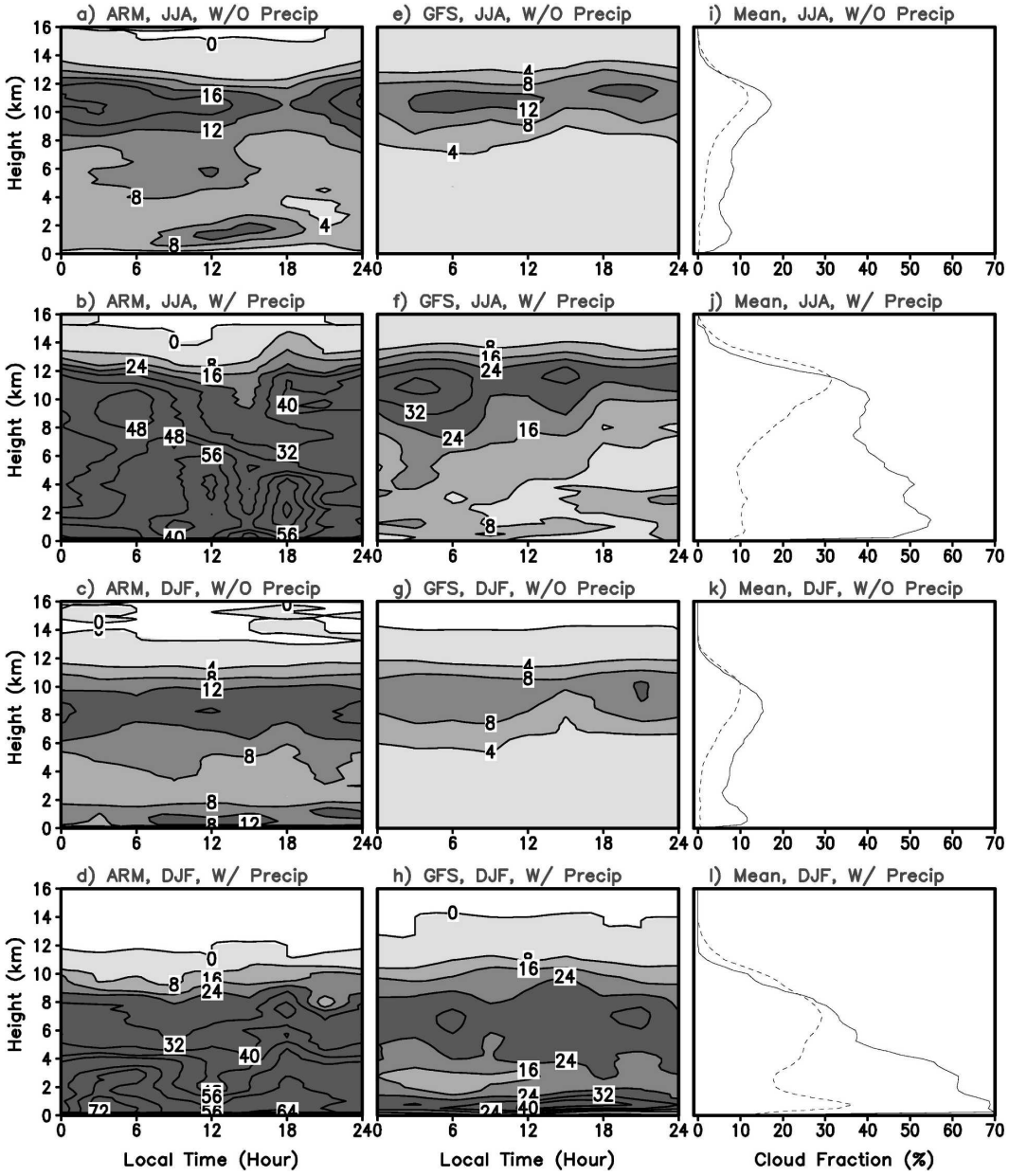

FIG. 10. Diurnal and vertical distributions of clouds in JJA and DJF, same as in Fig. 9 except that the cases with and without precipitation occurring on the ground are separated. (a)–(h) The CIs are 4% for the cases without precipitation and 8% for the cases with precipitation. (i)–(l) Dashed lines are for the GFS forecasts and solid lines are for ARM observations.

cipitation at the surface for both the forecasts and observations in June–August (JJA) and December–February (DJF; Fig. 10). In the ARM observations there were many nonprecipitating low clouds during the day for both seasons. The GFS forecasts captured the nonprecipitating high clouds, but entirely missed the daytime nonprecipitating low clouds (Figs. 10e,g). The deficiency in the model's shallow convection scheme is probably responsible for this bias. When precipitation did occur at the surface, clouds from the forecasts were still underestimated in both seasons, but especially in the summer. The observed summertime penetrative convective clouds occurred in the afternoon and evening (Fig. 10b); however, the forecasts failed to capture such clouds (Fig. 10f). Both the shallow and penetrative convection from the GFS forecasts seem to be less active than those from the observations. On the other hand, it should be pointed out that the ARM

