Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project

Dag Lohmann,[1] Kenneth E. Mitchell,[1] Paul R. Houser,[2] Eric F. Wood,[3] John C. Schaake,[4] Alan Robock,[5] Brian A. Cosgrove,[6] Justin Sheffield,[3] Qingyun Duan,[4] Lifeng Luo,[5,7] R. Wayne Higgins,[8] Rachel T. Pinker,[9] and J. Dan Tarpley[10]

Received 19 February 2003; revised 4 October 2003; accepted 24 February 2004; published 9 April 2004.

[1] This paper is part of a series of papers about the multi-institutional North American Land Data Assimilation System (NLDAS) project. It compares and evaluates streamflow and water balance results from four different land surface models (LSMs) within the continental United States. These LSMs have been run for the retrospective period from 1 October 1996 to 30 September 1999 forced by atmospheric observations from the Eta Data Assimilation System (EDAS) analysis, measured precipitation, and satellite-derived downward solar radiation. These model runs were performed on a common 1/8° latitude-longitude grid and used the same database for soil and vegetation classifications. We have evaluated these simulations using U.S. Geological Survey (USGS) measured daily streamflow data for 9 large major basins and 1145 small- to medium-sized basins from 23 km^2 to 10,000 km^2 distributed over the NLDAS domain. Model runoff was routed with a common distributed and a lumped optimized linear routing model. The diagnosis of the model water balance results demonstrates strengths and weaknesses in the models, our insufficient knowledge of ad hoc parameters used for the model runs, the interdependence of model structure and model physics, and the lack of good forcing data in parts of the United States, especially in regions with extended snow cover. Overall, the differences between the LSM water balance terms are of the same magnitude as the mean water balance terms themselves. The modeled mean annual runoff shows large regional differences by a factor of up to 4 between models. The corresponding difference in mean annual evapotranspiration is about a factor of 2. The analysis of runoff timing for the LSMs demonstrates the importance of correct snowmelt timing, where the resulting differences in streamflow timing can be up to four months. Runoff is underestimated by all LSMs in areas with significant snowfall. *INDEX TERMS:* 1860 Hydrology: Runoff and streamflow; 1863 Hydrology: Snow and ice (1827); 1878 Hydrology: Water/energy interactions; 1836 Hydrology: Hydrologic budget (1655); 1818 Hydrology: Evapotranspiration; *KEYWORDS:* LDAS, streamflow, water balance

Citation: Lohmann, D., et al. (2004), Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project, *J. Geophys. Res.*, *109*, D07S91, doi:10.1029/2003JD003517.

1. Introduction

[2] The multi-institutional North American Land Data Assimilation System (NLDAS) project was initiated by *Mitchell et al.* [2000, 2004] to provide a continuous timeline of background states of the land surface to initialize coupled atmosphere-ocean-land models. The ability of land surface models (LSM) to accurately reproduce measured fluxes at the surface is an important corner stone in the development of such an LDAS. *Mitchell et al.* [2004] list all NLDAS-related papers. The paper from *Cosgrove et al.* [2003a]

[1]Environmental Modeling Center, National Centers for Environmental Prediction, National Oceanic and Atmospheric Administration–National Weather Service, Camp Springs, Maryland, USA.
[2]Hydrological Sciences Branch, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[3]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA.
[4]Office of Hydrologic Development, National Oceanic and Atmospheric Administration–National Weather Service, Silver Spring, Maryland, USA.
[5]Department of Environmental Sciences, Rutgers University, New Brunswick, New Jersey, USA.
[6]Science Applications International Corporation–Hydrological Sciences Branch, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.
[7]Now at Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA.
[8]Climate Prediction Center, National Centers for Environmental Prediction, National Oceanic and Atmospheric Administration–National Weather Service, Camp Springs, Maryland, USA.
[9]Department of Meteorology, University of Maryland, College Park, Maryland, USA.
[10]Office of Research and Applications, National Environmental Satellite Data and Information Service, Camp Springs, Maryland, USA.

0148-0227/04/2003JD003517

describes the creation of forcing data from October 1996 to realtime used by the four participating LSMs. On small scales, *Luo et al.* [2003b] and *Robock et al.* [2003] look at the forcing data and the performance of LSMs over the Southern Great Plains. *Sheffield et al.* [2003] and *Pan et al.* [2003] investigate the snow cover extent and snow water equivalent of the models at the point scale and over large spatial areas. This paper focuses on the ability of land surface models to reproduce measured streamflow in 1145 small- to medium-sized and 9 large basins and intercompares the large-scale water budget of the participating LSMs.

[3] In previous off-line tests of land surface or hydrological models it has been shown that most models are generally capable of reproducing streamflow time series on a monthly to annual timescale for large river basins up to 10^7 km^2 [*Lohmann et al.*, 1998b; *Oki et al.*, 1999; *Maurer et al.*, 2002; *Bowling et al.*, 2003; *Nijssen et al.*, 2003]. The resulting errors of the models can be attributed to an incorrect amount of runoff or an incorrect timing of the modeled runoff. The reasons for the overprediction or underprediction of the total runoff amount on annual or seasonal timescales were addressed in the following major off-line studies. The Global Soil Wetness Project (GSWP) [*Dirmeyer et al.*, 1999] experiment showed that biases in the precipitation forcing led to biases of mean annual runoff [*Oki et al.*, 1999; *Chapelon et al.*, 2002]. The biases in the resulting modeled streamflow were identified as a function of the precipitation station density. The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) phase 2(e) showed that differences in the total amount of runoff were highly influenced by the sublimation physics of the models [*Bowling et al.*, 2003; *Nijssen et al.*, 2003] and were mainly due to snow surface roughness. Models with high sublimation lose their snow pack too early and consequently underpredicted observed runoff. The PILPS phase 2(c) [*Lohmann et al.*, 1998b] demonstrated that differences in runoff production parameterization affect the seasonal cycle of runoff. Models with almost no runoff production during summer precipitation events produced more realistic streamflow time series in the summer. These models produced runoff mainly by subsurface runoff. However, during periods of intense runoff production the resulting timing of runoff in these models was delayed. It was argued that this problem might be solved with a careful calibration of the model.

[4] The question of runoff timing was also addressed in these off-line studies. In most of these studies a simple linear river routing algorithm was used to transform modeled runoff into modeled streamflow (see, e.g., *Lohmann et al.* [1998a, 1998b] or *Oki et al.* [1999]). All these models are mathematically identical linear models and therefore can be described by the impulse response function of the governing equations. Differences in runoff timing were attributed to different factors. Snowmelt timing differences were significant in the studies by *Bowling et al.* [2003] and *Boone et al.* [2004]. The storage of snowmelt in either the snowpack, surface ponding, or in the soil column influenced the timing of the runoff, but not the absolute magnitude [*Bowling et al.*, 2003]. The resulting differences in peak runoff timing between models were on the order of days to up to 3 months. *Boone et al.* [2004] confirmed these results and documented that snowmelt timing on large spatial scales where the variability of orography is significant can be improved by the introduction of snow bands. Snow bands describe the technique to split one large-scale grid cell up into a number of elevation-dependent subgrids, in which the forcing data are corrected for each of the subgrids (snow bands) on the basis of the differences in mean elevation for each of the bands compared to the mean elevation of the whole grid cell. Another study found that differences in runoff production parameterizations introduced differences in streamflow peaks [*Lohmann et al.*, 1998b]. Models with more subsurface runoff production showed time delays for the peak streamflow on the order of 1 day to about one week for major flow events. Delays were mainly the result of different vertical water transport parameterizations in the LSMs. Differences in the routing parameters lead to different horizontal travel times of water in the river channels in a study by *Oki et al.* [1999]. Horizontal travel times in river channels for large basins are typically on the order of 0.5 to 5 m/s. Assuming a meandering ratio of 1.4 [*Oki et al.*, 1999], this means that a flood wave will pass through one NLDAS grid cell (1/8°) in about 1–10 hours. We therefore expect a maximum timing uncertainty for basins for up to 10,000 km^2 to be on the order of a couple of days for an uncalibrated routing model. In a previous NLDAS-related study for large U.S. basins, these uncertainties were on the order of weeks [*Maurer et al.*, 2002]. It should be noted that a full implementation of the physically based hydraulic St-Venant equations [*Chow*, 1959] could improve this runoff timing, but would be computationally more expensive and more difficult to set up since more parameters are required.

[5] Although it has been demonstrated that land surface models can successfully reproduce streamflow on daily to annual timescales for many river basins around the globe, *Entekhabi et al.* [1999] point to the limitations of current land surface and hydrology models which are used at grid scales from 1 km to 300 km. Most models are lumped single-column models that operate outside of the spatial range for which the governing equations were derived. The underlying assumption is that the equations still capture the basic behavior of the system for which we can find effective parameters. We believe that the four participating LSMs represent, to a large extent, our current knowledge of how to model the land surface. While two of the models (VIC, Sacramento model) came from the hydrologic community, the Noah and the Mosaic models were developed to be coupled to weather prediction and climate models. This paper investigates our ability to model the land surface over the continental United States based on a priori parameter choices and calibrated parameters from previous modeling experiences. In detail we would like to address the following specific questions:

[6] 1. What are the differences between the LSMs in partitioning the water balance terms (evapotranspiration, runoff, storage change) as a function of geography? Can we explain some results from the model physics, the model setup, or the a priori model parameters? To answer these questions, we will look at model output on annual and monthly time steps in different geographic regions.

[7] 2. What is the spatial distribution of the ability of the LSMs to reproduce streamflow in small- to medium-sized

(23 km^2 to 10,000 km^2) catchments? What are the major reasons for each model to overpredict or underpredict streamflow? To answer this question, we will compare daily streamflow time series with measured streamflow.

[8] 3. Are there systematic biases in all four LSMs that can be attributed to the forcing data? We will evaluate precipitation and streamflow data to answer this question, and also make references to related NLDAS studies.

[9] 4. Are there model components missing in the LSMs? Are there specific parameterizations in one model that are superior to the other models? We can give preliminary answers these questions by looking at the overall model performance of the LSMs.

[10] 5. How robust are model parameter estimates to a change of the experiment setup? The VIC model was previously run over the NLDAS area and calibrated to match streamflow [*Maurer et al.*, 2002]. The parameters from this calibrated run were then used for the model runs described in this paper. We will look at the results from these two runs that were set up differently.

[11] To keep the impact of model spin-up to a minimum, we decided to analyze only the model output from October 1997 to September 1999, the last 2 out of 3 years of model results. A NLDAS study by *Cosgrove et al.* [2003b] has shown that after the first year the effect of initialization errors for all four LSMs is rather minimal.

2. NLDAS Setup and NLDAS Models

[12] The NLDAS configuration and models are described in more detail by *Mitchell et al.* [2004], here only a short summary is given. NLDAS is an data assimilation system in which four land surface models are driven by hourly atmospheric forcing data from the Eta Data Assimilation System (EDAS) [*Rogers et al.*, 1999] and unified gauge-based precipitation analysis [*Higgins et al.*, 2000] and satellite retrieval [*Pinker et al.*, 2003] as described by *Cosgrove et al.* [2003a] on a 1/8° latitude-longitude resolution over a domain that covers the continental United States, part of Canada, and part of Mexico (125°–67°W, 25°–53°N). Table 1 lists the sources of the primary forcing (precipitation, air temperature, air-specific humidity, air pressure at the surface, wind speed, incoming solar radiation, and incoming longwave radiation), backup forcing (used when primary forcing fields are not available), auxiliary forcing, and GOES derived skin temperature for future data assimilation work. Hourly output fields from the LSMs include surface state variables such as soil moisture, soil temperature, snow water equivalent and surface fluxes such as latent, sensible, and ground heat flux, and runoff [*Mitchell et al.*, 2004]. The following models were used in the NLDAS system.

[13] The Noah model is the LSM of the National Centers for Environmental Prediction (NCEP/EMC) also used as the lower boundary condition in many atmospheric models [*Chen et al.*, 1996; *Koren et al.*, 1999]. It participated in all major off-line land surface experiments conducted under the Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) [*Henderson-Sellers et al.*, 1993], the Global Soil Wetness Project (GSWP) [*Dirmeyer et al.*, 1999], the Distributed Model Intercomparison Project (DMIP) (M. Smith, Distributed model intercomparison project (DMIP), 2002, available at http://www.nws.noaa.gov/oh/hrl/dmip), and the Rhone Aggregation Project [*Boone et al.*, 2004].

Table 1. Content of Hourly Land Surface Forcing Files for Retrospective NLDAS Runs

	Source			
	EDAS	GOES	Gauge	Radar
Primary forcing				
2-m air temperature, K	X			
2-m air-specific humidity, kg/kg	X			
10-m u-wind component, m/s	X			
10-m v-wind component, m/s	X			
Surface pressure, mbar	X			
Downward longwave radiation, W/m^2	X			
Downward shortwave radiation, W/m^2		X		
Total precipitation, kg/m^2			X[a]	X[a]
Backup forcing				
Downward shortwave radiation, W/m^2	X			
Total precipitation, kg/m^2	X			
Auxiliary forcing				
Total precipitation: WSR-88D, kg/m^2			X[b]	X[b]
Photosynthetically active radiation, W/m^2		X		
Convective precipitation, kg/m^2	X			
For validation (plus future assimilation)				
LST: land surface skin temperature, K		X		

[a]Daily total is gauge-only. Radar estimate used to temporally partition daily into hourly.

[b]Radar-dominated estimate (some gauge data used), known as "stage II/III/IV."

[14] The Mosaic land surface model developed by *Koster and Suarez* [1996] is a surface-vegetation-atmosphere transfer scheme (SVATS) that accounts for the subgrid heterogeneity of vegetation and soil moisture with a "mosaic" approach. It also participated in some of the off-line intercomparison studies.

[15] The variable infiltration capacity (VIC) model [*Liang et al.*, 1996; *Liang and Xie*, 2001; *Cherkauer and Lettenmaier*, 1999; *Cherkauer et al.*, 2003] has been widely applied to large continental river basins, for example, the Columbia [*Nijssen et al.*, 1997]; the Arkansas-Red [*Lettenmaier et al.*, 1996], the Weser river [*Lohmann et al.*, 1998a, 1998b] River, the Elbe River [*Lobmeyr et al.*, 1999] and the Upper Mississippi [*Cherkauer and Lettenmaier*, 1999], as well as at continental scales in the study by *Maurer et al.* [2002] and global scales [*Nijssen et al.*, 2003]. It has also participated in most other off-line projects within PILPS and GSWP.

[16] The Sacramento model (SAC) is run together with the SNOW-17 temperature index model, both part of the National Weather Service River Forecast System [*Burnash et al.*, 1973; *Anderson*, 1973]. SAC is a conceptual rainfall-runoff model. It has a two-layer structure, and each layer consists of tension and free water storages. The input data requirement of the SAC model is different from all the other models. The basic inputs needed to drive SAC are rain plus snowmelt from SNOW-17 and potential evapotranspiration (PE). The outputs include estimated evapotranspiration (ET), runoff, as well as the model states. The main distinctive feature of the SAC model is that it doesn't compute an energy balance. For the NLDAS runs, the PE was obtained from the Noah model output.

