Assessing the Performance of Multiple Regional Climate Model Simulations for Seasonal Mountain Snow in the Upper Colorado River Basin

NADINE SALZMANN* AND LINDA O. MEARNS
National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 13 September 2010, in final form 20 September 2011)

ABSTRACT

This study assesses the performance of the regional climate model (RCM) simulations from the North American Regional Climate Change Assessment Program (NARCCAP) for the Upper Colorado River basin (UCRB), U.S. Rocky Mountains. The UCRB is a major contributor to the Colorado River’s runoff. Its significant snow-dominated hydrological regime makes it highly sensitive to climatic changes, and future water shortage in this region is likely. The RCMs are evaluated with a clear RCM output user’s perspective and a main focus on snow. Snow water equivalent (SWE) and snow duration, as well as air temperature and precipitation from five RCMs, are compared with snowpack telemetry (SNOTEL) observations, with National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCAR) Reanalysis II (R2), which provides the boundary conditions for the RCM simulations, and with North American Regional Reanalysis (NARR). Overall, most RCMs were able to significantly improve on the results from the NCEP–NCAR reanalysis. However, in comparison with spatially aggregated point observations and NARR, the RCMs are generally too dry, too warm, simulate too little SWE, and have a too-short snow cover duration with a too-late start and a too-early end of a significant snow cover. The intermodel biases found are partly associated with inadequately resolved topography (at the spatial resolution of the RCMs), imperfect observational data, different forcing techniques (spectral nudging versus no nudging), and the different land surface schemes (LSS). Attributing the found biases to specific features of the RCMs remains difficult or even impossible without detailed knowledge of the physical and technical specification of the models.

1. Introduction

The assessment of climate impacts and the development of adequate adaptation measures are among the greatest and most important challenges facing mankind in the coming decades. While climate change is a global problem, impacts are mainly felt by society on regional to local scales. Therefore, climate impact assessments require primarily regional- to local-scale climate data for the past and the present and scenarios for the future. Regional climate models (RCMs) are among the most promising tools to simulate climate on the regional scale. Various initiatives worldwide—Coordinated Regional Climate Downscaling Experiment (CORDEX; Giorgi et al. 2009), North America: North American Regional Climate Change Program (NARCCAP; Mearns et al. 2009), Europe: Prediction of Regional Scenarios and Uncertainties for Defining European Climate Change Risks and Effects (PRUDENCE; Christensen et al. 2002) and Ensemble-Based Prediction of Climate Changes and Their Impacts (ENSEMBLES; van der Linden and Mitchell 2009), South America: Regional Climate Change Scenarios for South America (CREAS; Marengo and Ambrizzi 2006), and Asia: Regional Climate Model Intercomparison Project for Asia (RMIP; Fu et al. 2005)—have produced and are producing multiple RCM simulations by using identical initial and boundary settings for each of the RCMs, in order to serve the climate impacts community with an ensemble of simulations. However, the effective benefit of each of these RCMs and their ensembles for specific climate impacts assessments remains to be proven for individual impact studies. The climate impacts community as a target user of RCM output is a highly heterogeneous group consisting...
of natural scientists, social scientists, and policy makers. Similarly diverse are their demands for climate information because each system (natural or social) has its specific and mostly nonlinear climate sensitivities. Accordingly, for climate impacts studies (i) the significance of one climate variable varies for different (physical) processes; (ii) there are processes, where, for example, changes in variability are more important than changes in mean and vice versa; and (iii) process sensitivities may depend on the considered spatial and temporal scale. Furthermore, model performance is not necessarily transferable between variables and regions (Takle et al. 1999, 2007; Giorgi and Mearns 1999); that is, a model with good performance for one variable or region does not necessarily do well for another variable or region. This is also true for models that do well for current climate conditions compared with their “performance” under scenario conditions. Consequently, prior to the use of results from RCMs for a specific application, the performance and uncertainties of a specific variable generated by an RCM need to be evaluated.

For standard climate variables and parameters, like mean air temperature or precipitation, many evaluation studies exist. These analyses are usually done for large areas and with topographically relatively homogenous ground. For mountain regions and for more specific variables and parameters, the number of evaluation studies is much smaller. For instance, for seasonal snow in mountainous regions on fine scales only few studies exist (e.g., Leung and Qian 2003; Leung et al. 2004), despite its significance regarding climate change impacts. Mountain snow plays an important role in the climate system and in the hydrological cycle at all spatial scales (Kukla 1981; Vavrus 2007). Because of its proximity to melting conditions, snow is highly sensitive to climatic changes (Räisänen 2008) with effects on altitude-dependent trends for snow cover duration and total snow amount (Laternser and Schneebeli 2003). Changes in the timing and in the amount of a seasonal mountain snow cover greatly influence the hydrological cycle in many basins worldwide (Stewart 2009) with important impacts on the availability of freshwater for domestic use, irrigation, or power generation. These impacts are moreover not limited to the mountainous areas themselves, but often include large adjacent lowland areas and can thus affect a great part of the world’s population (Barnett et al. 2005).

Seasonal mountain snow is an important climate impact factor; however, measuring and modeling of snow in rough terrain is challenging. Measuring is often considerably complicated by difficulties in accessing and maintaining the measurement station, and because of the high spatial and temporal variability of seasonal snow (Liston and Elder 2006), modeling in mountain topography is challenging. During the past years, important efforts have been undertaken in developing, intercomparing, and advancing snow modules (e.g., Slater et al. 2001; Essery et al. 2009). However, for RCM output where the snow modules are embedded in an RCM structure, only very few regional- to local-scale snow evaluations studies are available.