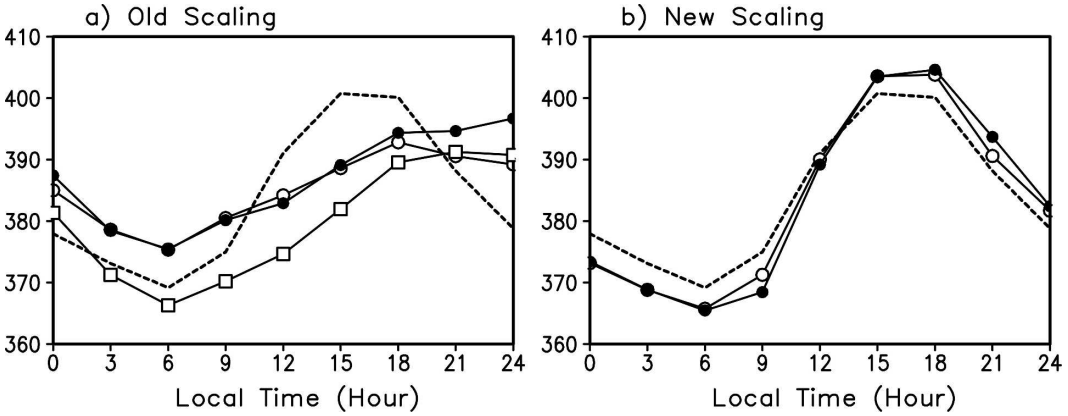

FIG. 11. Diurnal cycles of SDLW (W m^{-2}) in JJA 2004. (a), (b) The dotted lines are for ARM observations. In (a), the line with open circles is for the GFS forecasts, and the lines with filled circles and open squares are for the SCM forecasts with and without clouds. In (a), the SDLW fluxes for both the GFS and SCM forecasts were scaled by a factor that is a function of skin temperature. In (b), the lines with open and filled circles are for the SCM forecasts, for which the SDLW was scaled by a factor that is either a function of the air temperature at the lowest model layer or a function of the mass-weighted mean air temperature at the lowest four model layers.

ARSCL cloud data represent all types of hydrometers including water droplets, ice particles, rain, and snow. In the GFS, only cloud water droplets and ice particles are accounted for in the calculation of cloud fractions, rain and snow are not included. This may partly explain the model errors, especially when precipitation occurs.

5. Further analyses of radiative fluxes

a. Diurnal cycle of surface downward longwave flux

Figure 5 showed that, on average, the observed all-sky surface downward longwave flux (SDLW) peaked at 1500–1800 LST, while the forecast peaked at 1800–2100 LST and its magnitude was about 14 W m^{-2} smaller than the former. To facilitate further investigation of this problem, we performed a set of sensitivity experiments using the NCEP SCM that has no large-scale dynamics but includes all the GFS physical processes (Luo et al. 2005). The SCM was initialized with the surface and atmospheric states, including cloud fraction and cloud water amount, saved from the GFS forecasts. During the time integrations, the SCM was driven by the archived GFS dynamical forcing terms, which include vertical velocity and the time derivatives (tendencies) of surface pressure, temperature, specific humidity, zonal and meridional momentum, and cloud water. These forcing terms were accumulated from the dynamic processes of the GFS forecasts and saved as 3-hourly means. They were linearly interpolated to each SCM time steps. All experiments were initialized at 0000 UTC each day and were integrated forward for 36 h. Results from the last 24 h of the integrations were used to examine the diurnal cycles of the SDLW. For brevity, in Fig. 11 we present the results for JJA 2004. Similar results were obtained for other seasons.

The SCM was first run with the original GFS physics packages. The diurnal cycle of the SDLW from this experiment (the line with filled circles in Fig. 11a) follows closely the one from the GFS forecasts (the line with open circles). This demonstrates the credibility of the SCM and reproduces the problem found in the GFS forecasts. It was shown in section 5 that the GFS poorly simulated the diurnal cycle of clouds in the boundary layer in JJA (Fig. 9f). To test for the possibility that the problem in the diurnal cycle of the SDLW was caused by errors in cloud distributions, we set all cloud fractions to zero in the second SCM experiment. The resulting SDLW diurnal cycle shown in Fig. 11a still does not follow the observed cycle and is similar to the one from the first experiment.