[17] In this study, the LSMs were not calibrated, but many parameters for all LSMs were derived from the same

Table 2. NLDAS Model Configuration and Parameters

	Common NLDAS Classification			
	Vegetation	Soil	Elevation	Model-Specific Parameters
	1-km, global, AVHRR-based, University of Maryland [*Hansen et al.*, 2000], 13 vegetation classes	1-km STATSGO database, *Miller and White* [1998] 5-min ARS, 11 layers with variable thickness, 16 texture classes	GTOPO30 database of *Verdin and Greenlee* [1996]	
Noah	predominant vegetation class	predominant soil type	not used	standard parameters used in Eta/EDAS [*Ek et al.*, 2003], multiyear monthly climatology for AVHRR-based vegetation fraction
VIC	subgrid tiles, look-up table for vegetation fraction	predominant soil type	snow bands	*Maurer et al.* [2002], multiyear monthly AVHRR-based climatology for LAI
Mosaic	subgrid tiles for each vegetation type with 5% or more fractional coverage	predominant soil type, but weighted averages from the 11-layer soil textures for porosity	not used	monthly actual AVHRR-based values for vegetation fraction and LAI, nonstandard soil layer configuration (10 cm, 30 cm, 160 cm)
SAC	not explicit	*Koren et al.* [2000]	not used	*Koren et al.* [2000]

common database [*Mitchell et al.*, 2004] for vegetation and soil types. Each modeling group was free to choose their own parameters on the basis of these classifications, as well as model geometry, other physical parameter values (e.g., for runoff production), and their seasonal cycle of vegetation. This was done to benefit from the years of experience in each modeling group through participation in major uncoupled and coupled intercomparison studies over major river basins. Table 2 lists the main data sources for the soil and vegetation parameters and their references. *Mitchell et al.* [2004] describe these in more detail. The Mosaic and the Sacramento model did not run with their standard geometry or parameters. Mosaic changed its soil layer geometry to fixed layers of 10 cm, 30 cm, and 160 cm, rather than to make it dependent on the vegetation type. The Sacramento model was run for the first time with a priori parameters based on the work of *Koren et al.* [2000]. The VIC model ran with the parameters from a previous study from *Maurer et al.* [2002], where the VIC parameters were derived from model calibration based on daily uniform precipitation and a 3-hour model time step. For the Noah model runs we mapped the 13 vegetation types from the NLDAS configuration to the standard 13 vegetation types used in the operational coupled Noah model. We also mapped the NLDAS 16 soil texture classes to the 9 soil texture classes defined in the Noah model. We did this to ensure that we run the Noah model as close as possible to the Noah model coupled to the Eta model of NCEP/EMC.

3. Streamflow Data and Flow Direction Mask

[18] Daily streamflow data for the time period of the retrospective forcing for the entire NLDAS domain were obtained from the USGS Web site (http://waterdata.usgs.gov/nwis.sw) and the Army Corps of Engineers. We selected 1145 small basins for which data were available from 1 October 1996 to 30 September 1999. Criteria for the selection were basin size (smaller than 10,000 km^2), no visible signs for reservoir operation (OHD/NWS/NOAA) and no missing data. The 1145 basins represent 15041 grid points of the NLDAS domain, about 25% of the total land area of the reduced (cutoff at 50°N) NLDAS grid. Figure 1 shows the spatial distribution of mean annual measured runoff for these basins. The USGS and Army Corps streamflow data is stored in cubic feet per second (cfs), for this paper we re-mapped these values to mm/yr to get an idea about the spatial distribution of annual average runoff. The distribution follows closely the distribution of the mean annual precipitation as shown by *Cosgrove et al.* [2003a] with maximum values in the southeast and the northwest sections of the United States.

[19] The river flow direction mask was provided for 12 River Forecast Centers (RFC) by the Office of Hydrologic Development of the National Weather Service (S. Reed, personal communication, 2002). They used a modified method of *Wang et al.* [2000] to assign to each NLDAS grid point an integer value between 1 and 8 to characterize the eight main flow directions within each grid cell. This is sometimes referred to as a D8 model. These 12 maps were merged at NCEP/EMC into one NLDAS map and error corrected for loops and incorrect flow directions. Similar data sets have been used on various scales by above cited studies and by other authors [*Vörösmarty et al.*, 1989; *Oki et al.*, 1999]. Figure 2 shows the simulated river network for the Arkansas River. To show the reasonable agreement with the natural river network, we also plotted the river reach file RF1 data set from the Environmental Protection Agency (EPA). The complete simulated NLDAS river network is shown in Figure 3. We plotted the log10 of the upstream area in km^2 for each grid cell within the 12 RFCs.

4. Routing Model

[20] The routing model used for this study is identical with the one used in previous PILPS experiments [*Lohmann et al.*, 1998b; *Bowling et al.*, 2003]. It calculates the timing of the runoff reaching the outlet of a grid box, as well as the transport of the water through the river network. It can be coupled directly into a land surface scheme, thus adding a state variable ''surface water'' to that LSM, or it can be used off-line (like in this study) from the LSM with no further feedback. It is assumed that water can leave a grid cell only

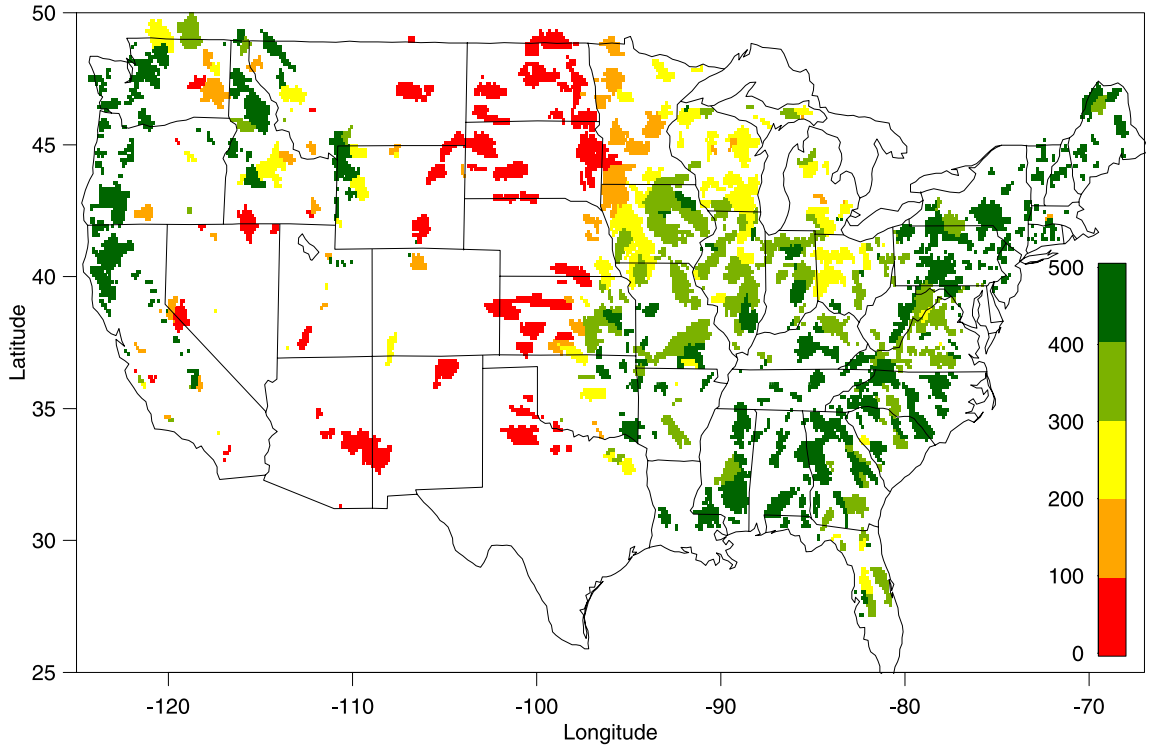

Figure 1. Annual mean observed runoff in mm/yr for 1145 small basins in the NLDAS domain for the time period 1 October 1997 to 30 September 1999. Data were provided by the USGS through their Web site http://www.usgs.gov.

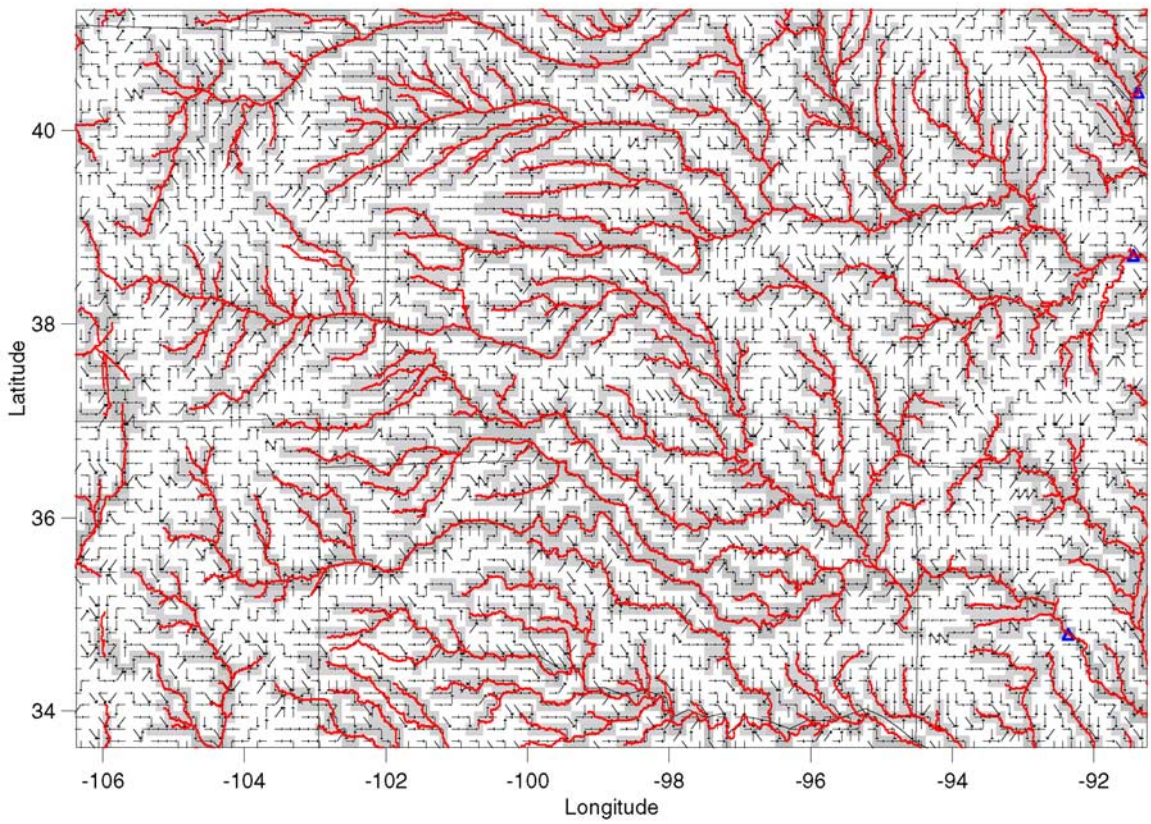

Figure 2. Example of the NLDAS 1/8° simulated river flow direction. The red lines indicate the location of the real rivers from the EPA river reach file RF1. The blue triangles are the basin outlets (Arkansas, Missouri, and upper Mississippi River).

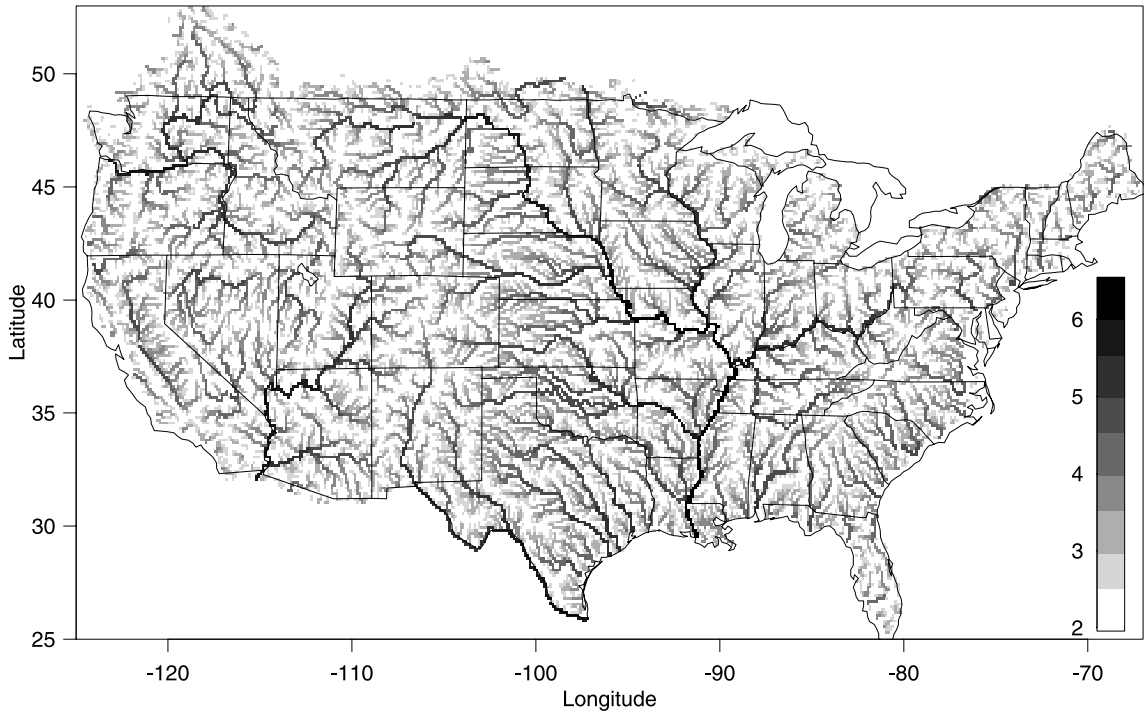

Figure 3. Logarithm to the base 10 (log 10) of the upstream area in km^2 for all grid cells.

in one of its eight neighboring grid cells, given by the river flow direction mask. Each grid cell can also function as the sink of runoff from its upstream area, like in the Great Basin (Utah, Nevada). Both with-in grid cell and river routing time delays are represented using linear, time-invariant and causal models [*Lohmann et al.*, 1998a], which are represented by nonnegative impulse-response functions.

[21] The equation used for the transport within the river is the linearized St. Venant equation.

$$\frac{\partial Q}{\partial t} = -D \cdot \frac{\partial^2 Q}{\partial x^2} + C \frac{\partial Q}{\partial x} \tag{1}$$

where Q is the discharge, D is a dispersion of diffusion coefficient, and C is the velocity. The coefficients were set in equation (1) to $C = 2$ m/s and $D = 50$ m^2/s.

[22] For this study we used the distributed approach of *Lohmann et al.* [1998b] for the major basins and small basins; as well as a simplified lumped approach for the small basins, in which we optimized the routing parameters for each model and each basin separately. The lumped approach convolutes the sum of each models runoff in each basin with one impulse response function $UH(t)$. This function is solved by deconvoluting

$$streamflow_{meas}(t) = \frac{\Delta\tau}{86.4} \sum_{\tau=0}^{\tau\,\max} \left(\sum_i area^i \cdot R^i_{t-\tau} \right) \cdot UH_\tau \tag{2}$$

where streamflow is the measured streamflow in m^3/s, $\Delta\tau$ is the time interval of the measurements (1 day), $area^i$ is the area of a grid cell in a basin in km^2, R^i is the modeled runoff of a grid cell in mm/day, 86.4 is the factor to account for the different units. τ max is the length of the impulse response function $UH(t)$ in units of $\Delta\tau$, which did not exceed 7 days for all basins. It reflects the maximum concentration time of runoff within the basins. Equation (2) was typically applied for a time period of 1 year to calculate the resulting impulse response function $UH(t)$.

[23] Figure 4 shows the travel time distribution of water in river channels for the United States. With the current parameters of the distributed routing model all runoff produced by the LSMs reaches the outlet of the river basins within maximal 50 days.