In this study we assess the performance of RCM simulations provided within NARCCAP (Mearns et al. 2009) with regard to the seasonal snow regime in the Upper Colorado River basin. This study differs somewhat from “standard” climate model evaluations as it is carried out from a climate model output user’s perspective. In addition to air temperature and precipitation, which are both closely linked to snow parameters, we analyze snow water equivalent (SWE; directly provided by the RCMs) and snow cover duration (derived and defined via SWE). NARCCAP results are compared with in situ observations and with data from reanalyses. With increasing relevance of impacts in climate change research, this shift in evaluation from particularly climate-relevant variables and parameters to impact-relevant ones will increasingly become important. In this manner, a second goal of the study is to provide a supportive basis for further climate impact studies in the Colorado River area—a region that is likely going to face significant water availability shortages caused in large part by a changing snow regime in the Upper Colorado River basin.

2. Study area

The study focuses on the Colorado River and the Upper Colorado River basin (UCRB), Colorado (Fig. 1). The Colorado River flows some 1400 km from the headwaters in the Rocky Mountains of Wyoming and Colorado to the Gulf of California, and thereby encompasses portions of seven U.S. states and Mexico. The Colorado River provides water for more than 27 million people, irrigation water to nearly 4 million acres in the United States, and is often described as the most-regulated and over-allocated river in the world (USDOI 2000).

The high-elevation snowpack in the Rocky Mountains is the major contributor (about 70%) of the annual runoff, and the seasonal runoff pattern is heavily dominated by winter snow accumulation and spring melt (Christensen et al. 2004). On average, 90% of the annual streamflow is generated in the Upper Colorado River basin. The dominance of seasonal snow as a contributor to runoff, the high allocation ratios, and the semiarid nature of the basin make the Colorado River particularly sensitive to climate change. Recent studies, most of them based on large-scale general circulation models (GCMs)
(e.g., Christensen et al. 2004; Christensen and Lettenmaier 2006), project for the Colorado River basin increased air temperature and, somewhat contrary to older studies (e.g., Nash and Gleick 1991), unaltered precipitation. As a result, runoff will decrease and a remaining key question is how much runoff will still occur. Thus, changes in the seasonal mountain snow dynamics in the Upper Colorado River basin are of major interest. Compared to the other regions in the west of the United States, the intermountain states of Wyoming, Utah, and Colorado do not show a clear and statistically significant trend in decreasing SWE for 1950–99 (Regonda et al. 2005). However, it must be noted that the measurement stations in the intermountain region are generally placed on higher-elevation sites (mostly above 2500 m) than in the other regions. Also, negative runoff changes are observed at lower-elevation stations in the intermountain regions, and this trend is highly linear with elevation. It is likely that at higher elevations (roughly above 2500 m) air temperature is still cold enough to prevent significant changes in SWE.

Another factor that should be considered for the Colorado Plateau is the influence of large-scale and long-term climate signals such as the Pacific decadal oscillation (PDO) and El Niño–Southern Oscillation (ENSO; Tootle et al. 2005; Hunter et al. 2006). In years with major El Niño events (e.g., 1982/83, 1986/87, 1991/92, or 1994/95), precipitation on the Colorado Plateau tends to be higher than average during November–December and lower than average in January–March (USGS 2002; Ropelewski and Halpert 1986).

3. Data and methods

In mountain areas the distribution of seasonal snow varies significantly in space and time. From snow accumulation (snowfall) to snow ablation (melt), various processes on different spatial and temporal scales influence the distribution and redistribution of snow (Liston and Elder 2006). The initial distribution of snow accumulation is mainly related to large-scale atmospheric stability and regional- to local-scale orographically caused precipitation. Consequently, the local patterns of snow distribution show little interannual variation, while snow amount (SWE) can vary significantly. Snow redistribution is mainly caused by local processes such as wind and avalanches, which are typically dictated by the local topography. Finally, snowmelt again modifies the snow distribution (Home and Kavvas 1997). Here, the main drivers are processes related to radiation.
differences dictated mainly by the topography (aspect and slope).

There are different methods in use for measuring snow parameters, including in situ point measurements and satellite- and air-based remote sensing techniques (Lundberg and Halldin 2001). Each method has its limitations either regarding record length, area coverage, or operational applicability. For long-term and operational use, the classic ground-based station measurement is still the standard method, and in many regions these are the only available datasets with significant long records for SWE. Their main deficit is, however, lack of spatial representativeness, particularly in mountain areas (Molotch and Bales 2006).

For the UCRB region, the only long-term snow observations available are the station measurements from the snowpack telemetry (SNOTEL) network (see following section 3a). In addition to these standard in situ measurements, we also include regional [North American Regional Reanalysis (NARR)] and global [National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR)] reanalyses (see section 3b) in this study, where the latter also provides the boundary data for both the NARCCAP simulations and NARR.

a. SNOTEL observations

The ground-based point measurements network SNOTEL is operated by the Natural Resources Conservation Service (NRCS). NRCS maintains over 660 automated stations in mountain snowpack zones of the western United States and collects snowpack and related climate data such as air temperature and precipitation. Forty-seven SNOTEL stations are located within the UCRB, most of them in operation since 1981. Figure 1 shows the distribution of the SNOTEL sites within the UCRB. SNOTEL provides daily observations of SWE, air temperature, and precipitation.