In the GFS the longwave radiative transfer routine is called once every 3 h; however, the actual SDLW [$LW^{\downarrow}(t)$] that drives the land surface model at each model time step is adjusted by the following factor:

$$LW^{\downarrow}(t) = LW^{\downarrow}(t0)\left[\frac{TG(t)}{TG(t0)}\right]^4, \qquad (1)$$

where $TG(t)$ is the ground skin temperature at time t, and $LW^{\downarrow}(t0)$ and $TG(t0)$ are the SDLW and skin temperature, respectively, at the time when the longwave radiative transfer routine is last called. Since the SDLW is mostly determined by the air temperatures near the surface instead of skin temperature, this scaling might

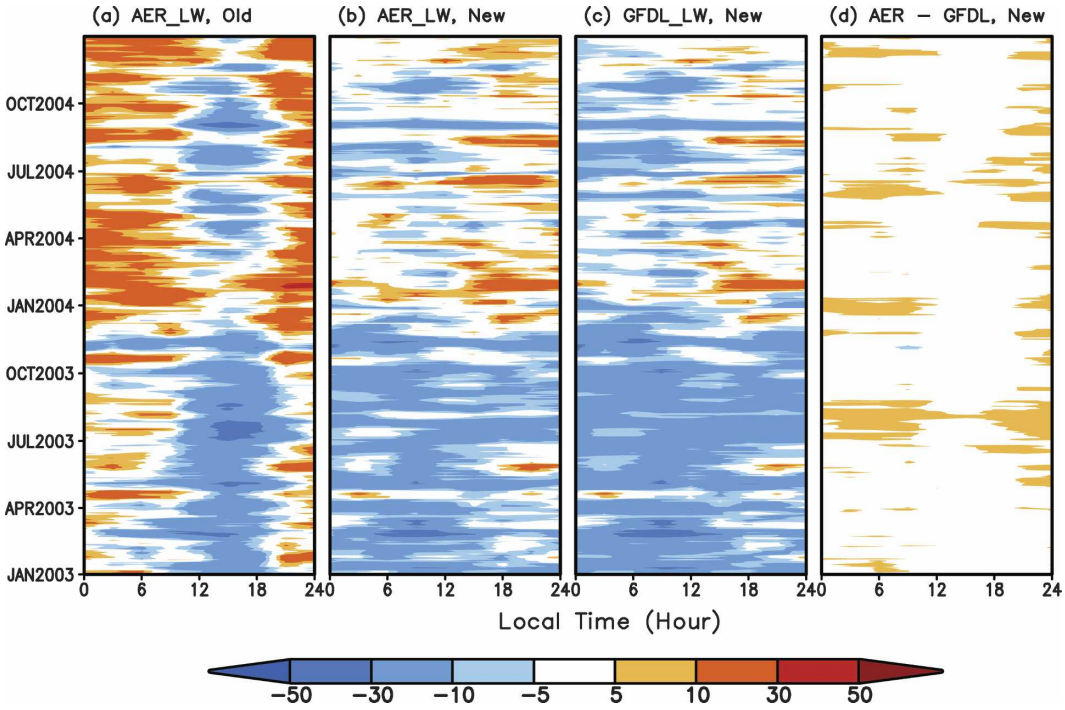

FIG. 12. SDLW from a set of SCM sensitivity experiments. (a)–(c) The differences between the SCM forecasts and ARM observations. The SCM was run with (a) the RRTM longwave routine and the old SDLW scaling method, (b) the RRTM routine and the new SDLW scaling method, and (c) the GFDL routine and the new SDLW scaling method. (d) The differences between the experiments in (b) and (c). Given the same atmospheric conditions, the GFDL routine often produces 5–10 W m^{-2} less SDLW than does the RRTM routine.

be inaccurate. The scaling factor should be a function of air temperature instead of skin temperature.

We performed two more SCM experiments using revised scaling factors. The skin temperature in Eq. (1) was replaced by the air temperature at the lowest model layer in the first experiment and by the mass-weighted mean air temperature of the lowest four model layers in the second experiment. Indeed, both the phase and magnitude of the SDLW diurnal cycle (Fig. 11b) were greatly improved with this scaling method. The magnitude of the bias relative to the ARM observations was reduced from about 20 W m^{-2} in Fig. 11a to about 5 W m^{-2} in Fig. 11b. Additional tests indicated that including more layers when calculating the mean air temperature for the scaling factor does not further reduce the bias. We are now implementing the revised method into the NCEP operational GFS.

b. Systematic SDLW biases

The longwave radiative transfer routine in the GFS was switched from the GFDL Schwarzkopf and Fels (1991) scheme to the RRTM (Mlawer et al. 1997) on 28 August 2003. Figures 3 and 6 showed that the forecast SDLW made a distinct transition at about the time of this switch. The forecast daytime fluxes were largely underestimated before that time, and matched more closely to, but were still smaller than, the observed fluxes after the switch (Fig. 3). After the switch, however, the forecast nighttime fluxes became less realistic and were overestimated by 10–30 W m^{-2}.

To assess the difference between the GFDL and RRTM routines and to test how the systematic bias might be related to the scaling factor described by Eq. (1), we performed three SCM experiments. The SCM was initialized and driven by the same GFS forecasts as described in section 5a and integrated for the years of 2003 and 2004. In the first experiment, the RRTM longwave routine was used in combination with the old scaling factor [Eq. (1)]. In the second and third experiments the SCM was run with either the GFDL or the RRTM longwave routines, but both used the revised SDLW scaling factor described in section 5a.