[24] Model streamflow is compared to the measured streamflow with the relative runoff bias

$$Bias = \frac{\overline{\mathrm{mod}} - \overline{meas}}{\overline{meas}} \tag{3}$$

and with the Nash-Sutcliffe efficiency criterion

$$Efficiency = 1 - \frac{\sum_i (meas_i - \mathrm{mod}_i)^2}{\sum_i (meas_i - \overline{meas})^2}, \tag{4}$$

where mod_i is the modeled streamflow with mean $\overline{\mathrm{mod}}$ and $meas_i$ is the measured streamflow with mean $\overline{meas}$ for any given time period. The Nash-Sutcliffe coefficient is a measure of the prediction skill of the modeled streamflow compared to mean daily observed streamflow. Efficiency below zero indicates that the streamflow climatology is a better predictor for the measured streamflow than the modeled streamflow; in this case the variance of the measured streamflow is smaller than the error variance. A perfect model prediction has a score equal to one.

[25] In most cases we found that using a simple lumped unit-hydrograph model for the small basin improved the resulting modeled streamflow as compared to distributed

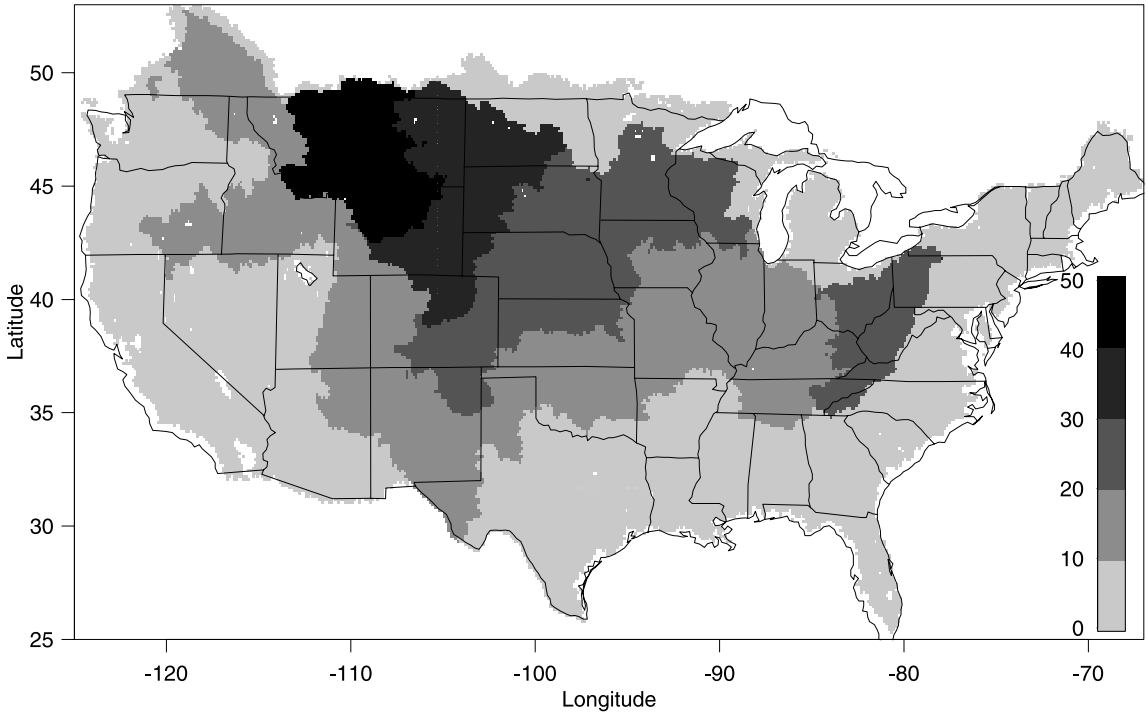

Figure 4. Distribution of travel times in days within the NLDAS domain for all grid cells to the outlet of each basin for the default parameters of the distributed routing model.

routing with default parameters. Of course, we could also optimize the distributed routing to achieve the same result, but for the basin size we chose we cannot expect too much improvement as compared with lumped routing. The reason why the lumped routing is relatively successful is because precipitation, vegetation and soil types are spatially highly correlated in the small basins. Therefore runoff production is spatially highly correlated for the models within the NLDAS domain.

5. Results and Discussion

[26] We analyzed the water budget and streamflow on two different spatial scales, large (10,000 km^2 to continental) and small to medium (23 km^2 to 10,000 km^2) river basins and three different timescales (daily, monthly, and annual).

5.1. Large-Scale Water Balance Intercomparison

[27] Figure 5 shows the spatial distribution of the modeled mean annual evapotranspiration of the time period 1 October 1997 to 30 September 1999 from the four models. Figure 6 shows the corresponding spatial distributions of mean annual runoff. The model results vary significantly in the eastern United States, but show similarities in the water limited drier western part. The similarity between the Mosaic model and Sacramento model is quite remarkable, given that the Sacramento model uses the PE computed by the Noah model as a surrogate for the atmospheric forcing. To highlight these differences, we divided the United States into four quadrants (NW, NE, SW, and SE), where 40°N and −98°W are the dividing lines. Figure 7 shows the mean annual sum of evapotranspiration and runoff for these four areas, the diagonal line is the mean annual precipitation. Over one or more annual periods the storage change of water is negligible to the other water balance terms, the sum of runoff and evapotranspiration therefore is equal or close to the precipitation amount. Model symbols below the diagonal line indicate a positive storage change for the analysis time period. Mean annual runoff in the NE quadrant varies by a factor of 4 between the VIC model and the Sacramento model, and by a factor of 3 in the SE between the VIC model and the Mosaic/Sacramento model, with the Noah model falling in between. Similar differences between models have been found in virtually all major off-line studies. Figure 8 analyzes this water balance for the 1145 small basins from Figure 7 and excludes all other grid cells. The vertical lines indicate the mean annual measured runoff values. The magnitude of evapotranspiration and runoff are very similar to the ones in Figure 7, especially in the eastern United States where almost 50% of the area is covered by the basins. In the western part the coverage with small basins is not as dense, and the basins do not seem to represent the water balance of the whole quadrant sufficiently. In order to quantify the error of the modeled mean annual evapotranspiration and runoff, we have to consider the error in the streamflow measurements (about 10% or higher for flood events) and the precipitation data set (unknown). In the SE and the NE the density of precipitation gauges is much higher than in the western regions and the influence of snowfall measurement errors is much smaller. Even if we assign a relatively large error to the precipitation measurements in the eastern regions, only the Noah model would fall within the error bounds. The Mosaic and the Sacramento model produce less runoff than observed and the VIC model produces more. However, the models underpredict runoff in the NW

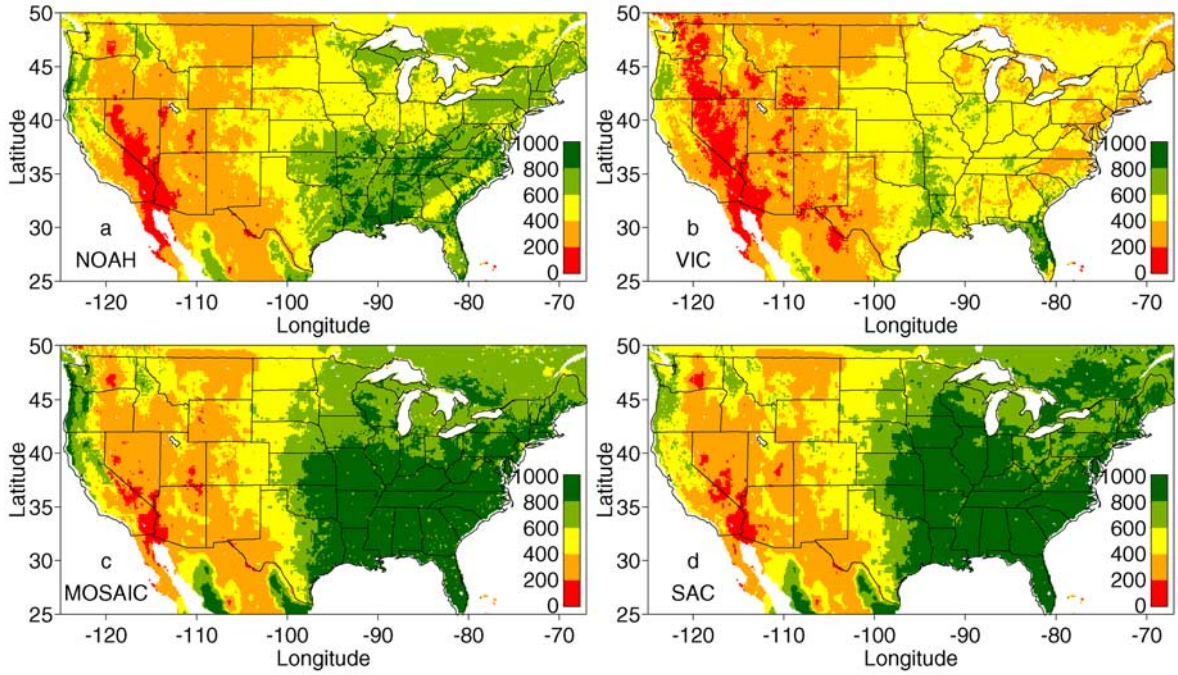

Figure 5. Mean annual evapotranspiration (mm/yr) for the NLDAS domain for the time period October 1997 to September 1999. (a) Noah, (b) VIC, (c) Mosaic, and (d) SAC.

quadrant compared to the measurements by a factor of about 1.3 (VIC) to 2 (Mosaic, Sacramento) with the Noah model in-between. This low bias in NLDAS is mainly due to the errors in the precipitation and solar insolation forcing in mountainous areas (see papers by *Mitchell et al.* [2004], *Sheffield et al.* [2003], and *Pan et al.* [2003] in this special section).

[28] Figure 9 shows the monthly water budget for October 1997 to September 1998 for each of the models in the four quadrants. The black line is the precipitation and the deviation of the red triangles from the solid black line indicates snow processes as follows: A red triangle is as much above (below) the solid black line as snowmelts (accumulates). Evapotranspiration (ET) is green, runoff is

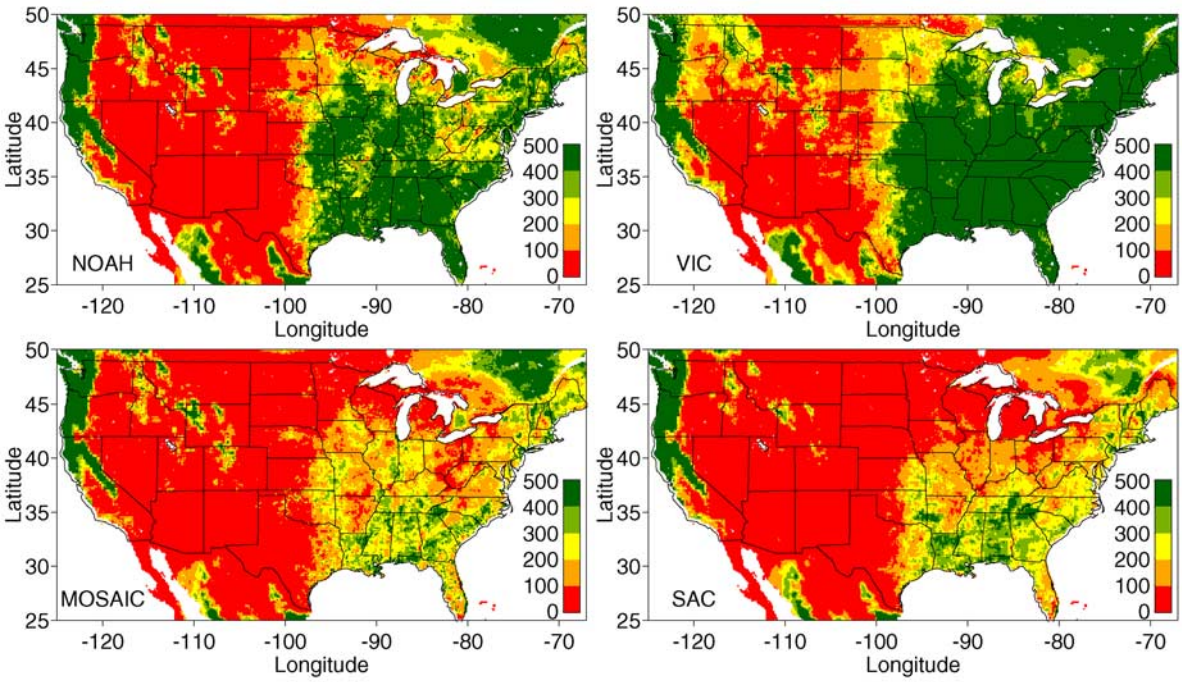

Figure 6. Mean annual runoff (mm/yr) for the NLDAS domain for the time period October 1997 to September 1999.

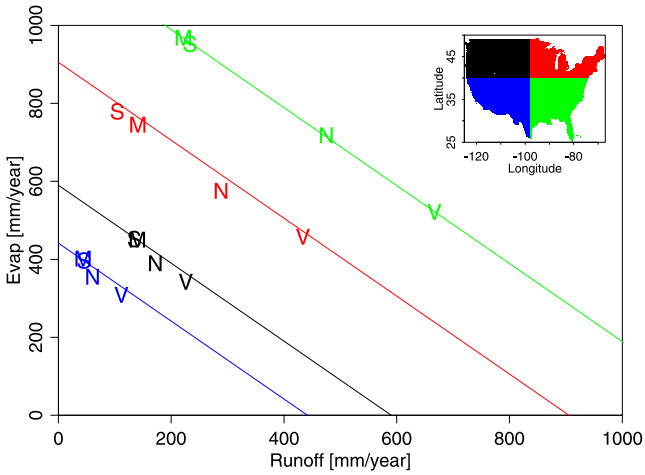

Figure 7. Partitioning of precipitation of the four NLDAS models (N, Noah; V, VIC; M, Mosaic; S, Sacramento) in four different quadrants for October 1997 to September 1999 into runoff (x axis) and evapotranspiration (y axis). The diagonal line is the mean of the precipitation in each area. Models whose symbol falls below the line have a positive storage change for the time period, also described in the NLDAS spin-up paper by *Cosgrove et al.* [2003b].

blue and storage changes are red. We further divided runoff into surface and subsurface runoff, and storage change into upper and lower zone storage change (see figure caption). In general the monthly storage changes are of the same magnitude as monthly runoff, and about half as large as monthly evapotranspiration.

[29] For the SE area the Mosaic model has the largest soil water storage changes within its annual cycle, associated with a higher evapotranspiration rate in July and August than any other model, about 20% of its total evapotranspiration. This is surprising given that Mosaic used a constant root depth of 0.4 m for all vegetation types, as compared to 1–2 m in Noah and 1.35–3 m in VIC. The subroot zone of the Mosaic model, which extends from 0.4 m to 2 m, accounts for most of this storage change. The only upward water transport between the layers in the Mosaic model is upward diffusion. The model physics (upward diffusion) therefore seems to counterbalance the model setup and vegetation parameters (constant 40 cm root layer in the two upper soil layers). Also, the Mosaic model has a much larger positive storage change in the SE in January, almost 80 mm, as compared to 10 mm in the VIC model, 20 mm in the NOAH model, and 40 mm in the Sacramento model. This corresponds to variations between 5 mm and 40 mm for ET and between 30 mm and 120 mm in runoff. The spread amongst the models in these monthly water balance terms is of the same magnitude than the water balance terms themselves.

