Ecological Applications, 14(2), 2004, pp. 527–545

VALIDATION OF SIMULATED RUNOFF FROM SIX TERRESTRIAL ECOSYSTEM MODELS: RESULTS FROM VEMAP

W. S. GORDON,[1,6] J. S. FAMIGLIETTI,[2] N. L. FOWLER,[3] T. G. F. KITTEL,[4] AND K. A. HIBBARD[5,7]

[1]*Graduate Program in Plant Biology, University of Texas, Austin, Texas 78712 USA*
[2]*Department of Earth System Science, University of California, Irvine, California 92697 USA*
[3]*Section of Integrative Biology, University of Texas, Austin, Texas 78712 USA*
[4]*Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado 80307 USA, and Natural Resource Ecology Laboratory, Colorado State University Fort Collins, Colorado 80523 USA*
[5]*Global Carbon Program, Climate Change Research Center University of New Hampshire, Durham, New Hampshire 03824 USA*

Abstract. Vegetation/Ecosystem Modeling and Analysis Project (VEMAP) Phase 2 model experiments investigated the response of biogeochemical and dynamic global vegetation models (DGVMs) to differences in climate over the conterminous United States. This was accomplished by simulating ecosystem processes using historical climate and atmospheric CO_2 records from 1895–1993. We evaluated the behavior of six models (Biome-BGC, Century, GTEC, LPJ, MC1, and TEM) by comparing simulated runoff in 13 watersheds to gauged streamflow from the Hydro-Climatic Data Network. Metrics used to assess the ''goodness of fit'' between simulated and observed values were: (1) Pearson's r to evaluate the overall data set, (2) Kendall's τ to gauge seasonality trends as derived from a time-series analysis of monthly runoff, and (3) three measures of absolute and relative error.

We found small differences in performance among the six models over all watersheds. However, the models yielded highly divergent results depending upon the watershed analyzed. Performance of the ensemble of models in a watershed was positively correlated with observed streamflow: models in the wettest watersheds in this study were associated with the highest model correlations and largest absolute errors, and models in the driest watersheds were associated with the lowest correlations and smallest absolute errors. Mean relative error was small and nearly constant across watersheds. A bias estimator showed that the models tended to underestimate runoff in wet watersheds and overestimate runoff in dry watersheds. Analysis of long-term trends in runoff using a moving-average approach demonstrated the ability of the models to reproduce temporal variation in observed data, even though quantitative differences among models were large.

Models relying on prescribed vegetation (Biome-BGC, Century, and TEM) outperformed the two DGVMs (LPJ and MC1); GTEC gave the poorest fit to observations due to the absence of an evaporation function and a snow routine. Across all 13 watersheds, TEM ranked the highest in model performance. The validation results presented here suggest that improvements in the simulation of hydrologic processes in land-surface models will come, in part, from a more realistic representation of subgrid-scale soil moisture and from a more detailed understanding and representation of subsurface processes.

Key words: Biome-BGC, Century, GTEC, LPJ, MC1, and TEM compared; climate change; dynamic global vegetation models; HCDN (Hydro-Climatic Data Network [USGS]); model intercomparisons and validation; Nash-Sutcliffe coefficient of efficiency; runoff estimation; streamflow; terrestrial ecosystem models; United States, conterminous; VEMAP Phase 2 model experiments; watersheds.

INTRODUCTION

Modeling sensitivity to altered climate conditions is currently an important focus of the climate change research community because of uncertainties about the speed and extent of future shifts in temperature and precipitation. In recent years, ecologists, resource managers, and policy makers have worked to identify the potential effects of climate change on ecosystem structure and function. While climate change studies are proliferating, it is difficult to gauge the accuracy of their results because these studies are modeling novel states. Model intercomparison studies (e.g., Wood et al. 1998, Cramer et al. 1999) and the reconstruction of past climates (e.g., Coe and Bonan 1997, Claussen et al. 1999) are among the techniques that have been used to try to assess the validity of underlying land-surface models used in climate change modeling experiments. As they can highlight short-

Manuscript received 3 September 2002; revised 24 March 2003; accepted 5 June 2003; final version received 17 July 2003. Corresponding Editor: M. L. Goulden.

[6] E-mail: wgordon@mail.utexas.edu

[7] Present address: College of Forestry, Oregon State University, Richardson Hall 344, Corvallis, Oregon 97331 USA.

comings and inconsistencies, model intercomparisons are a useful adjunct to validation but not a substitute.

One project that has combined both intercomparison and validation in its quest to understand the responses of terrestrial ecosystems to differences in climate over the conterminous United States is the Vegetation/Ecosystem Modeling and Analysis Project (VEMAP). VEMAP's objective has been to force biogeochemical (i.e., ecosystem function) and biogeographical (i.e., ecosystem structure) models with a common set of inputs derived from historical climate and projected climate and atmospheric CO_2 scenarios in order to understand how the models differ in attributes and responses (VEMAP Members 1995, Schimel et al. 2000). A common set of inputs facilitates model intercomparison while the use of historical climate data allows validation.