The differences between the SDLW in the three experiments and the ARM observations are shown in Fig. 12. There are actually three factors that contribute to the systematic SDLW bias found in Fig. 3. First, the experiment with the old scaling factor and RRTM longwave routine (Fig. 12a) has a persistent cold bias during the day and a warm bias at night, like the results shown

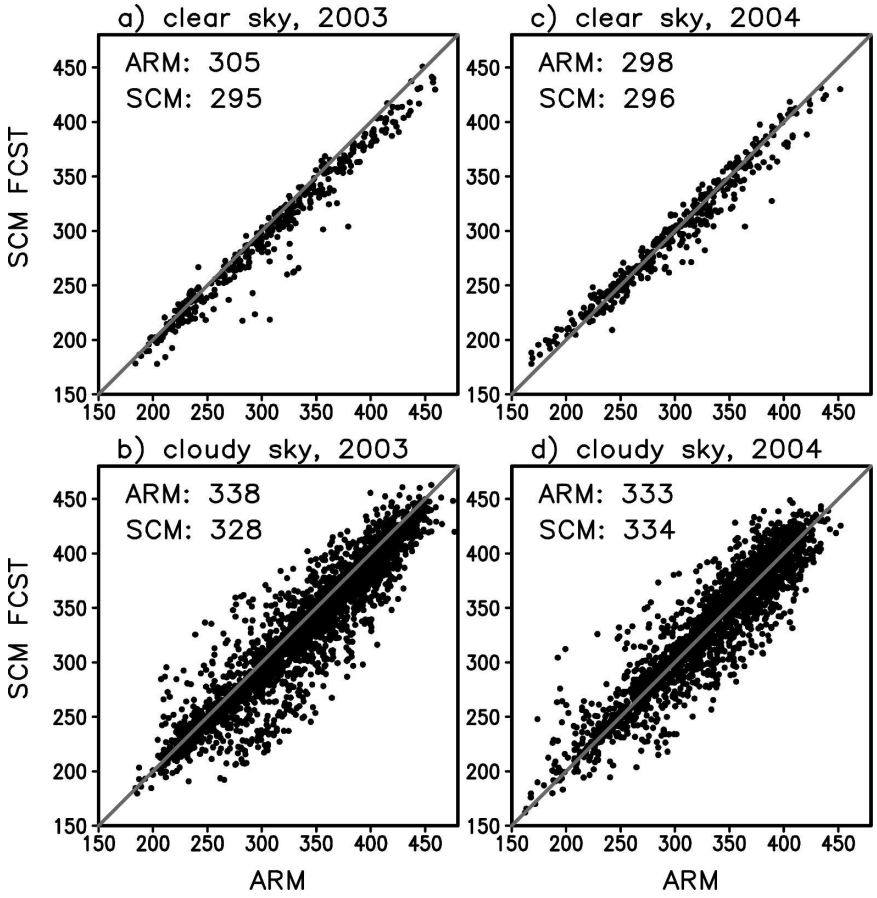

FIG. 13. SDLW: SCM forecasts with the RRTM routine against ARM observations in (a), (b) 2003 and (c), (d) 2004 and (a), (c) under clear-sky and (b), (d) cloudy-sky conditions. The annual means of SDLW for the observations and forecasts are also given in each panel.

in Fig. 3. With the revised scaling factor, the diurnal dependence of the SDLW bias disappeared no matter whether the SCM was run with either the RRTM or the GFDL longwave routine (Figs. 12b,c). Second, the RRTM and GFDL routines do have differences. The SDLW from the run with the RRTM routine is often about 5–10 W m^{-2} larger than that with the GFDL routine, especially at night (Fig. 12d). This explains in part why the warm bias in Fig. 3 became even worse at night after the switch from the GFDL routine to the RRTM routine on 28 August 2003. It should be pointed out that, even though the RRTM routine was designed to deal with the transmittances of trace gases such as methane and CFCs, only clouds, water vapor, ozone, and carbon dioxide are included in the NCEP GFS calculations, as the GFDL routine was not designed to treat trace gases. Third, there was a distinct transition at the end of 2003 (Figs. 12b,c) with both the GFDL or RRTM routines. Before that date, the SDLW was systematically underestimated during both the day and night. After that, the bias was smaller and mostly positive. This transition was coincidental and was not caused by the change in longwave routine.