[30] Although the mean annual evapotranspiration (ET) and runoff of the Mosaic and the Sacramento model are almost identical (Figures 5 and 6), their interannual distribution of evapotranspiration is very different. The winter ET from the Sacramento model is significantly larger compared to all other models, specifically in the SE. This might be because the Sacramento model does not impose restrictions on the transpiration based on the commonly used *Jarvis* [1976] or *Sellers et al.* [1986] approach, which limits transpiration dependent on the atmospheric (air temperature and humidity, solar radiation) conditions like all the other models. Another possible reason for the difference is that the Sacramento model can always evaporate from both its soil moisture storages, even under bare soil conditions or in the cold season with little vegetation, while all other models allow direct evaporation only from the top layer or have limited access to deeper reservoirs because of sparse vegetation coverage. Also, the Noah PE is known to be higher than the NOAA PE, which is used by the River Forecast Centers (RFC) in operations, causing a higher ET and less runoff. The VIC model has the lowest ET in all areas from November to April, for example, 20 mm in April in the NE area versus more than 80 mm in the Sacramento, and 50 mm in the Noah and Mosaic model. On the other hand, in the summer months, VIC's ET is not smaller than the ET from all other models in the NW and SW regions. Throughout the year the monthly storage changes of the Noah and the VIC model are very similar in all regions. Therefore the most significant seasonal difference between the Noah and the VIC model is the partioning of water from precipitation and storage change into evapotranspiration and runoff.

[31] Two parallel VIC modeling efforts were carried out as part of the NLDAS project. *Maurer et al.* [2002] performed a 50-year retrospective LSM run over the NLDAS domain, at a 1/8th degree spatial and 3-hourly temporal resolution. E. F. Wood and colleagues at Princeton University performed the VIC real-time NLDAS runs, which are analyzed in this and other NLDAS papers in this special section. The real-time VIC NLDAS runs use essentially the same parameters as *Maurer et al.* [2002]. One significant difference in the Maurer et al and real-time VIC

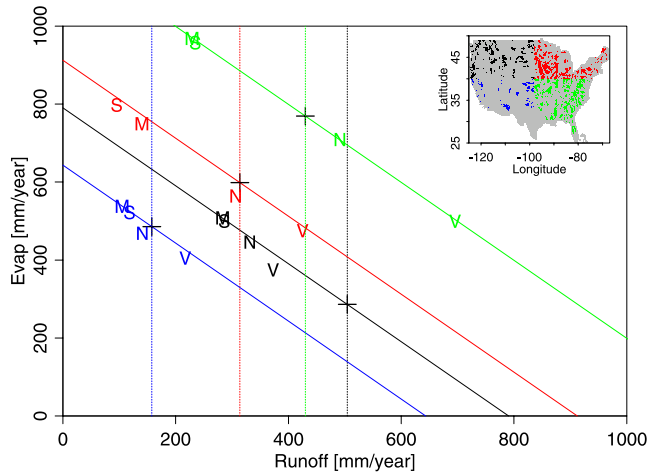

Figure 8. Partitioning of precipitation of the four NLDAS models (N, Noah; V, VIC; M, Mosaic; S, Sacramento) in all basins that fall into the four different quadrants for October 1997 to September 1999. The partitioning of precipitation is quite similar to Figure 7. Each vertical line is the averaged measured runoff from the 1145 small basins within the NLDAS area. The runoff in the northeast varies by a factor of 4 between VIC and the Sacramento model. All models underpredict runoff in the northwest.

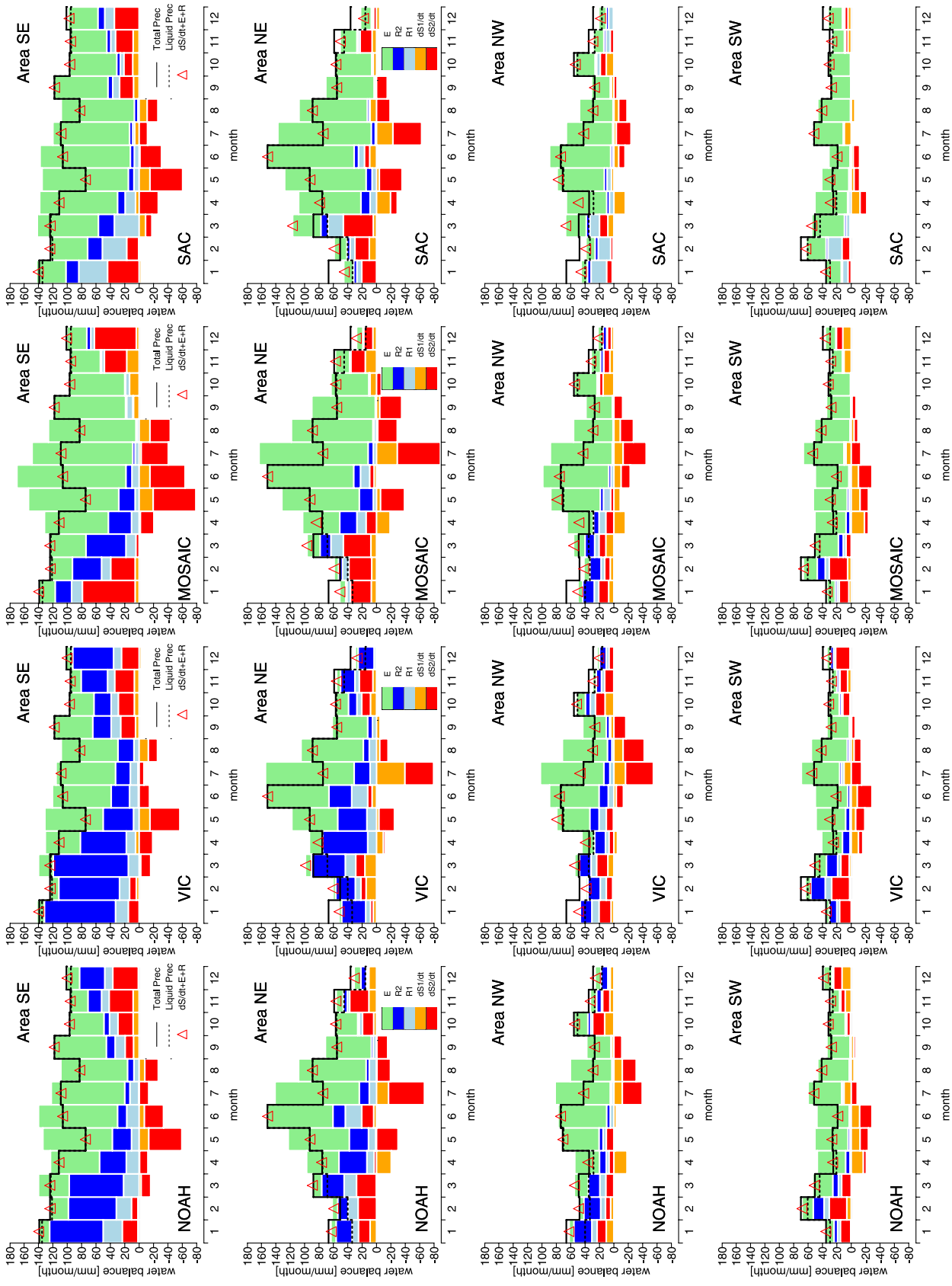
Area SE
Area NE
Area NW
Area SW
NOAH
VIC
MOSAIC
SAC
water balance [mm/month]
month
Total Prec
Liquid Prec
dS/dt+E+R
E
R2
R1
dS1/dt
dS2/dt

Figure 9.

runs is that the 50-year runs were performed at 3-hour time step, and used an equal partitioning of the daily gridded station data into 3-hour time intervals within the day. *Maurer et al.* [2002] analyzed the impact of the equal distribution of daily precipitation within 3-hour time steps as opposed to a more realistic disaggregation scheme based on observed hourly precipitation. The results (Figure 2 of Maurer et al.) show that runoff amounts were larger on the order of 10% for the model run with stochastically disaggregated precipitation in the lower Mississippi River basin for 1996–1999. Also, comparisons between the retrospective and real-time VIC runs show that the impact of the temporal disaggregation, and the impact of another difference in the runs, namely spatial disaggregation of precipitation, which was implemented in the real-time runs, but not by Maurer et al., can have a significant effect. This apparently has to do with (a) the difference between 3-hourly time steps, used by Maurer et al., and hourly time steps in the real-time runs; and (b) interactive effects of temporal and spatial disaggregation of precipitation. The differences (shown for a transect across the eastern and central U.S. at www.hydro.washington.edu/Lettenmaier/Models/VIC/VIChome.html) are most evident in portions of the country with a high fraction of convective precipitation and full canopy cover. *Maurer et al.* [2002] point out that the model could have been re-calibrated for the stochastically disaggregated precipitation run, resulting in equal evapotranspiration and runoff amounts compared to the run with daily evenly distributed precipitation. This clearly points to the interdependence of model setup and parameter choice, and to the lack of model parameter robustness. A more detailed examination of the differences will be conducted.

[32] From Figure 9 we can also see the influence of snow processes in the models. The Noah model has the smallest snow storage change. Since snow can only either sublimate or melt in the Noah model (no horizontal transport), the only explanation is that the Noah model melts snow earlier than all the other models, since its evapotranspiration (which includes sublimation) in the winter months is significantly smaller than the snowfall. Also, as *Mitchell et al.* [2004] show, the Noah model also has the highest sublimation of all models. The reason for the early Noah snowmelt is mainly the low albedo values in grid boxes with snow cover. This leads to a positive feedback, since a lower albedo will increase the available energy at the surface, and therefore increase snowmelt and sublimation. Less snow cover will then in return cause an even lower albedo.

[33] Other noticeable differences between the models are that (1) the Sacramento model produces dominantly surface runoff, while all other models produce mainly subsurface runoff; (2) the variability of relative differences between the models of total storage change is much larger in the western part of the United States than the eastern part. In the NW, VIC's total storage change is about 4 times larger than SAC's. In the SW the storage change of SAC is about one third of the storage change of the other models. This might be because ET in the western parts of the United States is limited by the availability of precipitation. The Sacramento models capability to evaporate more during the cold season results in a seasonality of ET that follows much more closely the seasonal precipitation curve, since much less water gets stored into the soil that could be available for ET during the summer months.

[34] Overall the results show that the modeled mean values of the water balance terms are of the same magnitude as the spread of the models around them. Main areas of model differences seem to be (1) the interaction of model physics with model setup, especially the large upward diffusion of the lower zone into the root zone in Mosaic; (2) seasonality of evapotranspiration, the Sacramento model has a high cold season ET and VIC a low ET; (3) early snowmelt with feedback in the Noah model; and (4) different utilization of soil water storage for seasonal cycle of ET in areas where ET is limited by precipitation.

5.2. Small-Scale Runoff Validation

[35] Figure 10 shows the observed (black curve) and modeled streamflow for the Nehalem River near Foss in Oregon (USGS code 14301000) for all four NLDAS models. In this basin all models have relatively low biases (less than 5%) and high correlation (R) values. Noticeable differences between modeled and observed data are the high base flow of VIC during the summer. There are about 20 basins within the NLDAS area for which all four models perform similarly well. Figure 11 shows the derived lumped unit-hydrograph (from equation 2) for the 4 models, which minimizes the least square difference between modeled and measured data for each simulation. It was calculated using the iterative deconvolution technique explained in detail by *Lohmann et al.* [1998a, 1998b] with modeled runoff instead of effective precipitation (see equation (2)). This unit-hydrograph represents the best linear lumped routing procedure for this catchment for each model and takes into consideration the different runoff production mechanisms of the four models. It is a measure of the distribution of the residence time of surface water in the catchment after it has been produced as runoff from the LSM. The different hydrographs for each model can be explained as follows: the Sacramento model and the VIC model produce more surface runoff than other two models. To match the measured flow, they therefore need to keep this runoff longer in a horizontal routing model. This interplay between runoff production and horizontal transport is often neglected. We used this optimization procedure for all 1145 basins. While for most basins the resulting model predictions improve as compared to a distributed

Figure 9. Monthly water balance of the NLDAS models in the four quadrants of Figures 3 and 4 for the time period October 1997 to September 1998. Orange, upper soil storage change; red, lower soil storage change; light blue, surface runoff; dark blue, subsurface runoff; green, evapotranspiration; black solid line, precipitation; black dotted line, liquid precipitation; red triangles, storage change + evapotranspiration + runoff. The deviation of the red triangles from the solid black line indicates snow processes. A red triangle is as much above (below) the solid black line as snow melts (accumulates).

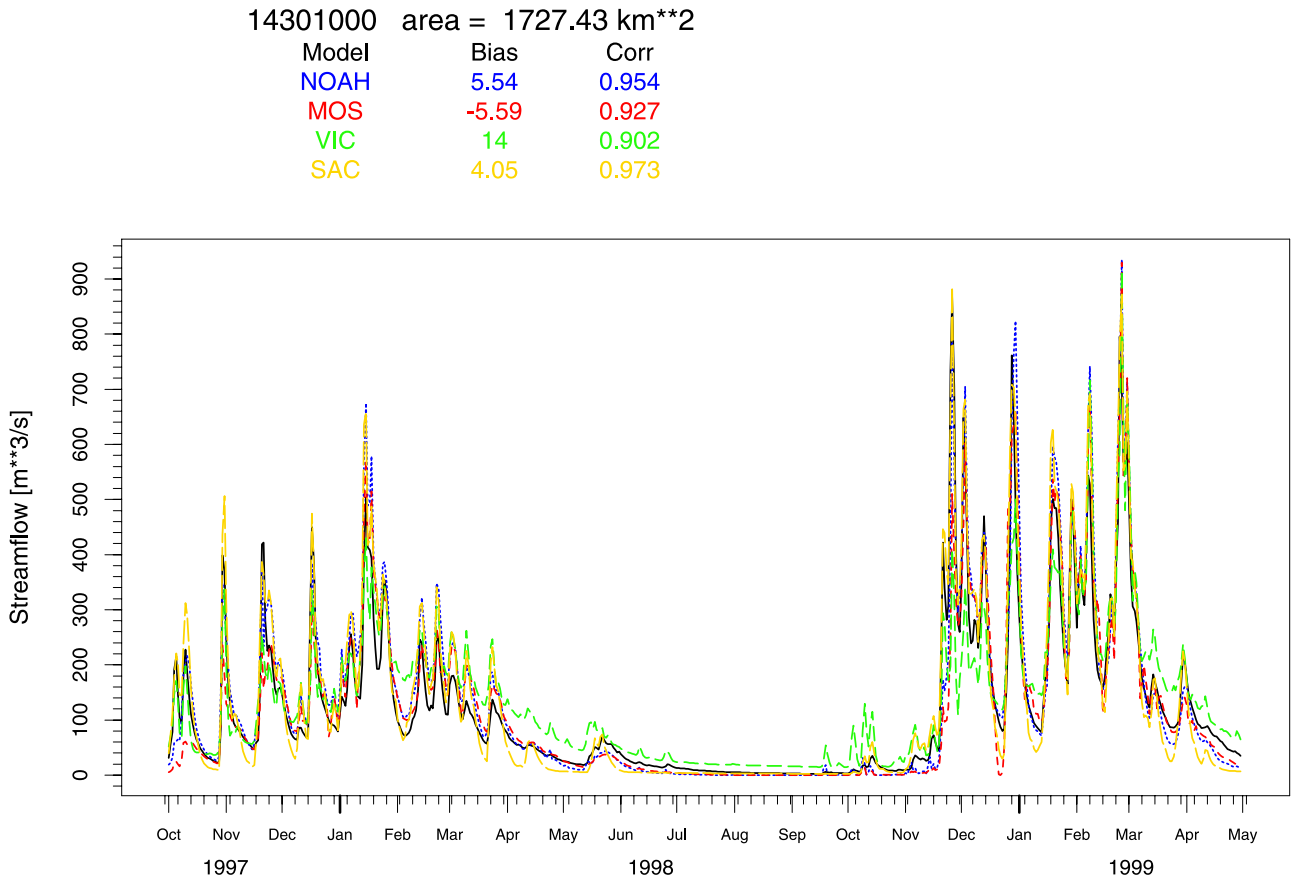

Figure 10. Observed (black curve) and modeled streamflow for the Nehalem River near Foss in Oregon. Noah (blue), VIC (green), Mosaic (red), and SAC (yellow). The bias is in $m^3\ s^{-1}$. Corr is the correlation coefficient.