In addition to the challenge of comparing point measurements with gridded data (cf. section 3d), there are other SNOTEL-specific issues to be considered. The SNOTEL network was initially installed in the late 1970s primarily as a water-supply-forecasting hydroclimatic data collection network. As a result, SNOTEL sites are not necessarily distributed to represent physiographic and snowpack conditions on a watershed, as shown by Molotch and Bales (2006) for the Rio Grande headwaters. SNOTEL sites were mostly specifically placed in areas where the streamflow originates—that is, where snow accumulation is normally high and the snowpack duration is long. As a result, the average of SNOTEL data in an area is probably somewhat high compared to what the “true” area average would be. Furthermore, particularly during the initial stage of the SNOTEL network operation, there were several irregularities found in reporting of the data, plus differences in instruments and their configurations and installations between the various sites. A basic data quality check has been operationally done for the SNOTEL data, but they do not undergo a higher level of correction and/or homogenization (T. Pagano, NRCS, 2008, personal communication). Therefore, SNOTEL data must be used with some caution, particularly when it comes to climate trend analyses, where trends might be generated by data inhomogeneities rather than by observed climatic change.

b. Reanalysis

Reanalyses are increasingly recognized and used by the climate impacts community. They provide a continuous stream of three-dimensional fields of meteorological variables of the past through advanced data assimilation techniques of various observations (Bengtsson and Shukla 1988). Data gaps are filled with dynamically and physically consistent model-generated information. In remote high-mountain regions and data-sparse areas in general, reanalyses may thus provide valuable additional or even the only time series for meteorological variables. In addition to global reanalyses such as the NCEP–NCAR reanalysis (Kalnay et al. 1996) or 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005), there are also higher-resolution regional reanalyses available such as NARR (Mesinger et al. 2006), which can provide data for local to regional climate and climate impacts studies.

1) NCEP–NCAR REANALYSIS

The NCEP–NCAR reanalysis (see Kalnay et al. 1996) is a global reanalysis with a horizontal resolution of T62 (about 210 km) and covers the years since 1948. Daily averages for SWE, air temperature, and precipitation are available. The NCEP–NCAR reanalysis also provides the initial and lateral boundary conditions for the NARCCAP simulations and also serves as the base data for NARR. It represents the UCRB by one grid box.

2) NARR

NARR is a long-term (1979–present), dynamically consistent, high-resolution (32 km; 45 layers), high-frequency (3 hourly), data-assimilation-based atmospheric and land surface hydrological dataset for the North American domain. NARR’s lateral boundaries are taken from the global NCEP–NCAR reanalysis. The NCEP Eta Model and its data assimilation system—a recent version of the Noah land surface model—are used, as well as numerous additional datasets (Mesinger et al. 2006). In particular, NARR has assimilated high-quality and detailed precipitation observations with the Parameter-Elevation
Regressions on Independent Slopes Model (PRISM; Daly et al. 1994) into the atmospheric analysis. As a result, NARR provides a much improved analysis of land hydrology and land–atmospheric interactions. NARR has good representations of extreme events, such as floods and droughts, and interfaces well with hydrological models (Mesinger et al. 2006). Because of the relatively short time that NARR has been available to users, experience with NARR, particularly regarding snow in mountain areas, is very limited. The present study therefore also aims to contribute in this regard.

In NARR, snow processes are mainly assimilated by satellite data (K. Mitchell, NCEP/Environmental Modeling Center, 2008, personal communication). The Noah land surface scheme (LSS) simulates snowpack states of water content (SWE), density, and fractional coverage via the process of sublimation, snowfall, and snowmelt and the snowpack surface energy fluxes of radiation, sensible/latent heat flux, subsurface heat flux, and phase-change heat sources/sinks. SWE is updated daily at 0000 UTC from the daily global (47 km) snow depth analysis of the U.S. Air Force Weather Agency (SNODEP) (Mesinger et al. 2006). SNODEP uses in situ observations, a Special Sensor Microwave Imager (SSM/I)-based detection algorithm, and its own climatology to produce a global analysis of physical snow depth, once per day at 47-km resolution (Kopp and Kiess 1996).

NARR has a higher horizontal resolution (~32 km) than the NARCCAP simulations (~50 km) and represents the UCRB region with 25 grid boxes. Only 18 of these grid boxes have SNOTEL sites located within them. Most of the other seven grid boxes are low-altitude, valley grid cells. We found no significant differences in our results between using the 25- and 18-gridbox average of snow for the area. Nonetheless, we only considered the 18 grid boxes with SNOTEL sites for the further analyses in order to avoid any possible influence from low-altitude, valley grid cells. Furthermore, we only used the average values of the 18 grid boxes of NARR; that is, we did not compare single NARR grid boxes to NARCCAP grid boxes because of the difference in resolution between the two datasets.

3. Regional climate model data (NARCCAP simulations)

The international program NARCCAP (Mears et al. 2009) aims to serve the regional climate change scenario needs in North America. The program provides simulations from six RCMs driven by NCEP–NCAR reanalysis and four GCMs for current and future climate conditions. Here, we only analyze the NCEP–NCAR-driven simulations of five of the RCMs. The Hadley Centre regional model version 3 (HadRM3) is not included in this study because of delays in processing some of the data.

Each RCM has been run with a horizontal resolution of ~50 km. The NCEP–NCAR-driven simulations cover 1979–2004. The year 1979 is used for spinning up the models. Table 1 provides an overview and some specifications of the models used.