Since longwave routine was not the cause for the systematic SDLW bias, other possible causes were forecast deficiencies in clouds, atmospheric temperature, or water vapor amount. First, let us determine if the biases were caused by clouds. Figure 13 shows scatterplots of SDLW from the SCM experiments with the RRTM longwave routine against ARM observations under clear- and cloudy-sky conditions for the years 2003 and 2004. Here clear sky means that neither the forecasts nor the observations have clouds, and cloudy sky means that either the forecasts or the observations have clouds. In 2003, the forecasts of SDLW were underestimated under both the clear- and cloudy-sky conditions by about 10 W m^{-2}, while the forecasts were much better in 2004 than in 2003 under both sky conditions. This indicates that clouds were not the source of the SDLW error evident in Figs. 12b,c because the results

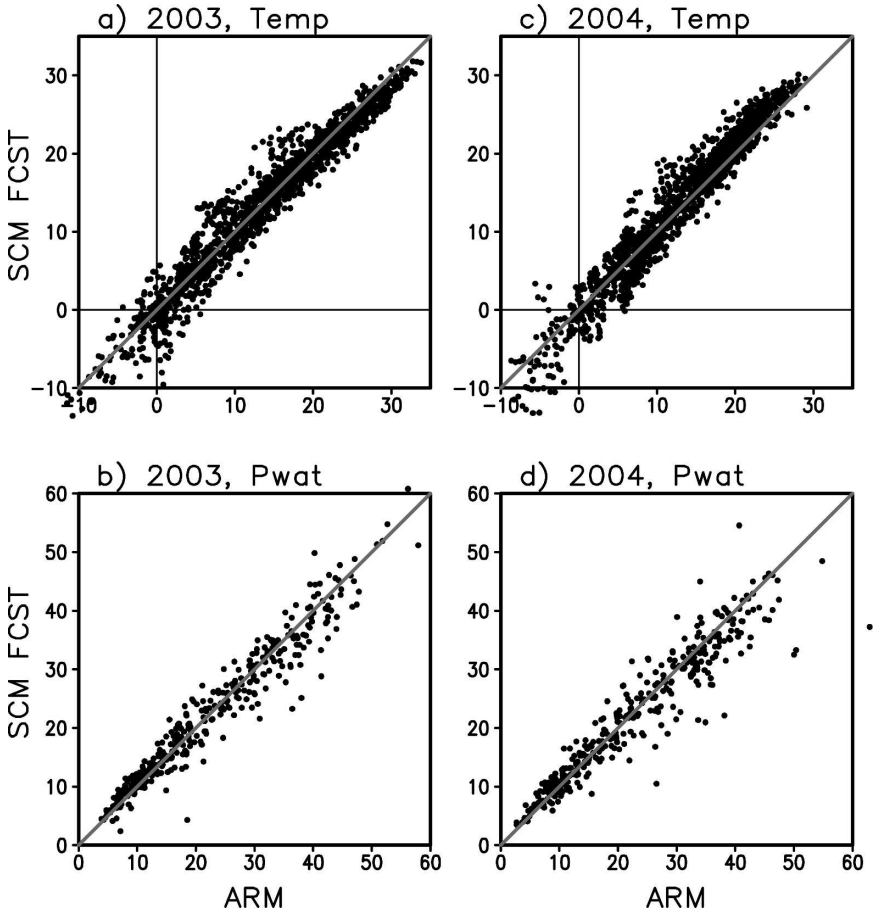

FIG. 14. Daily mean air temperatures in the lowest 1 km of the (a), (c) atmosphere and (b), (d) column-integrated water vapor amount: SCM forecasts with the RRTM routine against ARM observations in (a), (b) 2003 and (c), (d) 2004.

for both the clear- and cloudy-sky conditions were similar.

In Fig. 14 a further comparison was made between the SCM forecasts and the ARM observations in 2003 and 2004 for the daily mean air temperatures in the lowest 1 km of the atmosphere and the column-integrated water vapor amount, respectively. The forecast column water vapor amount had no systematic biases in either 2003 or 2004. The air temperature forecasts had cold biases in 2003 and minor warm biases in 2004. Therefore, the change of sign in the SDLW biases shown in Fig. 12 was probably caused by the different temperature biases in each year. However, at this time it is still not clear why the bias of the near-surface air temperature changed from negative in 2003 to positive in 2004.

c. *Surface downward shortwave flux*

Clouds have a much stronger impacts on the surface downward shortwave fluxes (SDSW) than on the SDLW at the SGP site. To assess the accuracy of the GFS shortwave routine, we evaluated the SDSW forecasts in 2003 and 2004 under clear- and cloudy-sky conditions (Fig. 15). The definitions of sky conditions are the same as those used in section 5b for the SDLW. The forecast clear-sky fluxes were relatively accurate (Fig. 15a). Under cloudy-sky conditions, the forecasts largely overestimated the flux on average (Fig. 15b). The SDSW biases identified in Figs. 3 and 5 were caused by inaccurate forecasts of cloud properties such as cloud fraction, cloud water and/or ice paths, and cloud particle size and so on.