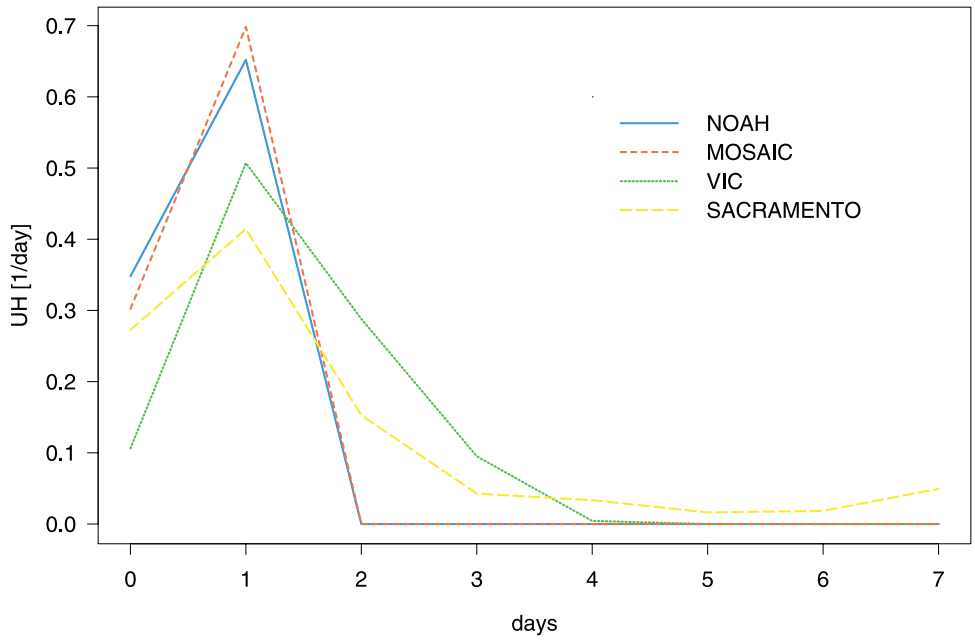

Figure 11. Derived unit-hydrograph for the four NLDAS models for the Nehalem River near Foss in Oregon. Noah (blue), VIC (green), Mosaic (red), and SAC (yellow).

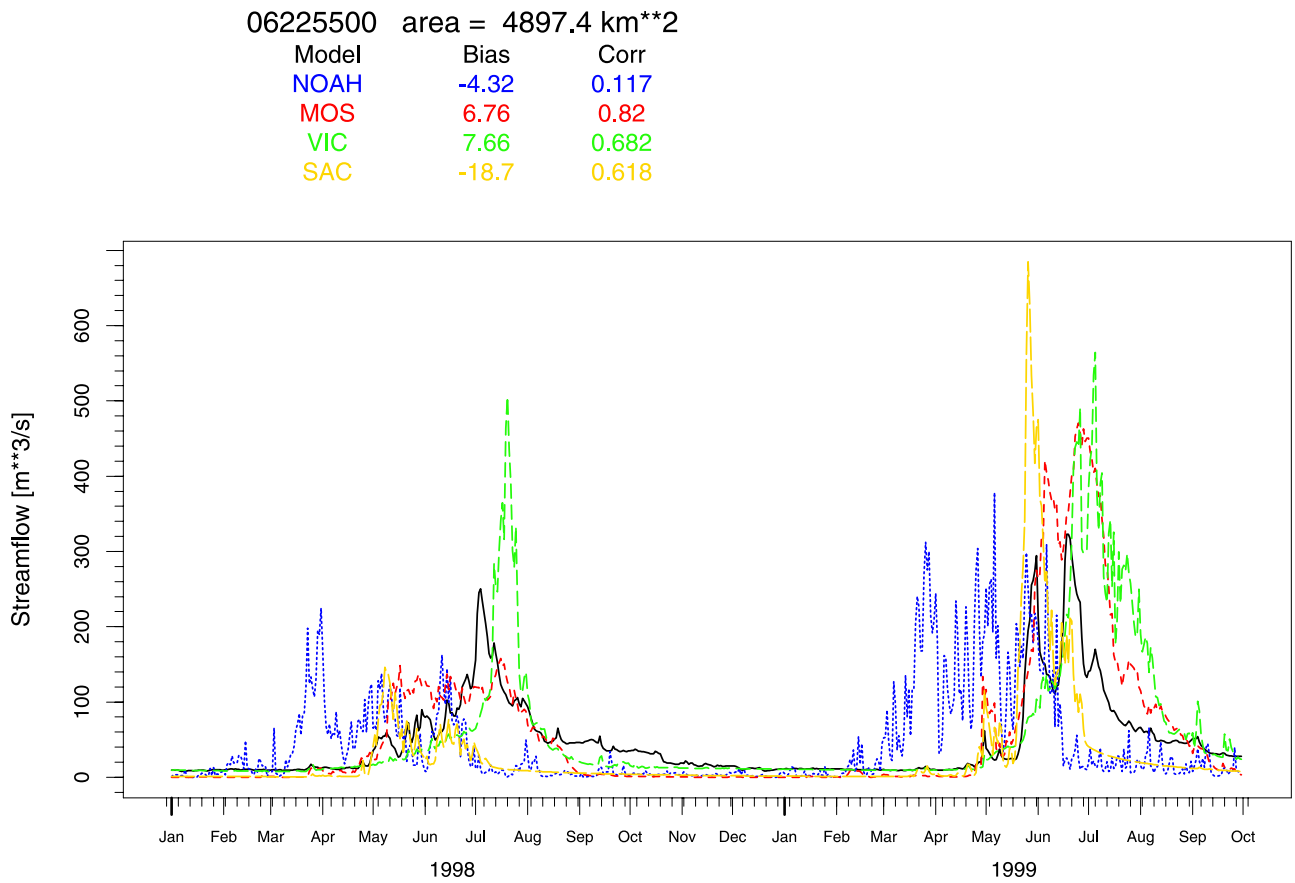

Figure 12. Observed (black curve) and modeled streamflow for the Wind River near Crowheart, Wyoming. The station is 1718 m above sea level and shows the impact of snowmelt on runoff. Noah (blue), VIC (green), Mosaic (red), and SAC (yellow). The bias is in $m^3\ s^{-1}$. Corr is the correlation coefficient.

model with default parameters, the iterative deconvolution scheme failed whenever there was no parsimonious unit-hydrograph to transform modeled runoff into streamflow. This mainly occurred in areas with significant snowfall or low runoff ratios. In these cases the distributed runoff routing model was used to calculate the error statistics.

[36] To illustrate the effect of snowmelt on runoff timing, Figure 12 shows the observed (black curve) and modeled streamflow for the Wind River near Crowheart, Wyoming (USGS code 06225500). The streamflow measurement station is 1718 m above sea level and the runoff timing is therefore highly influenced by snowmelt. The Noah model starts melting snow in early March and has melted the entire snow pack by the end of June. The other models start melting the snow significantly later. SAC and Mosaic start melting their snow late April and early May for both years, but Mosaic melting period lasts until the end of August, while SAC has a shorter melt period until the end of June. The VIC model starts melting about the same time as SAC and Mosaic, but starts slower and has its peak from snowmelt runoff in July, about three weeks later than the observed streamflow. VIC's snowmelt period lasts about as long as Mosaic's. Previous model experiments by *Boone et al.* [2004] showed that VIC's snowband parameterization not only provided more realistic simulations of snowmelt, but that it was also almost scale independent. The main reason for the early snowmelt of the Noah model is the low albedo over snow-covered areas, as mentioned earlier. The papers by *Mitchell et al.* [2004], *Pan et al.* [2003], and *Sheffield et al.* [2003] in this special section investigate this further.

[37] It is also important to note that *Pinker et al.* [2003] found a substantial high bias in NLDAS solar insolation over areas with winter snow cover. The low NLDAS precipitation bias over the Northwest combined with the high solar insolation bias suggests that the Noah model would improve its snowmelt timing with revised forcing data and that also the Mosaic and the VIC model would melt the snow even later in the season.

[38] Figure 13 shows the relative bias of the mean annual runoff of all four NLDAS models for the time period from 1 October 1997 to 30 September 1999. All models underpredict the mean annual runoff in the northern Rocky Mountains in most basins by 20% to around 80%. The main reason for this can be found in the papers by *Sheffield et al.* [2003] and *Pan et al.* [2003] in this special section. They show that for 110 screened SNOTEL stations within the NLDAS area the NLDAS precipitation forcing is more than 50% too low compared to station measurements. They showed that by increasing the amount of precipitation by a constant factor of 2.1693 (from regression analysis) most of the errors in the snow water equivalent can be reduced

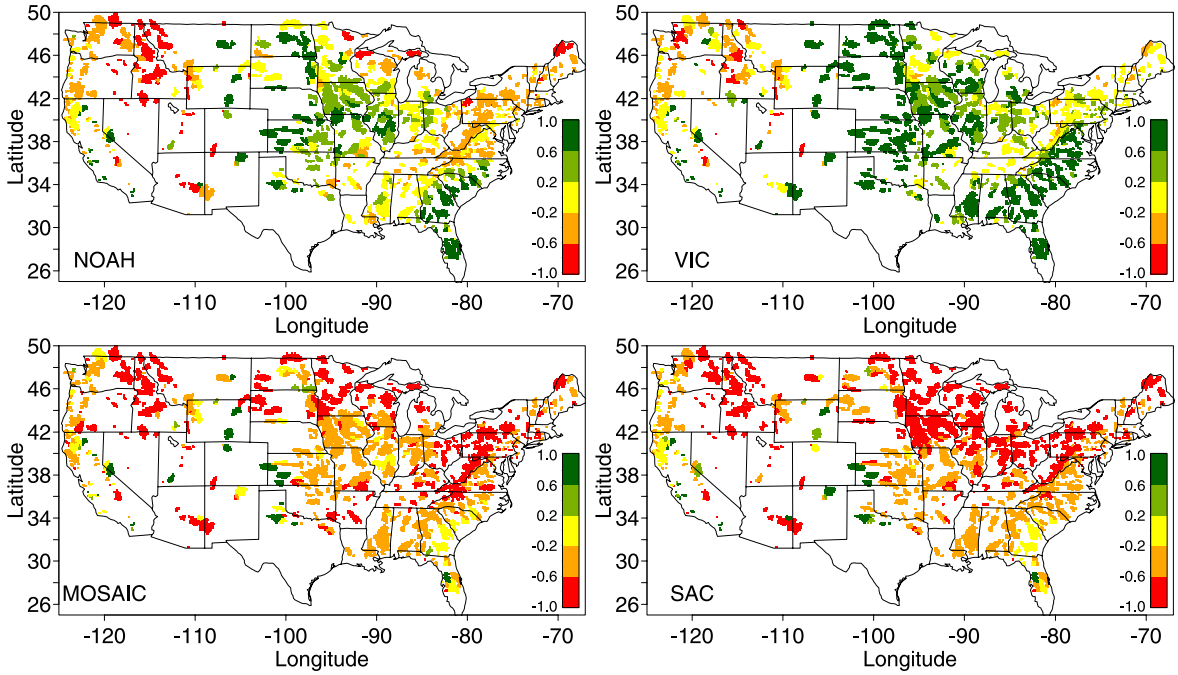

Figure 13. Relative runoff bias for the four NLDAS models for the time period 1 October 1997 to 30 September 1999. Notice that all models show less runoff than observed in the western snow-covered areas.

significantly. However, other potential forcing errors (e.g., solar radiation) also would need to be addressed.

[39] In the eastern part of the United States all models show a similar gradient of relative biases. The relative biases increase from north toward south. The VIC model calculates the right annual runoff correctly in the northeast, but produces too much runoff going southward. The Noah model underpredicts in the northeast, but overpredicts in the South, and is correct in-between. The SAC and Mosaic model have the smallest absolute bias in the southeast corner, but underpredict total annual runoff northward.

[40] The VIC model also overestimates runoff by more than 60% in the southeast and Midwest, but has the lowest bias of all models in the Rocky Mountains. The Noah model has a similar spatial structure of relative bias gradients; however the biases tend to be smaller and of either sign, resulting in the lowest regional bias in NE, SE and SW as shown in Figure 8. The Mosaic and the SAC model have very similar patterns of relative runoff biases. Both consistently underpredict runoff throughout the NLDAS area but the Southeast corner. The only overprediction of all models can be found basins in North Texas, New Mexico, and Oklahoma, where all models show too much runoff. The reasons for this have not been investigated in this paper, but one possible reason could be irrigation and the influence of farming. These effects are not included in the current NLDAS setup.

[41] To gain more insight into seasonal differences between the models, we computed the cold (Figure 14) and the warm season (Figure 15) runoff biases of the models. The Noah model produced the correct amount of runoff in many basins in the east to the Midwest in the cold season, but shows the same north-south gradient as in the relative annual bias. The pattern looks fairly similar in the warm season. The most noticeable difference is the warm season bias in the Midwest and the negative bias in the Rocky Mountains. The VIC model also shows a similar spatial pattern during the cold and the warm season, but with a larger positive bias in the Midwest during the warm season. VIC's runoff during this time is mainly subsurface flow. The Mosaic and the Sacramento model also show in their seasonal pattern in the east the same north-south gradient as in the annual relative bias pattern. However, the two models have a distinctly different seasonal characteristic. While the Mosaic model produces less runoff in the eastern region in the cold season than the Sacramento model does, especially in the southeast, this role gets reversed in many small basins during the warm season. Together with Figure 9 we can therefore make the observation that the in the east during the warm season the Mosaic model has more evapotranspiration and runoff than the Sacramento model, but less evapotranspiration and runoff during the cold season. The total annual runoff however is about the same. The reason for this can be found in Figure 9, where it can be seen that in the SE the total storage change for the warm season of the Mosaic model exceeds the Sacramento model storage by more than 90 mm, which is about one quarter of the total annual runoff for this area (Figure 8) and almost half the total annual runoff amount produced by the two models. *Schaake et al.*'s [2004] investigation of soil moisture results from the NLDAS models over 17 sites in Illinois showed that the Mosaic model had a storage range that was about 50% larger than observed, with the other models being closer to the observations. This is consistent with a similar observation from *Robock et al.* [2003] for 72 sites in Oklahoma. The soil moisture anomalies of the Noah and the VIC were closest to the observations, the Mosaic model had a much larger than observed soil moisture anomaly, and the

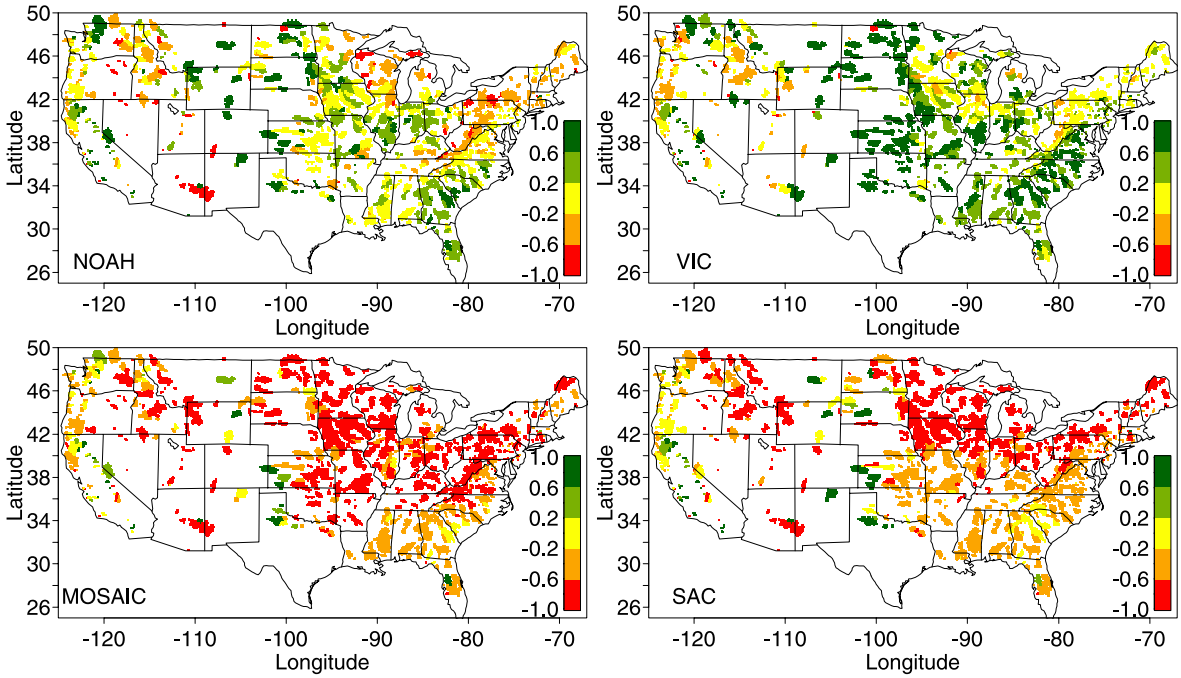

Figure 14. Cold season relative bias (October-March) in runoff for the time period 1 October 1997 to 30 September 1999.