The simulation of snow processes within RCMs is mainly determined through the respective LSSs (Slater et al. 2001; Liston 2004). While the Weather Research and Forecasting Model (WRF) and Experimental Climate Prediction Center Model (ECPC) use Noah (like NARR) (Ek et al. 2003; Mitchell et al. 2004), the Canadian Regional Climate Model (CRCM) uses the Canadian Land Surface Scheme (CLASS; Verseghy 1991) and the Regional Climate Model (RegCM) uses the Biosphere–Atmosphere Transfer Scheme (BATS; Dickinson et al. 1993; Yang et al. 1997) (see Table 1). The differences in how these LSSs simulate snow processes are mainly given by their different parameterizations and their treatment of subgrid variability of vegetation and elevation, which affects snow through subgrid variations in precipitation, temperature, and radiation. LSSs usually have standard parameter modes (e.g., for the range of albedo values used for snow aging and melting processes). However, these standard ranges can be varied by the modelers, and it is often nontransparent to the user of the RCM output which mode was used for a certain RCM simulation. Furthermore, it has been shown by several studies (e.g., Desborough 1997), that the internal model structure and the related coupling procedures influence the final simulation results. That is, a similar or equal parameterization embedded in a different model structure can lead to significantly different simulation results.

The UCRB is represented by 10 NARCCAP grid boxes (Fig. 1a). The topography of the NARCCAP RCM for the UCRB region and its comparison with SNOTEL elevations is shown in Fig. 1b.

d. Analysis setup

1) POINT-TOGRID COMPARISON

The data used in this study span a range of spatial scales, from point measurements (SNOTEL) to gridded datasets with horizontal resolutions of 32 (NARR), 50 (NARCCAP), and 210 km (NCEP–NCAR). The comparability of data with such different scales is somewhat problematic, especially in mountainous terrain. Molotch and Bales (2005) showed that SNOTEL point data do not necessarily represent well area averages of snow distribution, and there is also no overlap between correlations found for accumulation and melt. Nevertheless, SNOTEL
and NARR are probably the best possible datasets for the UCRB. To our knowledge no better practical options for daily observations and particularly for snow are available, which is a common problem in mountain areas. Furthermore, SNOTEL data have been used in other studies (e.g., by Simpson et al. 1998) for validation of snow cover products derived from remote sensing data, by Serreze et al. (1999) for regional snowpack and precipitation analysis in the mountainous west, and by Pan et al. (2003) for validation of the North American Land Data Assimilation System’s snow cover simulation. These authors all state that despite the scale issue, SNOTEL is the only practical option for validation in this region. Nevertheless, the scale issue has to be addressed and in the present study this is done in a number of ways as briefly described below. Regarding SNOTEL, small-scale variability has been removed through spatial and temporal data aggregation, as also suggested by Pan et al. (2003) and Andreadis and Lettenmaier (2006). Furthermore, only the averages of all SNOTEL sites within one grid box (with an average of 4.7 SNOTEL sites per RCM grid box) and the average of all stations within the whole UCRB were used. For many of our analyses we used monthly means, as well as monthly means aggregated over a 5-yr period. Finally, and most importantly, we do not assume that the observations are the absolute “truth,” but more as being indicative of the true values. We do not intend to formally rank model results regarding quality but rather to intercompare modeled and measured datasets.

2) ANALYZED TIME PERIODS

The two time periods 1981–86 and 1991–96 are analyzed. For these two time periods, SNOTEL and NARR data are based on different numbers of observations. This also allows us to analyze the effects of a varying number of observations. During the second period, 47 SNOTEL stations have been active, whereas fewer stations have been available during the first period. Also for NARR, the availability of observations for data assimilation was lower in the first period compared to the second. Moreover, the two time periods also include some major El Niño events in the years 1982/83, 1991/92, and 1994/95.

4. Results

Whether precipitation finally hits the ground in solid (snow) or liquid (rain) form is mainly governed by the local air temperature, but is also influenced by humidity and saturation, respectively. The air temperature furthermore influences significantly the duration and dynamics of the seasonal snow cover through melt and freeze processes. Because of these interdependences
between air temperature and snow, and thus also the importance of air temperature to explain departures found for snow (section 4c), the analysis here starts with air temperature.

a. Air temperature

The monthly mean 2-m air temperatures spatially averaged over the 10 UCRB grid boxes are shown in Fig. 2. Generally, during winter (the main season of interest in this study) the differences between the models are larger than during summer. The NARCCAP results compared with SNOTEL and NARR exhibit a distinct warm bias during summer (WRF being the warmest). The WRF results analyzed here were simulated with the WRF using the Kain–Fritsch parameterization. With this setting, WRF has shown a significant warm bias in other regions caused by the Kain–Fritsch convection scheme, which produces too-few convective clouds and thus too-high radiation. On the other end of the range, CRCM has overall the lowest winter air temperatures—lower than the measured SNOTEL temperature in each of the analyzed years. Overall, WRF and RegCM3 are the warmest models. The NCEP–NCAR reanalysis has particularly warm summer air temperatures compared to SNOTEL. ECPC and CRCM have lower biases for temperature, while RegCM and WRF have relatively warm biases. In summary, compared to SNOTEL, the results of NARCCAP, NARR, and NCEP–NCAR show higher air temperatures in winter on the order of 8–10 K. During summer the biases vary in a range of about 6 K, with SNOTEL situated usually in the lower third of the range. For the whole analyzed time spans, Table 2 provides the $R^2$ and RMSE statistics, which additionally supports the above findings. In general, the variations across the models are relatively constant, with slightly more constancy for the 1991–96 than for the 1981–86 period, perhaps as a result of more and probably improved data during the 1991–96 period.