To gauge how well simulations perform requires rigorous assessment, and setting benchmarks against which to measure success. Model validation is essential to the interpretation of simulation results. It illuminates under what circumstances a model reproduces events accurately and under what circumstances it performs unsatisfactorily. Validation is also critical to the improvement of models; the modeling community cannot improve models if it does not know how, where, and when they fail. Calls for the evaluation or validation of climate and related models have been present in the literature for decades (e.g., Willmott et al. 1985, Koster et al. 1999, Cramer et al. 2001).

Lack of scientific consensus about which methods are most appropriate for determining model accuracy has been one obstacle to the widespread adoption of validation techniques (Willmott et al. 1985, Rastetter 1996). A second challenge, which is often model or question specific, has been determining what constitutes success. Finally, few data sets are available for model validation of continental-scale simulations of ecological processes, particularly at appropriate spatial and temporal scales (Scurlock et al. 1999). As Rastetter (1996) noted, tests of long-term phenomena against data derived from short-term experiments may be inappropriate because processes that dominate at one temporal scale (e.g., months to years) may not be important to long-term behavior (e.g., decades to centuries). Moreover, processes that control long-term responses may not be apparent from short-term data. It is also well known that data to verify ecologically meaningful variables such as evapotranspiration or net primary production may be available at scales ranging from leaf to plot, but scaling up to the landscape level is fraught with uncertainty.

Before turning to the task of understanding sensitivity of land-surface processes to possible altered forcings (e.g., land-use change, CO_2, climate), the scientific community must be capable of realistically simulating past and present states. One widely accepted validation technique in the area of land-surface modeling is to compare simulated runoff to observed streamflow records. Streamflow serves to integrate a host of local- and regional-scale processes, and as such is appropriate for the validation of continental-scale simulations of land-surface processes (e.g., Vörösmarty and Moore 1991, Abdulla et al. 1996, Bonan 1998, Arora et al. 2000, Olivera et al. 2001). Assuming a simple hydrologic budget where streamflow equals precipitation minus evapotranspiration, streamflow can serve as a proxy for a wide range of hydrologic processes, including surface and base runoff, evapotranspiration, and precipitation. In addition, because many of these processes are influenced by the local ecosystem and physical properties of the areas in which they occur (e.g., vegetation, topography, and soil hydrologic characteristics), streamflow is an integrator of the physical and natural environment.

In this study we used streamflow records as a yardstick against which to measure the effectiveness of six terrestrial ecosystem models in reproducing temporal and spatial patterns of observed runoff. We wanted to determine under what conditions the models accurately simulated monthly runoff and under what conditions the models performed poorly. We evaluated the behavior of six VEMAP models—Biome-BGC (BGC = BioGeochemical Cycles; Hunt and Running 1992, Running and Hunt 1993), Century (Parton et al. 1987, 1988, 1993), Global Terrestrial Ecosystem Carbon Model (GTEC; Post et al. 1997), Lund-Potsdam-Jena Dynamic Global Vegetation Model (LPJ; Haxeltine and Prentice 1996, Sitch 2003), MC1 Dynamic Global Vegetation Model (MC = modified Century; Daly et al. 2000), and Terrestrial Ecosystem Model (TEM; McGuire et al. 1992, Melillo et al. 1993, Tian et al. 2000)—by comparing simulated runoff from the VEMAP Phase 2 historical (20th century) experiments to gauged streamflow from the Hydro-Climatic Data Network (HCDN; Slack and Landwehr 1992). These streamflow observations cover much of the 20th century, providing a lengthy record of monthly flows against which to validate the models. We applied several metrics to gauge the ''goodness of fit'' between modeled and observed data for 13 watersheds representing a range of vegetation types and climate zones. In particular, we were interested in knowing how well the models reproduced the overall observed data set and how well they accounted for seasonal differences in runoff. Our results present an important assessment of the performance of the water balance of the constituent VEMAP models.

Methods

VEMAP project and models

VEMAP Phase 2 model inputs consisted of temporally infilled and spatially interpolated measured temperature and precipitation data for 1895–1993 on a 0.5° of latitude × 0.5° of longitude grid for the conterminous United States (Daly et al. 1994, Kittel et al. 1997,

TABLE 1. Summary of terrestrial ecosystem models' hydrologic parameters.

Model†	Time step‡	Soil layers	Soil depth§	Evapotranspiration‖	Snow routine¶	Baseflow#
BBGC	daily	1	variable	Penman-Monteith (1973)	yes	no
CENT	monthly	up to 10	variable	Linacre (1977)	yes	yes
GTEC	daily	12	constant	none	no	no
LPJ	daily	2	constant	Monteith (1995)	yes	no
MC1	monthly	up to 10	variable	Linacre (1977)	yes	yes
TEM	monthly	1	variable	Jensen-Haise (1963)	yes	yes

† Six VEMAP models were evaluated: BBGC = Biome-BioGeochemical Cycles model (Hunt and Running 1992, Running and Hunt 1993), CENT = Century (Parton et al. 1987, 1988, 1993), GTEC = Global Terrestrial Ecosystem Carbon model (Port et al. 1997), LPJ = Lund-Potsdam-Jena model (Haxeltine and Prentice 1996, Sitch 2003), MC1 = modified Century (Daly et al. 2000), and TEM = Terrestrial Ecosystem Model (McGuire et al. 1992, Melillo et al. 1993, Tian et al. 2000). BBGC, Century, GTEC, and TEM are biogeochemical cycling models; LPJ and MC1 are dynamic global vegetation models.