Sacramento models response looked overamplified for individual precipitation events.

[42] Figure 16 shows the Nash-Sutcliffe efficiency for daily modeled streamflow for all basins for the time period 1 October 1997 to 30 September 1999. All models share the same spatial structure of efficiency in the eastern United States, with the SAC and the Noah model being slightly better than the other two models for most basins. The Mosaic and the SAC model have the highest efficiency scores in the northern Midwest, while the VIC model has the highest efficiency of all the models in the Rocky Mountains and the northern part of the east. Despite being almost equal in total annual runoff, the Sacramento model has a smaller error variance than the Mosaic model. This is consistent with the relative seasonal runoff biases of both models. The Sacramento model is closer to the observed in

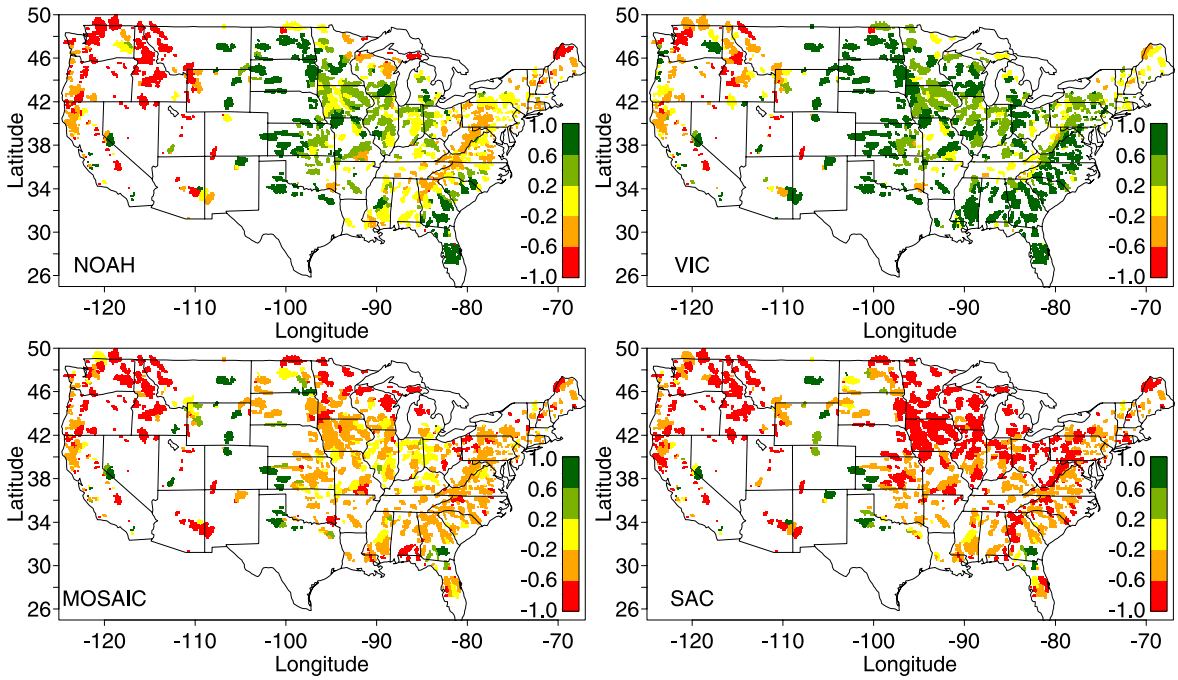

Figure 15. Warm season (April-September) relative runoff bias for the time period 1 October 1997 to 30 September 1999.

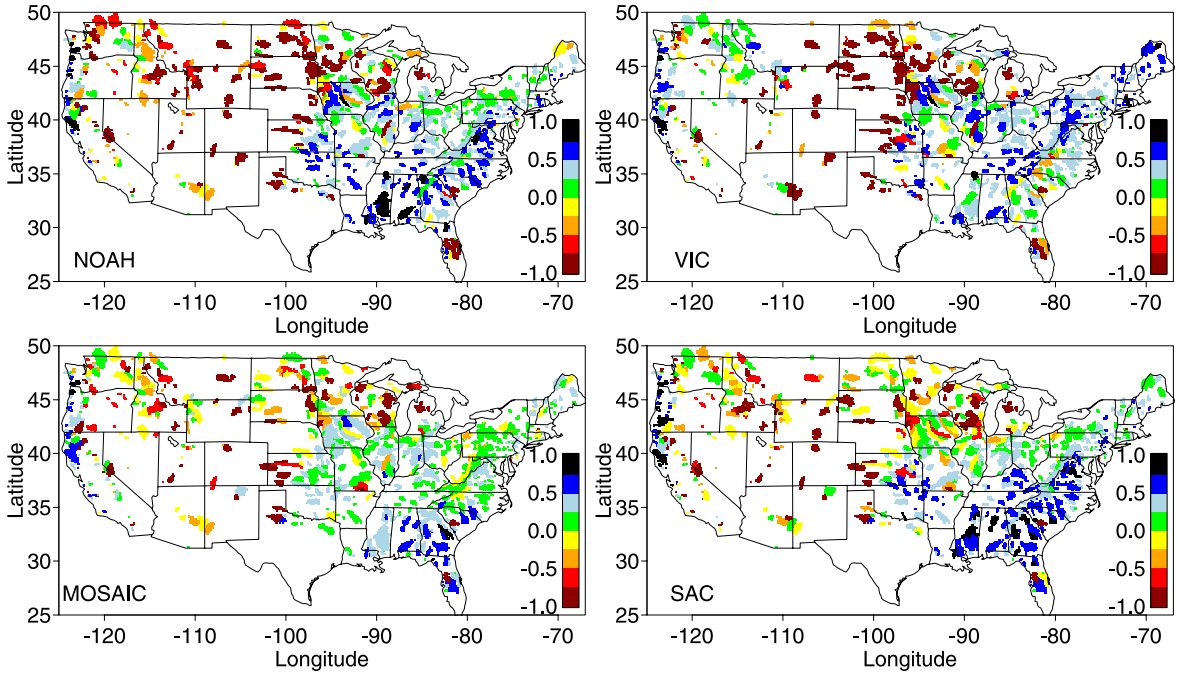

Figure 16. Nash-Sutcliffe efficiency for the NLDAS domain for the time period 1 October 1997 to 30 September 1999 for daily mean modeled and measured data.

the cold season, where most of the variance of streamflow is observed.

[43] The relationship between mean annual snowfall and the correlation between modeled and observed streamflow is shown in Figure 17. All models show their lowest correlation in basins with high snowfall, and noticeably the Noah models correlation decreases dramatically for basins with more than 100 mm/yr snowfall. This indicates that more efforts are needed to produce reliable forcing data sets for these areas, as well as the need for intensified research for large-scale snow models.

[44] To quantify the timing of the modeled streamflow, we computed the cross-correlation function between measured and observed streamflow data. Figure 18 shows the number of days from time zero to the maximum of the cross-correlation function. Negative numbers indicate the number of days that the modeled streamflow peaked before the observed streamflow. Positive numbers show the number of days that the model lagged behind the observation. Basically, over most of the country the models reproduce the streamflow peaks within plus or minus one day. This is also due to the fact that we optimized the routing model for large parts of the country. In a few basins we can see that all models have timing problems. We assume that the reason for this is either streamflow regulation or that the modeled and measured streamflow time series are autocorrelated and pick up the wrong maxima. For the Rocky Mountains and the northeast, the cross-correlation function clearly indicates model trends for snow-covered areas. The Noah model has many of its peak streamflows more than 2 months prior to the event. Mosaic and the Sacramento model have errors of about one month for many basins, while the VIC model seems to model runoff timing (and therefore snowmelt timing) very well, though sometimes it predicts snowmelt too late, as was also seen in Figure 12. These results are consistent with the NLDAS papers by *Sheffield et al.* [2003] and *Pan et al.* [2003] in this special section and also with results from the PILPS 2(e) [*Bowling et al.*, 2003; *Nijssen et al.*, 2003], and the Rhone GSWP experiment [*Boone et al.*, 2004], in which the VIC and Noah model participated.

[45] Figure 19 was inspired by the work of *Oki et al.* [1999]. They showed that for the Global Soil Wetness Project (GSWP) there was a high correlation between runoff biases and the precipitation station density. Note that the station density in the NLDAS project is about an order of magnitude larger than in the GSWP project. We used the average station density for July 1997 to compute the station density. Each precipitation station in grid cells that had more than 30% of their area within one basin was counted as a station for that basin. The resulting pattern in the NLDAS project is not as prominent compared to the GSWP pattern. It is possible that the biases are better explained by model physics (specifically snow, evapotranspiration and runoff parameterization) and precipitation amounts in mountain and snow-covered areas than by the density of the precipitation network itself. The red circles in the figure are the basins with more than 100 mm/yr snowfall. For all models the majority of the red circles show a clear negative bias in modeled runoff. For the basins with annual snowfall equal or less 100 mm/yr, the Noah model has the smallest overall bias.

5.3. Large-Scale Runoff Validation

[46] Figure 20 shows the location of the 9 large basins and their gauging stations that were used for the validation. The corresponding mean monthly streamflow is shown in Figure 21. The results for the major U.S. rivers are consistent with the analysis for the 1145 small basins, and also show new features in the western part. These large basins can be seen as integrators of all the headwater systems

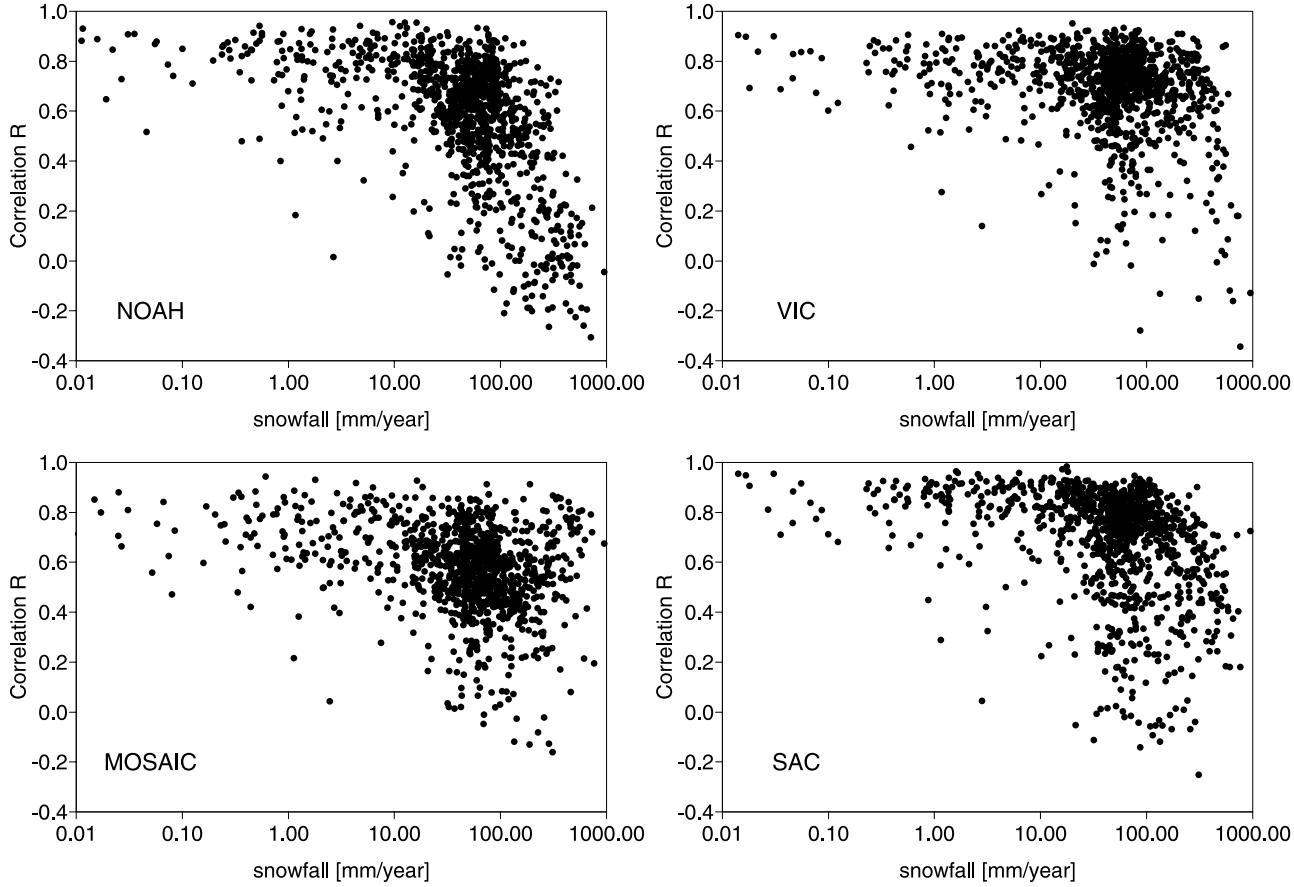

Figure 17. Relationship of mean annual snowfall in mm/yr and the correlation of simulated and observed runoff. Each of the circles represents one of the 1145 basins.