The consistencies found in the model data would theoretically allow for lapse-rate corrections. Such adjustments might be of particular interest when intending to use and apply RCM output for impact studies (e.g., Salzmann et al. 2007). However, as the difference in elevation between the SNOTEL sites and NARCCAP topography shows (Fig. 1b) there is only an elevation difference of about 200 m, which can thus only explain a small part of the effective air temperature deviations found. The lapse-rate issue is further discussed in section 4d.

b. Precipitation

1) ACCUMULATED ANNUAL PRECIPITATION

Analyses of the accumulated annual precipitation provide insights about the temporal evolution of precipitation amounts. Figure 3 shows the total cumulative precipitation per hydrological year (1 October–30 September) for each of the different models and for each grid box. In general, the spatial variability, here represented as the variability across the grid boxes, is relatively low. The modeled precipitation amounts, compared to SNOTEL, are lower except by NARR for the period...
FIG. 3. Annual cumulative precipitation (mm).
1991–96. The temporal distribution and dynamics of the precipitation over 1 yr is relatively similar for most of the datasets and have a tendency to lower precipitation amounts around fall. NCEP–NCAR and WRF have higher precipitation in spring or summer for most of the years, and they generally have a very similar pattern in absolute and relative terms. Overall, both are very dry.

Regarding ENSO signals, none of the datasets, including SNOTEL, shows clearly increased precipitation during November–December and decreased precipitation during January–March for the ENSO years 1982/83, 1986/87, 1991/92, or 1994/95.

2) MONTHLY PRECIPITATION

In Fig. 4, precipitation data is aggregated temporally to monthly means and spatially over the whole UCRB region. The general patterns found in Fig. 3 are also well recognized in the spatially and temporally aggregated data. NCEP and WRF (and NARR for 1981–86) have
low total precipitation, ENSO events are not discernible, and ECPC and CRCM exhibit patterns most similar to SNOTEL, particularly ECPC during the period 1981–86. The precipitation peak measured by SNOTEL in late 1983 (calendar year) is well simulated by ECPC and the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5), and also by RegCM3, although the peak is a bit too early in the year. CRCM and NARR also represent this event to some degree. Only WRF fails completely in this regard. The models are similarly able to simulate the other two “precipitation peak events” during the 1981–86 period (late 1985 and 1982). For the second analyzed period (1991–96), the precipitation event in early 1993, as measured by SNOTEL, is again simulated best by ECPC and MM5, while RegCM3 simulates it with some delay. The event in early 1995 is, however, best represented by WRF. Also remarkable are the four events in NCEP–NCAR with high precipitation amounts. Compared to SNOTEL, they capture the annual precipitation peaks, but receive the precipitation within a very short period each year. These peaks are not evident in the RCMs and NARR except perhaps for WRF. It seems, however, that the RCMs are able to temporarily stretch the precipitation amounts of these events.

c. Seasonal mountain snow

1) Annual Snow Cycle

The annual cycles of SWE as represented by the different datasets are shown in Fig. 5, with separate plots for each grid box and the averages over the 10 grid boxes. Note that the axis values are not identical in Fig. 5b. The plots show that the intermodel spreads within the analyzed 10 grid boxes are relatively low. Largest spreads are found for RegCM and particularly for WRF and NARR. RegCM shows much higher amounts for the three grid boxes 3, 6, and 9, which represent the westernmost grid boxes within the UCRB (see Fig. 1). This pattern is also found in the SNOTEL data during the period 1991–96, but there it is much less pronounced. Furthermore, WRF and NARR show overall not plausible patterns, with no clear annual cycles. The variability of SNOTEL data within one grid box is shown in Fig. 5c. In most grid boxes, the variability between the different SNOTEL stations is relatively low, except for grid boxes 1 and also 9.

Figure 5a shows the averages of all grid boxes in one single plot and thus enables a direct comparison of the snow cycles in absolute terms. There are numerous differences between the datasets as shown in Fig. 5a. SNOTEL very clearly has the highest SWE values throughout the analyzed period. However, from this it cannot be directly concluded that all other datasets underestimate snow amount. As discussed in section 3, SNOTEL values probably overestimate SWE because of their specific location in snowy areas. Furthermore, local processes, such as distribution and redistribution by wind and avalanches, influence measured observations (SNOTEL) but are not considered in the coarse resolution of current RCM simulations. Moreover, during the period 1981–86 the averaged SNOTEL SWE values are influenced by the very high values from grid box 1 (see Figs. 5b,c), which partly results from the very high values of the single SNOTEL station within grid box 1. There might be some data quality issues or errors in the SNOTEL measurements of this specific single station. As mentioned earlier, SNOTEL data are only controlled and quality checked by basic automated systems, and there are no spatial or temporal tests manually performed on the data. Although in an application-oriented study this specific station would probably be omitted, we did not omit it here, deliberately, because this study aims also at pointing out exactly such possible constraints.

Regarding the RCMs, the highest SWE values are simulated by CRCM. This is probably a result of a combination of the low (winter) temperatures of CRCM (that are even lower than those of SNOTEL; see section...
FIG. 5. (Continued)
4a) and/or the precipitation in CRCM. During 1981–86 RegCM, and during 1991–96, RegCM and ECPC both show low SWE values. For RegCM, again, this is probably due to air temperature, which is relatively high during winter. Similar arguments can be used for ECPC. The results for WRF appear to be unrealistically very low. Compared to the season-long low SWE values for NCEP–NCAR, WRF, however, achieves at least some variation within the snow seasons. NCEP–NCAR’s SWE values increase monotonically and rapidly in late fall, although only to a very low level, and decrease in the same manner in spring. In view of the unrealistic and low SWE values as simulated by NCEP–NCAR, most NARCCAP RCMs are able to improve significantly the results for the snow cycle and the added value of dynamical downscaling becomes here particularly evident.