6. Summary

This study evaluated the NCEP operational global NWP forecasts against ARM observations at the ARM SGP site for the years from 2001 to 2004. The spatial and temporal scales of the observations were examined to search for an optimum approach for comparing grid-

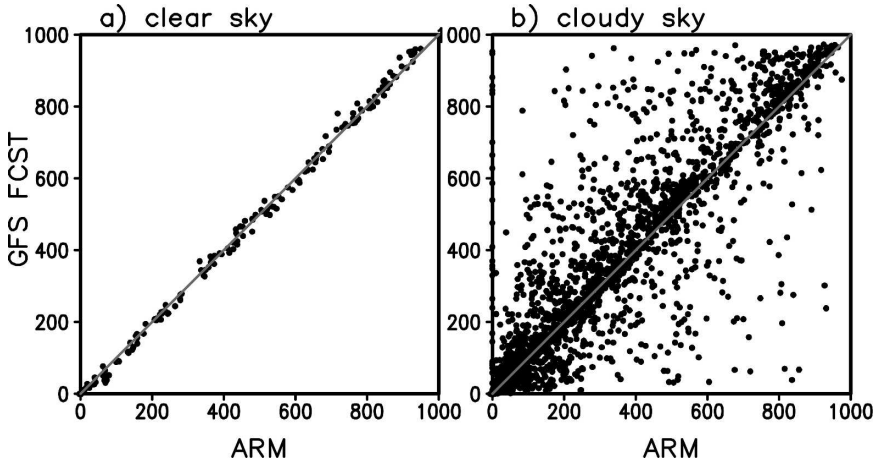

FIG. 15. SDSW: GFS forecasts against ARM observations in 2003 and 2004 under (a) clear- and (b) cloudy-sky conditions.

mean model forecasts with single-point observations. A SCM based on the GFS was used to perform sensitivity experiments to help us better understand the sources of the model errors. Unlike previous studies (e.g., Hinkelman et al. 1999; Morcrette 2002) that focused on specific synoptic events, we examined the long-term mean characteristics of the NCEP forecasts. The purpose was to diagnose systematic forecast errors, to determine the sources of errors, and to recommend changes for future model development.

The scale-dependence test on surface flux variables showed that comparing the model grid-mean forecasts with either the single-point observations made at the CF site or with the mean observations made over an area of the model grid size would give similar results. However, mean cloud occurrence frequency varied in magnitude with the time period used for computing the mean. When comparing model clouds with observations, the uncertainty of "observations" arising from the different definitions of cloud occurrence frequency needs to be taken into account. The method that Jakob et al. (2004) proposed for comparing model clouds with pointwise observations based on probabilistic distributions is one alternative approach to avoiding such uncertainties.

From 2001 to 2004, the GFS underwent a few major changes in configuration, including the implementation of a prognostic stratiform cloud condensation scheme and a new longwave radiative transfer module and increases in the model's horizontal and vertical resolutions. Overall, the performance of the model has improved. However, some forecast biases remain. Some of these biases that we found depend strongly on the season and/or time of the day. It is found that the balance of surface energy in the forecasts was a result of the cancellation of errors among the individual flux terms. The GFS largely overestimated evaporation and surface downward solar radiation, and underestimated sensible heat flux during the day. The GFS has a surface albedo lower than observed, especially over a snow-covered surface. The largest bias in latent heat flux was found in the spring and fall. The model did not simulate well the observed seasonal and interannual variations in sensible heat flux.

The GFS was able to capture the observed vertical cloud structures during major synoptic events. However, on average, the model underestimated the cloud fraction in the lower and midtroposphere in all seasons, and slightly overestimated the cloud fraction in the upper troposphere in all but the spring seasons. The diurnal cycles of clouds in the lower troposphere from the forecasts were weaker than those from the observations in all seasons, especially in the summer and fall. The model underestimated deep convective clouds in the afternoon and entirely missed the observed daytime nonprecipitating clouds in the lower troposphere. Both the shallow and penetrative convection schemes in the GFS require further attentions.

We performed a set of SCM experiments to investigate the source of errors in the forecast model's surface downward longwave fluxes. It was shown in section 3 that the diurnal cycle of the surface downward longwave flux (SDLW) from the forecast model was not in phase with that from the ARM observations, and that the nighttime SDLW was overestimated and the daytime SDLW was underestimated. Our SCM experiments demonstrated that the error was caused by an inaccurate scaling factor in the forecast model, which was a function of the skin temperature and was used to adjust the SDLW at each model time step to that com-

puted by the model's longwave routine once every 3 h. The use of a new scaling factor that is a function of near-surface air temperature eliminated this error. We also found that the SDLW biases changed from mostly negative in 2003 to positive in 2004 due to a corresponding change in the bias of the near-surface air temperature. We concluded that clouds and water vapor were not the major sources of the SDLW error. Our SCM sensitivity experiments also showed that under the same atmospheric conditions, the SDLW flux simulated by the newly implemented RRTM longwave routine is usually 5–10 W m^{-2} larger than that simulated by the earlier GFDL longwave routine.