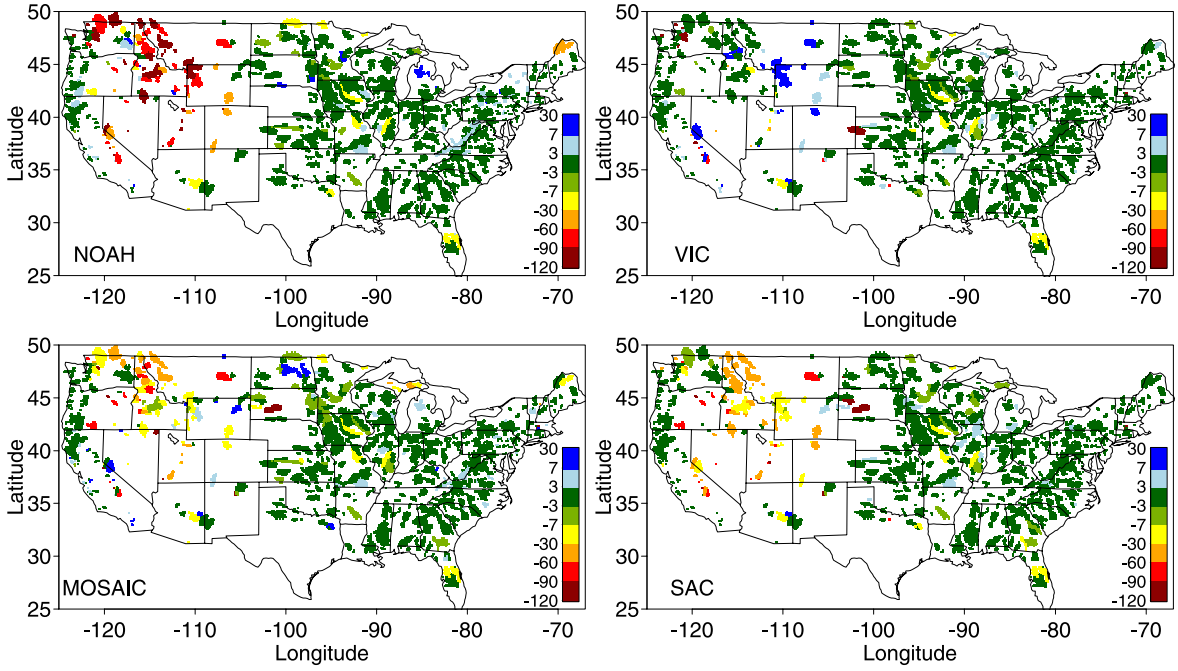

Figure 18. Time to peak delay (days) of streamflow for 1145 basins calculated as the maximum of the cross-correlation function of modeled and observed streamflow. Negative numbers indicate streamflow peaks earlier than observed peaks. In snow-covered areas, Noah, Mosaic, and the Sacramento model consistently show too early snowmelt and therefore produce streamflow too early [see *Pan et al.*, 2003].

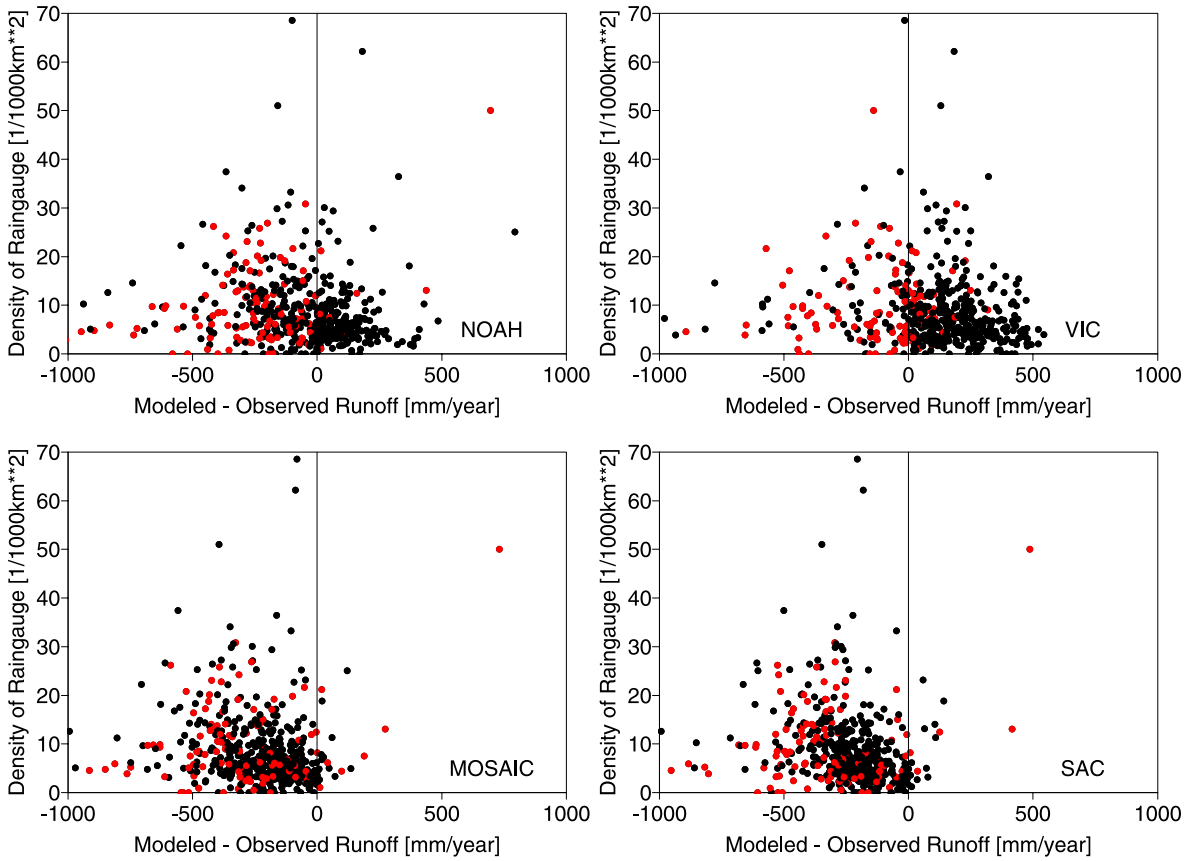

Figure 19. Density of the observing rain gauges as a function of the annual runoff bias for the time period 1 October 1997 to 30 September 1999. Basins with more than 100 mm/yr of snow are shown as red circles. This analysis was done for the first time for the GSWP experiment [*Oki et al.*, 1999], and that analysis showed that the absolute runoff bias was a function of the gauging station density. That relationship is also observed here, but to a much lesser degree.

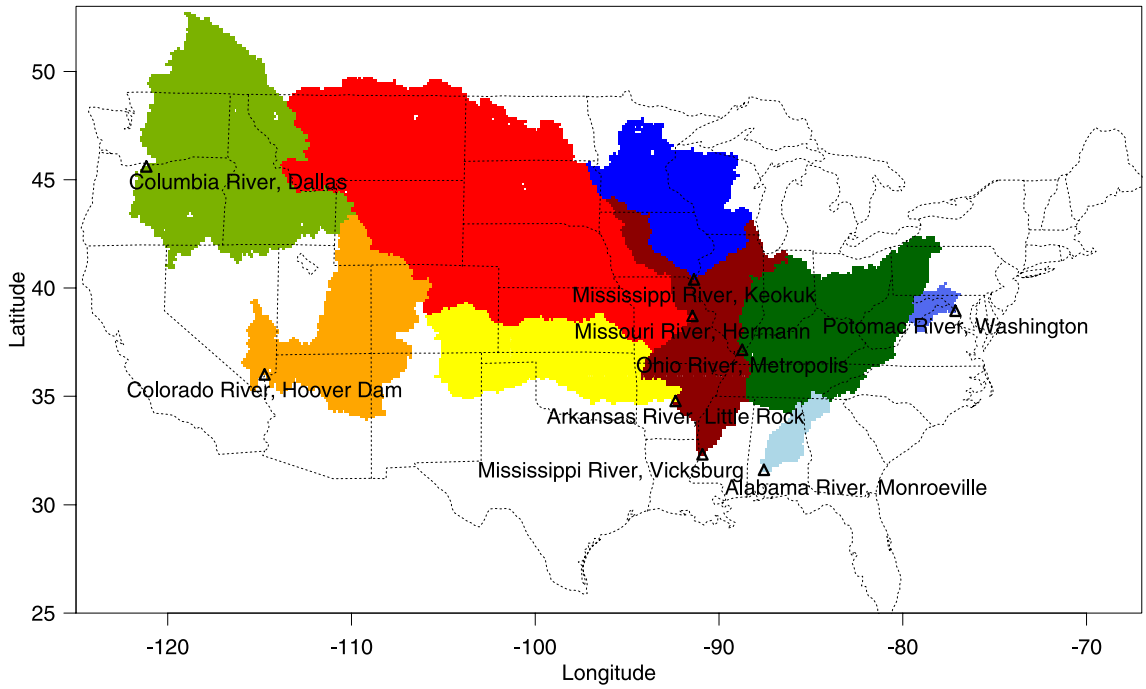

Figure 20. Location of the nine major basins and USGS gauging stations used for this intercomparison and validation study within the United States.

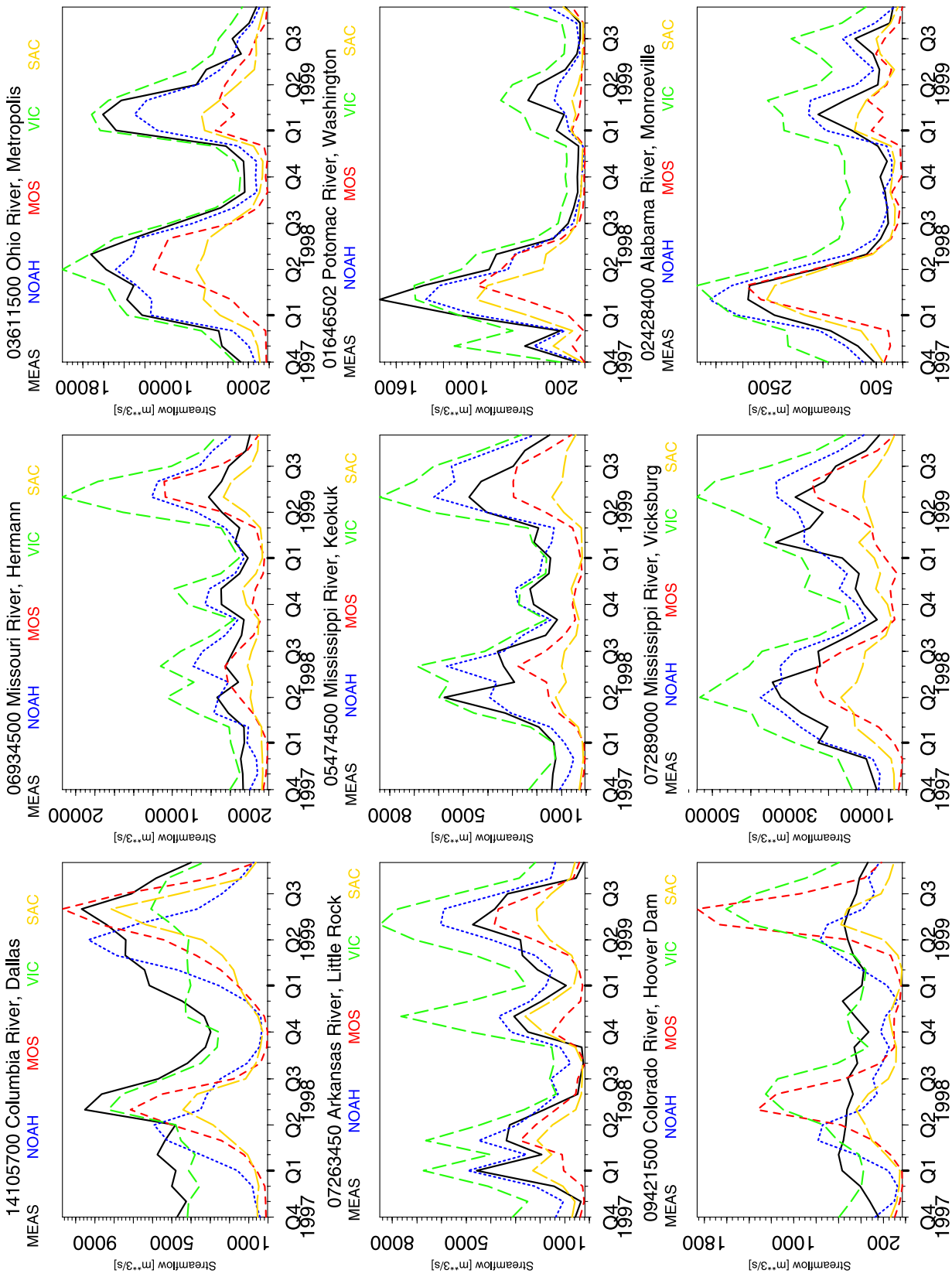

Figure 21. Monthly mean streamflow for the nine major basins from Figure 20 for the time period October 1997 to September 1999. Noah (blue), VIC (green), Mosaic (red), and SAC (yellow).

upstream. However, it is unclear whether our modeling efforts need to include groundwater systems that are larger than the current grid box in the NLDAS setup. None of the models include such a horizontal water transport.

[47] The largest differences between the modeled and the measured runoff are visible in the strongly regulated basins (Columbia, Arkansas, Colorado, and Missouri). All show a smaller seasonal signal than the modeled runoff. Previous modeling studies therefore used naturalized streamflow data [e.g., *Maurer et al.*, 2002; *Lohmann et al.*, 1998a, 1998b], which are reconstructed time series were the influence of dams and reservoirs is removed. There are two major features of the modeled streamflow in the Columbia and Colorado River though. The first is the difference in runoff timing between the models. Consistent with earlier results from the small-scale basins the Noah model has a much earlier runoff production in the snowmelt season than the other models and the observations. The VIC model has a much more sustained runoff from August to January, when all other models have their streamflow minimum. Similar to *Nijssen et al.* [1997], this might be caused by the choices of the base flow parameters and the model layer thicknesses. The second feature is the runoff amount of the models. The Mosaic model exceeds the Noah model's streamflow and produces almost as much as the VIC model. One explanation might be again Noah's low albedo in snow-covered areas that results in an earlier snowmelt and a higher sublimation. The water from the snowmelt then becomes available for transpiration. On the other hand, it also could be the interplay of the runoff parameterizations with soil moisture [*Koster and Milly*, 1997], where the Mosaic model produces more runoff at low soil moisture levels in these two basins, while Noah produces more runoff at higher soil moisture levels in the east.

[48] In all other less regulated basins the models capture the seasonality of runoff reasonably well. The VIC model produces too much runoff in almost all basins; however, the Ohio River basin is modeled fairly well. In these large eastern basins the integrative effect of the large areas seems to benefit the Noah model results, which showed a spatially varying underprediction and overprediction in the east, and therefore reflects the water balance of most major basins the best of all models. The Mosaic and the Sacramento model consistently underpredict streamflow in the eastern region. Also, the Mosaic model runoff has a small phase lag to the Sacramento model, consistent with SAC's larger runoff amounts in the cold season.

6. Conclusions

[49] The NLDAS project was initiated to foster the development of data assimilation of land surface and hydrology models that can be coupled to atmospheric models. One of the early milestones is the off-line test and validation of these models. An hourly forcing data set from October 1996 to September 1999 was used to drive four different land surface models over the continental United States on a common 1/8° latitude-longitude grid. The models used a priori parameters for these runs. This paper presented results from the streamflow validation and water balance intercomparison. The results are not necessarily encouraging; they show that we cannot model streamflow in most basins within the United States without more work done on (1) parameter estimation and calibration techniques; (2) a better understanding of the interplay between model structure, model parameters and model setup; (3) quantifying the effect of temporal and spatial disaggregation of precipitation; and (4) creation of bias corrected input data. The intermodel differences of the water fluxes have the same magnitude as the mean modeled or measured water fluxes. Regional differences between models are significant, up to a factor of 4 for mean annual runoff and a factor of 2 for evapotranspiration. For monthly water balance terms these relative differences were even larger. It is possible to explain some of the model results by the models physics and model setup, which might lead to model improvements. Some candidates for further research are: Mosaic's high upward soil moisture transport resulting in large soil water storage changes that influence the other water balance terms; Noah's low albedo over snow leading to a positive feedback mechanism with snowmelt; Sacramento's cold season evaporation and evaporation from two storages which might lead to problems in arid regions; VIC's (and possibly other models as well) sensitivity to the parameterization of spatial disaggregation of precipitation and the length of the time step. Also, the influence of using forcing data from different sources with different climatologies needs to be examined. This might have diminished the performance of the Sacramento model, which is the most flexible model among the four NLDAS models when it comes to the parameterization of runoff processes and has proven that it can reproduce streamflow.