The above findings are also clearly shown in the $R^2$ and RMSE statistics (Table 2) between RCMs and observations (SNOTEL and NARR) for temporally and spatially aggregated SWE values for the two analyzed time periods. The $R^2$ values range from 0.18 (WRF) to 0.75 (CRCM); $R^2$ is particularly low for SNOTEL compared with NARR. The values for RMSE for SNOTEL vary between 0.15 (CRCM) and 0.25 (WRF) and for NARR between 0.08 (ECPC) and 0.09 (all others).

On the one hand, the absolute values of SWE seem generally to be too low. On the other hand, the relative snow cycle is reasonably captured by most RCMs. In particular, again, CRCM’s snow cycle is very similar to the measured snow cycle by SNOTEL stations, which is also evident by the high $R^2$ and low RMSE value. WRF, on the other end of the range, is not achieving a clear and plausible snow cycle and has thus also the lowest $R^2$ and highest RMSE values. NARR has a very high and likely unrealistic temporal variability (Fig. 5). Because snow processes in NARR are mainly assimilated by satellite data, they are relatively poorly represented. Also, the large difference in precipitation amounts of NARR between 1981–86 and 1991–96 (Fig. 4) is not reflected in the SWE values of NARR (Fig. 6) and NARR air temperature is lower in 1991–96 than in 1981–86. In fact, NARR is only expected to deliver reliable snow cover data, rather than reliable SWE values, because SWE in NARR is mainly and indirectly derived from assimilation of satellite data (K. Mitchell 2008, personal communication). Therefore, a physically based relation between air temperature, precipitation, and snow is not clearly given in NARR.

2) SEASONAL SNOW COVER DURATION

For many climate impacts processes, not only the amount of snow (SWE) but other snow parameters are critical, too, like the time of the initial snow cover, the time of snow cover end, or the time of maximum SWE. For instance, for annual hydrological runoff, the time of maximum SWE can be critical (Clow 2010), while for the ground thermal regime (e.g., permafrost) the time of the first significant snow cover in fall is of particular importance (Zhang 2005; Harris et al. 2009). The mentioned snow parameters, however, are directly provided neither by the RCMs, nor by the in situ observations, nor by the reanalyses. Therefore, we derived and defined snow cover duration as the number of days with a “significant snow cover” (ssc) for the average of all grid boxes, whereas ssc is SWE $\geq 0.01$ m for a minimum period of 10 continuous days. This definition of a significant snow cover was applied in order to avoid too many days with very low or zero SWE mainly at the beginning and end of the season (e.g., ECPC; Fig. 6). We did not include NCEP–NCAR and WRF here because of their very low SWE values (e.g., WRF hardly reaches values over 0.01 m) all year round and because no obvious seasonal evolution is apparent (NCEP–NCAR; see Fig. 5).

In accordance with the findings in the previous sections, SNOTEL stations overall report the most days in a row with a significant snow cover (Fig. 6 and Table 3). Relative to SNOTEL, the RCMs and NARR show significantly fewer ssc days (34%–74% of SNOTEL) during the two considered time periods. The time of maximum SWE is usually reached within the last third of the snow cover period by the RCMs and SNOTEL, while NARR tends to generate an earlier maximum SWE. Overall, again, CRCM corresponds best with SNOTEL.
In general, the NARCCAP RCMs and NARR match better the time of the initial ssc with SNOTEL than the time of the ssc end, where variations and delays are larger (Fig. 6 and Table 3). While SNOTEL report an ssc usually around mid-October (Table 3), NARR and the RCMs show more variability in the time of initial ssc and are mostly delayed by about 3–8 weeks. The last day of the ssc in the SNOTEL data falls usually into the period between mid-June and mid-July. NARR and the RCMs again show a much higher variability for this parameter. Here, the ssc ends between April and May (but for ECPC it mostly ends earlier in March).

The initial snow cover is among other factors coupled to air temperature. The temporal evolution of air temperature is quite similar across the NARCCAP RCMs and SNOTEL, and therefore explains part of the relatively small differences in the time of the initial snow cover. The relatively larger differences in the time of the ssc end in spring is partly associated with the significantly larger absolute SWE values of SNOTEL, which results in an extended melt period until complete disappearance of the ssc. Furthermore, the higher air temperature simulated by most of the RCMs, also in spring, explains another part of the earlier disappearance of the ssc. Because radiation fluxes dominate snowmelt in midlatitude mountainous regions (Cline 1997), snow melts more rapidly and the high amounts of snow from SNOTEL disappear at faster rates. Missing new snow accumulation in the RCMs is a less important factor during spring.

**d. Bias attribution**

Being able to attribute the detected biases to a specific source would enable major progress in regional climate modeling and their application for impacts studies. On the one hand, it would support and advance the development of de-biasing strategies for use of RCM output.
in impact studies and, on the other hand, would improve
the model performance by enhancing parameterization
of key processes in the RCMs.