The forecast SDSW was compared to ARM observations for clear- and cloudy-sky conditions. The forecasts were relatively accurate under clear-sky conditions. Under cloudy-sky conditions, the forecast SDSW was largely overestimated on average. The large SDSW errors were caused by inaccurate forecasts of cloud properties rather than by the radiative transfer routine itself.

This investigation focused on the surface fluxes and clouds. The intensive ARM observations at high temporal and vertical resolution allowed us to examine in detail the diurnal cycles of the surface fluxes and the vertical distribution of clouds. We were able to identify some of the forecast biases and link them to the model's representation of a particular physical process; however, for most of the forecast biases there are as yet no simple explanations. They might have resulted from the coupling between several physical processes. For example, an underestimation of cloud allows excessive solar radiation to reach the surface, which, in turn, produces large surface latent heat flux. In nature, this would lead to increased shallow convection and cloudiness. However, the shallow convection scheme in the GFS produces no cloudiness. This is an obvious deficiency. Better forecasts of clouds are crucial to improving the overall performance of the forecast model.

Acknowledgments. The authors wish to thank Steven Lazarus for archiving some of the early GFS column output, the staff at the ARM Archive Center for providing ARM observations, and Mary Hart for proofreading the manuscript. We would also like to thank Brad S. Ferrier and Michael Ek for their thoughtful comments. We greatly appreciate the insights of two anonymous reviewers, which helped to improve the manuscript. Fanglin Yang was supported by the U.S. Department of Energy (DOE) ARM Program and NCEP. Steven K. Krueger was supported by the DOE ARM Program under Grant DE-FG03-94ER61769.

REFERENCES

Ackerman, T., and G. M. Stokes, 2003: The Atmospheric Radiation Measurement Program. *Phys. Today,* **56,** 38–45.

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus ensemble with the large-scale environment, Part I. *J. Atmos. Sci.,* **31,** 674–704.

Barnett, T. P., J. Ritchie, J. Foat, and G. Stokes, 1998: On the space-time scales of the surface solar radiation field. *J. Climate,* **11,** 88–96.

Briegleb, B. P., 1992: Delta-Eddington approximation for solar radiation in the NCAR Community Climate Model. *J. Geophys. Res.,* **97,** 7603–7612.

Chou, M. D., and M. J. Suarez, 1999: A solar radiation parameterization for atmospheric studies. NASA Tech. Memo. 104606, Vol. 11, 40 pp.

Clothiaux, E. E., T. P. Ackerman, G. G. Mace, K. P. Moran, R. T. Marchand, M. Miller, and B. E. Martner, 2000: Objective determination of cloud heights and radar reflectivities using a combination of active remote sensors at the ARM CART sites. *J. Appl. Meteor.,* **39,** 645–665.

——, and Coauthors, 2001: The ARM Millimeter Wave Cloud Radars (MMCRs) and the Active Remote Sensing of Clouds (ARSCL) Value Added Product (VAP). DOE Tech. Memo. ARM VAP-002.1, U.S. Department of Energy, Washington, DC, 56 pp.

Clough, S. A., M. J. Iacono, and J.-L. Moncet, 1992: Line-by-line calculations of atmospheric fluxes and cooling rates: Application to water vapor. *J. Geophys. Res.,* **97,** 15 761–15 785.

Ebert, E. E., and J. A. Curry, 1992: A parameterization of ice cloud optical properties for climate models. *J. Geophys. Res.,* **97,** 3831–3836.

Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.,* **121,** 764–787.

Heymsfield, A. J., and G. M. McFarquhar, 1996: High albedos of cirrus in the tropical Pacific warm pool: Microphysical interpretations from CEPEX and from Kwajalein, Marshall Islands. *J. Atmos. Sci.,* **53,** 2424–2451.

Hinkelman, L. M., T. P. Ackerman, and R. T. Marchand, 1999: An evaluation of NCEP Eta model predictions of surface energy budget and cloud properties by comparison to measured ARM data. *J. Geophys. Res.,* **104,** 19 535–19 549.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.,* **124,** 2322–2339.

Hou, Y.-T., S. Moorthi, and K. A. Campana, 2002: Parameterization of solar radiation transfer in the NCEP models. NCEP Office Note 441, 46 pp.

Hu, Y. X., and K. Stamnes, 1993: An accurate parameterization of the radiative properties of water clouds suitable for use in climate models. *J. Climate,* **6,** 728–742.

Jakob, C., 2003: An improved strategy for the evaluation of cloud parameterizations in GCMs. *Bull. Amer. Meteor. Soc.,* **84,** 1387–1401.

——, R. Pincus, C. Hannay, and K.-M. Xu, 2004: Use of cloud radar observations for model evaluation: A probabilistic approach. *J. Geophys. Res.,* **109,** D03202, doi:10.1029/2003JD003473.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428.

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

Lazarus, S. M., S. K. Krueger, and G. G. Mace, 2000: A cloud climatology of the Southern Great Plains ARM CART. *J. Climate,* **13,** 1762–1775.