[50] Results from other NLDAS studies were confirmed that investigated the bias of the precipitation data over mountainous regions. In the current realtime NLDAS setting [*Mitchell et al.*, 2000], unlike the retrospective runs here, we use the PRISM climatology to interpolate precipitation spatially, which might help us with the low bias in the snow-covered areas and data sparse regions in the west. The VIC model showed the best snowmelt timing. This is consistent with results from the Rhone experiment [*Boone et al.*, 2004]. It therefore seems to be advisable that other models adopt VIC's approach and introduce elevation bands into their models.

[51] Parsimonious parameter estimation routines need to be implemented over large spatial areas to conduct these experiments. We should be able to reduce the spread amongst the models significantly and move closer toward the observations by calibrating model parameters. Considering the many off-line tests in recent years, and operational or calibrated runs in basins all over the United States, the models performance is disappointing; our a priori parameter estimations have to be refined and become more robust. However, it should be noted that the model setup for the Sacramento model and the Mosaic model where different from their standard configuration, and therefore the models might require adjusted parameter values. VIC's model results showed us the sensitivity of the model when new physical parameterizations (or model updates) are introduced.

[52] This model comparison is not necessarily fair and is also not complete. We can only investigate how the models were able to perform with their a priori parameters in this specific NLDAS setup. That does not allow us to rank the models and to conclude that one particular model is better

than another. A fair comparison would allow all modelers to objectively calibrate their model on the basis of given objective functions. Also, we would like to gain knowledge about the parametric uncertainty of each model since currently we cannot perform any type of error analysis. Running and calibrating the models multiple times with the NLDAS data, the PILPS 2(c) experiment [*Wood et al.*, 1998], the PILPS 2(d) experiment at Valdai, Russia [*Schlosser et al.*, 1997, 2000; *Slater et al.*, 2001; *Luo et al.*, 2003a], the PILPS 2(e) experiment [*Bowling et al.*, 2003; *Nijssen et al.*, 2003], the Rhone GSWP experiment [*Boone et al.*, 2004], and similar off-line data should be beneficial.

[53] **Acknowledgments.** The work on this project by NCEP/EMC, NWS/OHD, and NESDIS/ORA was supported by the NOAA OGP grant for the NOAA Core Project for GCIP/GAPP (co-PIs K. Mitchell, J. Schaake, and D. Tarpley). The work by NASA/GSFC/HSB was supported by NASA's Terrestrial Hydrology Program (P. Houser, PI). The work by Rutgers University was supported by NOAA OGP GAPP grant GC99-443b (A. Robock, PI), the Cook College Center for Environmental Prediction, and the New Jersey Agricultural Experiment Station. The work by Princeton was supported by NOAA OGP GAPP grant NA86GP0258 (E. Wood, PI). The work by NCEP/CPC was supported by NOAA/NASA GAPP Project 8R1DA114 (R. W. Higgins, PI). The work by University of Maryland was supported by grants NA56GPO233, NA86GPO202, and NA06GPO404 from NOAA/OGP and by NOAA grant NA57WC0340 to University of Maryland's Cooperative Institute for Climate Studies (R. Pinker, PI). We thank USGS for the streamflow data, which were provided to the project at no cost.

References

Anderson, E. A. (1973), National Weather Service River Forecast System: Snow Accumulation and Ablation Model, *NOAA Tech. Memo., NWS Hydro-17*, U. S. Natl. Weather Serv., Silver Spring, Md.

Boone, A., et al. (2004), The Rhone-Aggregation Land Surface Scheme Intercomparison Project: An overview, *J. Clim.*, *17*, 187–298.

Bowling, L. C., et al. (2003), Simulation of high latitude processes in the Torne-Kalix basin: PILPS Phase 2(e) 1: Experiment description and summary comparisons, *J. Global Planet. Change*, *38*, 1–30.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, technical report, 204 pp., Joint Fed. and State River Forecast Cent., U.S. Natl. Weather Serv. and Calif. State Dep. of Water Resour., Sacramento, Calif.

Chapelon, N., H. Douville, P. Kosuth, and T. Oki (2002), Off-line simulation of the Amazon water balance: A sensitivity study with implications for GSWP, *Clim. Dyn.*, *19*, 141–154.

Chen, F., K. Mitchell, J. Schaake, Y. Xue, H. Pan, V. Koren, Q. Duan, and A. Betts (1996), Modeling of land-surface evaporation by four schemes and comparison with FIFE observations, *J. Geophys. Res.*, *101*, 7251–7268.

Cherkauer, K. A., and D. P. Lettenmaier (1999), Hydrologic effects of frozen soils in the upper Mississippi River basin, *J. Geophys. Res.*, *104*, 19,599–19,610.

Cherkauer, K. A., L. C. Bowling, and D. P. Lettenmaier (2003), Variable infiltration capacity cold land process model updates, *Global Planet. Change*, *38*, 151–159.

Chow, V. T. (1959), *Open Channel Hydraulics*, McGraw-Hill, New York.

Cosgrove, B. A., et al. (2003a), Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project, *J. Geophys. Res.*, *108*(D22), 8842, doi:10.1029/2002JD003118.

Cosgrove, B. A., et al. (2003b), Land surface model spin-up behavior in the North American Land Data Assimilation System (NLDAS), *J. Geophys. Res.*, *108*(D22), 8845, doi:10.1029/2002JD003316.

Dirmeyer, P. A., A. J. Dolman, and N. Sato (1999), The Global Soil Wetness Project: A pilot project for global land surface modeling and validation, *Bull. Am. Meteorol. Soc.*, *80*, 851–878.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley (2003), Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, *J. Geophys. Res.*, *108*(D22), 8851, doi:10.1029/2002JD003296.

Entekhabi, D., et al. (1999), An agenda for land-surface hydrology research and a call for the Second International Hydrological Decade, *Bull. Am. Meteorol. Soc.*, *80*, 2043–2058.

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, and R. Sohlberg (2000), Global land cover classification at 1 km spatial resolution using a classification tree approach, *Int. J. Remote Sens.*, *21*, 1331–1364.

Henderson-Sellers, A., Z.-L. Yang, and R. E. Dickinson (1993), The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS), *Bull. Am. Meteorol. Soc.*, *74*, 1335–1350.

Higgins, R. W., W. Shi, and E. Yarosh (2000), Improved United States precipitation quality control system and analysis, *NCEP/Clim. Predict. Cent. Atlas 7*, 40 pp., Clim. Predict. Cent., Camp Springs, Md.

Jarvis, P. G. (1976), The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field, *Philos. Trans. R. Soc. London, Ser. B*, *273*, 593–610.

Koren, V., J. Schaake, K. Mitchell, Q.-Y. Duan, F. Chen, and J. Baker (1999), A parameterization of snowpack and frozen ground intended for NCEP weather and climate models, *J. Geophys. Res.*, *104*, 19,569–19,585.

Koren, V. I., M. Smith, D. Wang, and Z. Zhang (2000), Use of soil property data in the derivation of conceptual rainfall-runoff model parameters, paper presented at 15th Conference on Hydrology, Am. Meteorol. Soc., Long Beach, Calif., 10–14 Jan.

Koster, R., and P. Milly (1997), The interplay between transpiration and run-off formulations in land-surface schemes used with atmospheric models, *J. Clim.*, *10*, 1578–1591.

Koster, R. D., and M. J. Suarez (1996), Energy and water balance calculations in the Mosaic LSM, *NASA Tech. Memo., 104606*, vol. 9, 60 pp.

Lettenmaier, D. P., F. A. Abdulla, E. F. Wood, and J. A. Smith (1996), Application of a macroscale hydrologic model to estimate the water balance of the Arkansas-Red River basin, *J. Geophys. Res.*, *101*, 7449–7459.

Liang, X., and Z. Xie (2001), A new surface runoff parameterization with subgrid-scale soil heterogeneity for land surface models, *Adv. Water Resour.*, *24*(9–10), 1173–1193.

Liang, X., E. Wood, and D. Lettenmaier (1996), Surface and soil moisture parameterization of the VIC-2L model: Evaluation and modifications, *Global Planet. Change*, *13*, 195–206.

Lobmeyr, M., D. Lohmann, and C. Ruhe (1999), An application of a large scale conceptual hydrological model over the Elbe region, *Hydrol. Earth Syst. Sci.*, *3*, 363–374.

Lohmann, D., E. Raschke, B. Nijssen, and D. P. Lettenmaier (1998a), Regional scale hydrology, Part II: Application of the VIC-2L model to the Weser River, Germany, *Hydrol. Sci. J.*, *43*(1), 143–158.

Lohmann, D., et al. (1998b), The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) Phase-2(c) Red-Arkansas River Basin Experiment: 3. Spatial and temporal analysis of water fluxes, *J. Global Planet. Change*, *19*, 161–179.

Luo, L., et al. (2003a), Effects of frozen soil on soil temperature, spring infiltration, and runoff: Results from the PILPS 2(d) experiment at Valdai, Russia, *J. Hydrometeorol.*, *4*, 334–351.

Luo, L., et al. (2003b), Validation of the North American Land Data Assimilation System (NLDAS) retrospective forcing over the southern Great Plains, *J. Geophys. Res.*, *108*(D22), 8843, doi:10.1029/2002JD003246.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen (2002), A long-term hydrologically-based data set of land surface fluxes and states for the conterminous United States, *J. Clim.*, *15*, 3237–3251.

Miller, D. A., and R. A. White (1998), A conterminous United States multilayer soil characteristics data set for regional climate and hydrology modeling, *Earth Inter.*, *2*, Paper No. 2.

Mitchell, K., et al. (2000), Recent GCIP-sponsored advancements in coupled land-surface modeling and data assimilation in the NCEP ETA mesoscale model, paper presented at 15th Conference on Hydrology, Am. Meteorol. Soc., Long Beach, Calif., 10–14 Jan.

Mitchell, K., et al. (2004), The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system, *J. Geophys. Res.*, *109*, D07S90, doi:10.1029/2003JD003823.

Nijssen, B., D. P. Lettenmaier, X. Liang, S. W. Wetzel, and E. F. Wood (1997), Streamflow simulation for continental-scale river basins, *Water Resour. Res.*, *33*, 711–724.

Nijssen, B., et al. (2003), Simulation of high latitude hydrological processes in the Torne-Kalix basin: PILPS Phase 2e. 2: Comparison of model results with observations, *Global Planet. Change*, *38*, 31–54.

Oki, T. Nishimura, and P. Dirmeyer (1999), Assessment of annual runoff from land surface models using Total Runoff Integrating Pathways (TRIP), *J. Meteorol. Soc. Jpn.*, *77*, 235–255.

Pan, M., et al. (2003), Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model-simulated snow water equivalent, *J. Geophys. Res.*, *108*(D22), 8850, doi:10.1029/2003JD003994.

Pinker, R. T., et al. (2003), Surface radiation budgets in support of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data Assimilation System (NLDAS) project, *J. Geophys. Res.*, *108*(D22), 8844, doi:10.1029/2002JD003301.

Robock, A., et al. (2003), Evaluation of the North American Land Data Assimilation over the southern Great Plains during the warm season, *J. Geophys. Res.*, *108*(D22), 8846, doi:10.1029/2002JD003245.

Rogers, E., D. Parris, and G. DiMego (1999), Changes to the NCEP operational Eta Analysis, technical procedures bulletin, Off. of Meteorol., Natl. Weather Serv., Silver Spring, Md.

Schaake, J. C., et al. (2004), An intercomparison of soil moisture fields in the North American Land Data Assimilation System (NLDAS), *J. Geophys. Res.*, *109*, D01S90, doi:10.1029/2002JD003309.

Schlosser, C. A., A. Robock, K. Vinnikov, N. Speranskaya, and Y. Xue (1997), 18-year land-surface hydrology model simulations for a midlatitude grassland catchment in Valdai, Russia, *Mon. Weather Rev.*, *125*, 3279–3296.

Schlosser, C. A., et al. (2000), Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS Phase 2(d), *Mon. Weather Rev.*, *128*, 301–321.

Sellers, P., Y. Mintz, Y. Sud, and A. Dalcher (1986), A simple biosphere model (SiB) for use within general circulation models, *J. Atmos. Sci.*, *43*, 505–531.

Sheffield, J., et al. (2003), Snow process modeling in the North American Land Data Assimilation System (NLDAS): 1. Evaluation of model-simulated snow cover extent, *J. Geophys. Res.*, *108*(D22), 8849, doi:10.1029/2003JD003274.

Slater, A. G., et al. (2001), The representation of snow in land-surface schemes: Results from PILPS 2(d), *J. Hydrometeorol.*, *2*, 7–25.

Verdin, K. L., and S. K. Greenlee (1996), Development of continental scale digital elevation models and extraction of hydrographic features, paper presented at Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Natl. Cent. for Geogr. Inf. and Anal., Santa Fe, N. M., 21–26 Jan.

Vörösmarty, C. J., B. Moore, M. P. Gildea, B. Peterson, J. Melillo, D. Kicklighter, J. Raich, E. Rastetter, and P. Steudler (1989), A continental-scale model of water balance and fluvial transport: Application to South America, *Global Biogeochem. Cycles*, *3*, 241–265.

Wang, M., A. T. Hjelmfelt, and J. Garbrecht (2000), DEM aggregation for watershed modeling, *J. Am. Water Resour. Assoc.*, *36*, 3.

Wood, E. F., et al. (1998), The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) Phase-2(c) Red-Arkansas River basin experiment: 1. Experimental description and summary intercomparisons, *Global Planet. Change*, *19*, 115–135.

B. A. Cosgrove and P. R. Houser, Hydrological Sciences Branch, NASA Goddard Space Flight Center, Mail Code 974.1, Greenbelt, MD 20771, USA. (brian.cosgrove@gsfc.nasa.gov; paul.r.houser@nasa.gov)

Q. Duan and J. C. Schaake, Office of Hydrologic Development, NOAA/NWS, 1325 East-West Highway, SSMC2, Room 8356, Silver Spring, MD 20910, USA. (qingyun.duan@noaa.gov; john.schaake@noaa.gov)

R. W. Higgins, Climate Prediction Center, National Centers for Environmental Prediction, NOAA/NWS, 5200 Auth Road, Room 605, Camp Springs, MD 20746-4304, USA. (wayne.higgins@noaa.gov)

D. Lohmann and K. E. Mitchell, Environmental Modeling Center, National Centers for Environmental Prediction, NOAA/NWS, 5200 Auth Road, Camp Springs, MD 20746-4304, USA. (dag.lohmann@noaa.gov; kenneth.mitchell@noaa.gov)

L. Luo, J. Sheffield, and E. F. Wood, Department of Civil and Environmental Engineering, Princeton University, Room E208, E-Quad, Olden Street, Princeton, NJ 08544, USA. (lluo@princeton.edu; justin@princeton.edu; efwood@princeton.edu)

R. T. Pinker, Department of Meteorology, University of Maryland, College Park, 2213 Computer and Space Sciences Building, College Park, MD 20742-2425, USA. (pinker@atmos.umd.edu)

A. Robock, Department of Environmental Sciences, Rutgers University, 14 College Farm Road, New Brunswick, NJ 08901-8551, USA. (robock@envsci.rutgers.edu)

J. D. Tarpley, Office of Research and Applications, NESDIS, E/RA1 WWBG Room 712, 5200 Auth Road, Camp Springs, MD 20746, USA. (dan.tarpley@noaa.gov)