Main sources of biases in RCM simulations and partic-
ularly for mountain regions are introduced by scale mis-
matches and the LSS. While the 50-km resolution of the
NARCCAP RCMs is a substantial improvement over the
resolution of the GCMs, it is still too coarse for resolving
complex mountain topography and the associated local
and regional processes such as redistribution of snow by
wind or avalanches, interception, or lapse-rate effects.
Consequently, for local climate impact studies, de-biasing
is still needed prior to their use. In regions with steep to-
pography, a lapse-rate correction often leads to important
improvements. A first, rough assessment of potential im-
provements resulting from applying a lapse-rate correc-
tion might be done by comparing those grid boxes, with
an only small elevation difference between the RCMs
and SNOTEL. Here, this is given for grid boxes 6 and 10
(however, note that grid box 6 is only represented by one
single SNOTEL station). For successful lapse-rate cor-
rection, these two grid boxes would need to have signif-
ically better SWE agreement with SNOTEL than grid
boxes 3, 5, and 8, which have much greater average elevation
differences compared to SNOTEL. Based on Fig. 5b, however, there is in none of the RCMs an
obvious better (higher) SWE found by the respective
grid box in all models. Therefore, in this study, a general
lapse-rate correction would not significantly improve
the results.

While topography is similar for each NARCCAP
simulation, the RCMs use different LSSs. Because the
simulation of seasonal snow is mainly defined in the LSS,
it can be assumed that the choice of the LSS affects the
final results. Particularly, CLASS, the LSS of the CRCM,
which treats snow as a separate soil layer (Verseghy
2000), seems to improve the overall results for snow in the
CRCM. However, regarding the other LSSs used by the
NARCCAP RCMs, there was no obvious larger similarity
found between RCMs using the same LSS. In particular,
NARR, which uses the same LSS (Noah) as the RCMs
WRF, MM5, and ECPC, does not show specific similarities
with the respective RCMs. Therefore, it seems, except for
the LSS CLASS, which includes a specific snow layer to
better solve snow processes, that the LSS’s function in
general is marginal for the final output regarding snow and

### Table 3. First and last day of an ssc, the number of days with an ssc, and the day of maximum SWE for (a) each hydrological year and for
(b) the two time periods 1981–85 and 1991–95, including the differences to SNOTEL.

<table>
<thead>
<tr>
<th></th>
<th>SNOTEL</th>
<th>NARR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>ssc: first-last day</td>
<td>No. days ssc</td>
</tr>
<tr>
<td>1981/82</td>
<td>16 Oct–13 Jul</td>
<td>271</td>
</tr>
<tr>
<td>1983/84</td>
<td>15 Oct–9 Jul</td>
<td>269</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RegCM</th>
<th>CRCM</th>
<th>ECPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssc: first-last day</td>
<td>No. days ssc</td>
<td>Day max SWE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b)</th>
<th>SNOTEL</th>
<th>NARR</th>
<th>RegCM</th>
<th>CRCM</th>
<th>ECPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tot No. days ssc</td>
<td>267</td>
<td>172</td>
<td>140</td>
<td>198</td>
<td>104</td>
</tr>
<tr>
<td>Diff to SNOTEL (%)</td>
<td>95 (64%)</td>
<td>127 (52%)</td>
<td>69 (74%)</td>
<td>163 (39%)</td>
<td></td>
</tr>
<tr>
<td>Tot No. days ssc 1991–95</td>
<td>249</td>
<td>185</td>
<td>132</td>
<td>190</td>
<td>87</td>
</tr>
<tr>
<td>Diff to SNOTEL (%)</td>
<td>65 (74%)</td>
<td>118 (53%)</td>
<td>59 (76%)</td>
<td>163 (35%)</td>
<td></td>
</tr>
</tbody>
</table>

* Not continuous.
that the biases found cannot directly be attributed to the LSS used by the RCM. The biases found for precipitation (see section 4b) are most probably more important for the performance of seasonal snow than the LSS.

The overall most plausible results were achieved by the RCM CRCM. CRCM (and also ECPC) use spectral nudging, and this may also contribute to the good results by CRCM.

Many of generally known other sources that limit accuracy such as domain size or grid resolution (Laprise et al. 2008) can be neglected in this study, since these conditions are equal in all NARCCAP experiments. There are of course many other possible sources of uncertainty, for example radiation fluxes, which, however, have not been assessed here.

In summary, a quantitative detailed bias attribution is hardly possible based on RCM simulation results alone—that is, on the data that is typically provided to the impacts community. To prove if LSSs containing enhanced snow-specific processes lead most effectively to improvements in the final results, an additional experiment with CRCM not using spectral nudging would be required. Regarding spatial resolution, it can be assumed based on other studies (e.g., Leung and Qian 2003; Ikeda et al. 2010) that higher resolution improves generally the final simulation results.

5. Discussion and conclusions

The performance for seasonal mountain snow and related variables from five NCEP–NCAR-driven RCM runs realized within the frame of the NARCCAP program has been assessed for the Upper Colorado River basin—a significant snow-dominated hydrological basin in the U.S. Rocky Mountains. The assessment is mainly done from a user’s perspective rather than a climate modeler’s point of view. Therefore, the analyses and approaches of the chosen evaluation assessments are may not always in line with “classic” intercomparison studies, and the focus is mainly on less common parameters (SWE and snow cover duration), which are, however, of interest to the impacts community.

Assessments of the performance of climate model output for impact studies are often challenged by unavailability of directly comparable observational datasets. Consequently, datasets like observations in the form of point measurements (e.g., SNOTEL), gridded observations (e.g., CRU), or reanalysis (e.g., NCEP–NCAR) must be used that mostly don’t match spatial and temporal resolution, record length, or specific parameters of the climate model outputs.