Lenderink, G., and Coauthors, 2004: The diurnal cycle of shallow cumulus clouds over land: A single-column model intercomparison study. *Quart. J. Roy. Meteor. Soc.,* **130,** 3339–3364.

Long, C. N., 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, DC, 19 pp.

——, T. P. Ackerman, and J. E. Christy, 2002: Variability across the ARM SGP area by temporal and spatial scale. *Atmospheric Radiation Measurements and Applications in Climate,* J. A. Shaw, Ed., Vol. 4815, *Proceedings of SPIE,* The International Society for Optical Engineering, 51–57.

Luo, Y., S. K. Krueger, G. G. Mace, and K. M. Xu, 2003: Cirrus cloud properties from a cloud-resolving model simulation compared to cloud radar observations. *J. Atmos. Sci.,* **60,** 510–525.

——, ——, and S. Moorthi, 2005: Cloud properties simulated by a single-column model. Part I: Comparison to cloud radar observations of cirrus clouds. *J. Atmos. Sci.,* **62,** 1428–1445.

Mace, G. G., C. Jakob, and K. P. Moran, 1998: Validation of hydrometeor occurrence predicted by the ECMWF model using millimeter wave radar data. *Geophys. Res. Lett.,* **25,** 1645–1648.

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

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.,* **102,** 16 663–16 682.

Moorthi, S., H.-L. Pan, and P. Caplan, 2001: Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. Tech. Procedures Bull. 484, Office of Meteorology, National Weather Service, 14 pp. [Available online at http://205.156.54.206/om/tpb/484.htm.]

Morcrette, J.-J., 2002: Assessment of the ECMWF model cloudiness and surface radiation fields at the ARM SGP site. *Mon. Wea. Rev.,* **130,** 257–277.

Pan, H.-L., and L. Mahrt, 1987: Interaction between soil hydrology and boundary layer developments. *Bound.-Layer Meteor.,* **38,** 185–202.

——, and W.-S. Wu, 1995: Implementing a mass flux convection parameterization package for the NMC Medium-Range Forecast model. NMC Office Note 409, 40 pp. [Available from NCEP, 5200 Auth Road, Washington, DC 20233.]

Phillips, T. J., and Coauthors, 2004: Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction. *Bull. Amer. Meteor. Soc.,* **85,** 1903–1915.

Randall, D. A., and D. G. Cripe, 1999: Alternative methods for specification of observed forcing in single-column models and cloud system models. *J. Geophys. Res.,* **104,** 24 527–24 545.

Schwarzkopf, M. D., and S. B. Fels, 1991: The simplified exchange method revisited: An accurate rapid method for computation of infrared cooling rates and fluxes. *J. Geophys. Res.,* **96,** 9075–9096.

Sela, J., 1980: Spectral modeling at the National Meteorological Center. *Mon. Wea. Rev.,* **108,** 1279–1292.

Shi, Y., and C. N. Long, 2002: Best estimate radiation flux value added procedure: Algorithm operational details and explanations. DOE Tech. Memo. ARM-TR-008, U.S. Department of Energy, Washington, DC, 55 pp.

Stokes, G. M., and S. E. Schwartz, 1994: The Atmospheric Radiation Measurement (ARM) Program: Programmatic background and design of the cloud and radiation test bed. *Bull. Amer. Meteor. Soc.,* **75,** 1201–1221.

Sundqvist, H., E. Berge, and J. E. Kristjánsson, 1989: Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model. *Mon. Wea. Rev.,* **117,** 1641–1657.

Tiedtke, M., 1983: The sensitivity of the time-mean large-scale flow to cumulus convection in the ECMWF model. *Proc. ECMWF Workshop on Convection in Large-Scale Models,* Reading, United Kingdom, ECMWF, 297–316.

Troen, I., and L. Mahrt, 1986: A simple model of the atmospheric boundary layer; Sensitivity to surface evaporation. *Bound.-Layer Meteor.,* **37,** 129–148.

Wesely, M. W., D. R. Cook, and R. L. Coulter, 1995: Surface heat flux data from energy balance Bowen ratio systems. Preprints, *Ninth Symp. on Meteorological Observations and Instrumentation,* Charlotte, NC, Amer. Meteor. Soc., 486–489.

Xie, S. C., and M. H. Zhang, 2000: Analysis of the convection triggering condition in the NCAR CCM using ARM measurements. *J. Geophys. Res.,* **105,** 14 983–14 996.

Xu, K. M., and D. A. Randall, 1996: A semiempirical cloudiness parameterization for use in climate models. *J. Atmos. Sci.,* **53,** 3084–3102.

——, and Coauthors, 2002: An intercomparison of cloud-resolving models with the atmospheric radiation measurement summer 1997 intensive observation period data. *Quart. J. Roy. Meteor. Soc.,* **128,** 593–624.

Zhao, Q. Y., and F. H. Carr, 1997: A prognostic cloud scheme for operational NWP models. *Mon. Wea. Rev.,* **125,** 1931–1953.