The SNOTEL network was actually set in place to provide supportive data for hydrological runoff forecasting rather than climate observation. Nonetheless, we used SNOTEL data in this study for a number of reasons. SNOTEL provides the only measured long-term data for the UCRB, and we have handled the single point data problem by analyzing mainly temporally and spatially averaged data. Because of the relatively low variation found between the SNOTEL stations within one grid box (Fig. 5c) this approach seems to be sound. Furthermore, from an impact user’s point of view it makes sense to compare the RCM output with the type of data that the impact community usually uses. Since we were aware of possible errors and biases in the observations, we finally only intercompared rather than ranked the different modeled and measured datasets.

Taking into account the above-described constraints, the findings of the NARCCAP data analysis for the UCRB might be summarized as follows: The RCMs are generally too warm and too dry. The modeled ssc is significantly shorter than the ssc reported by the SNOTEL stations. The time of maximum SWE is usually too early in the year, although realistic in relative terms, considering the shorter ssc. The annual snow cycle of SWE is relatively well represented by most RCMs, but they underestimate SWE. Quantifying exactly the underestimation of ssc and SWE, however, is difficult because of the assumed overestimation of SWE by SNOTEL. Also, local processes affecting snow distributions and redistribution that are integrated in the SNOTEL observations can’t be considered with an RCM resolution of 50 km. Single attribution of the found differences to the LSS or the topography is difficult or even impossible. However, based on our results, improvements might be gained by enhanced consideration of snow-related processes in the LSS. Finally, this study shows that most RCMs improve on the NCEP–NCAR reanalysis. Moreover, the RCMs results seem to be more plausible for SWE (and derived ssc) than does NARR. It can thus be concluded that the NARCCAP RCMs definitely provide an additional and very important regional- to local-scale data source for the UCRB and likely for other parts of the Rocky Mountain as well.

From an impact modeler’s perspective, the sensitivity of related impact processes on each of the climate variables in its spatial and temporal resolution needs to be considered in addition to the RCM performance. The actual effect of the RCM biases on a specific application can only be judged in relation to the specific sensitivity of an impact process. For instance, for permafrost, it is not critical if an RCM is able to simulate the time of maximum SWE correctly, because a significant snow cover almost fully decouples the ground from the atmosphere during the whole snow cover period and variations in snow cover thickness affects permafrost temperatures only very marginally (Zhang et al. 2001). The
ability of RCMs to simulate the time of initial snow cover is of much greater importance for permafrost. In contrast, for hydrologic runoff, the time of maximum SWE is critical (e.g., Clow 2010; Moore et al. 2007) and hydrological runoff is thus sensitive regarding RCM errors in the simulation of maximum SWE. The time of the initial snow cover, however, has no significant influence on hydrologic runoff.

Regarding impact processes, an important result from studies like the present one is therefore the information provided on the range of uncertainties for specific RCM variables and their intermodel variability. These uncertainties must be set in relation to the uncertainties associated with the available observational datasets. Impact studies that use output from RCMs must carefully consider all of these issues.

Once the performance of an RCM is assessed and the biases put in relation to the sensitivity of the impact process, the RCM output might be also adjusted for a specific impact study (e.g., by weighting simulation results toward observations or applying lapse-rate correction). Adjusting the RCM output, for instance, to SNOTEL by applying a lapse-rate correction or by weighting the RCM results toward SNOTEL observations can be useful when a hydrological model that is normally driven by SNOTEL data is driven by climate model output for scenario simulations. Such adjusting may be done on different spatial and temporal levels and includes further regional knowledge such as slope and aspect of each station or regional lapse-rate values. For example, for the UCRB, a lapse-rate adjustment should probably best be made on a monthly basis, based on investigations by Pepin and Losleben (2002) where they showed that for the Colorado Rocky Mountains the lapse-rates vary monthly.

Another often-proposed technique to handle model biases and to improve the overall model output is the use of ensemble means instead of individual model outputs. This might be appropriate for a general model validation without a specific application purpose. However, for application-oriented performance analysis, and subsequent use of RCM output for a specific impact assessment, the use of single model results rather than the use of the ensemble mean might be preferable. On one hand, because an ensemble time series is not physically based anymore, it is a mix of different model outputs. On the other hand, regarding specific application purposes, it is not necessarily given that an ensemble achieves better “performance” than a single model (e.g., Cantelaube and Terres 2005; Thomson et al. 2006). In the present study for instance, the CRCM achieves more plausible results for snow and related variables than does the ensemble mean.

This study presents an intercomparison of all significant and available SWE datasets for the UCRB and a first assessment for the NARCCAP datasets for a major mountain area in the Rocky Mountains. Such detailed and local evaluation on specific RCM variables as provided here are usually not available from state-of-the-art evaluation by climate modelers. They typically evaluate RCMs on much larger scales, which, however, mostly do not serve sufficiently the needs for local climate impacts assessments. While it is clear that modeling and evaluation on spatial scales in the order of a few kilometers as provided by current RCMs are difficult, it is where the climate impacts community can benefit most from RCMs. Results from such studies may finally also improve knowledge on mountain cryospheric processes and lead to enhanced representation of seasonal snow processes in RCMs and/or earth system models.

Acknowledgments. Most of this work was done during a postdoctoral fellowship at NCAR, with funding from the Swiss National Science Foundation and the National Center for Atmospheric Research, Boulder, Colorado. We wish to thank the North American Regional Climate Change Assessment Program (NARCCAP) for providing the data and some preliminary results used in this paper. NARCCAP is funded by the National Science Foundation (NSF), the U.S. Department of Energy (DoE), the National Oceanic and Atmospheric Administration (NOAA), and the U.S. Environmental Protection Agency Office of Research and Development (EPA).

We thank the two anonymous reviewers for their valuable comments, which have considerably improved this article.

REFERENCES