‡ Time step refers to the frequency with which the hydrologic models were updated.

§ The soil depth was either constant for each grid cell or was retrieved from the VEMAP soils data set.

‖ The specification of evapotranspiration method does include stomatal processes that would have been calculated independently.

¶ Snow routines modeled snowpack accumulation and melt.

Baseflow occurred when water reaching the bottom of the soil profile was siphoned from the water balance rather than being added to surface runoff. Baseflow does not refer to the modeling of groundwater contributions.

2000). Other climate forcings including solar radiation and humidity were empirically estimated from daily temperature and precipitation (Kittel et al. 2000). Daily and monthly versions of the data were created to serve the input requirements of the different terrestrial ecosystem models used in the project. Daily values were disaggregated from the monthly records using a modified version of the stochastic weather generator WGEN (Richardson 1981, Richardson and Wright 1984, Kittel et al. 1995).

The VEMAP models investigated include four biogeochemical cycling models (Biome-BGC, Century, GTEC, and TEM), which simulate plant production and nutrient cycles, but rely on a static land-cover type. For these models, land cover is based on a vegetation map derived from Küchler's (1975) scheme of potential natural vegetation (Kittel et al. 1995), and has prescribed levels of disturbance (e.g., fire). The two dynamic global vegetation models (DGVMs), LPJ and MC1, combine biogeochemical cycling processes with

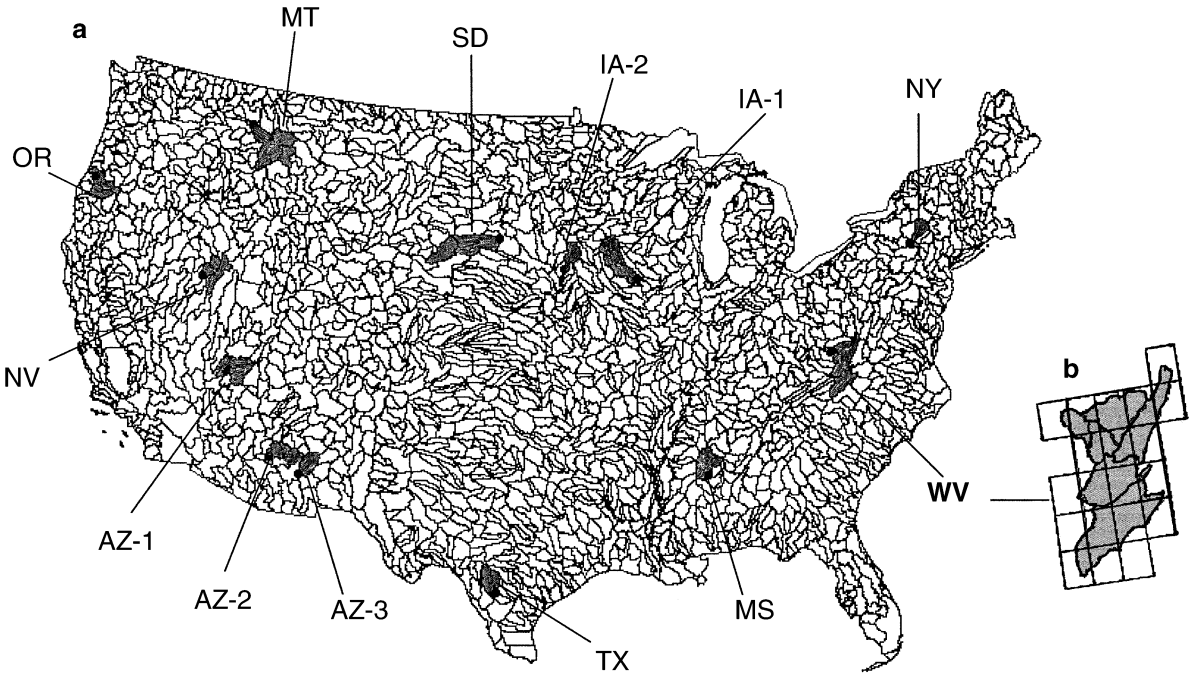

FIG. 1. The 13 watersheds used in this model-validation exercise. (a) Map of the United States showing USGS eight-digit hydrologic units and location of the selected watersheds. (b) Example of VEMAP 0.5°-grid overlay on a watershed. State codes: AZ = Arizona, IA = Iowa, MS = Mississippi, MT = Montana, NV = Nevada, NY = New York, OR = Oregon, SD = South Dakota, TX = Texas, WV = West Virginia.

TABLE 2. Location, climate data, and descriptive information from the HCDN (USGS's Hydro-Climatic Data Network; Slack and Landwehr 1992) for watersheds used in validation.

Watershed code name†	HUC of Gauge‡	Gauge location			
		Water body	Latitude	Longitude	Elevation (m)
NY	2050101	Butternut Creek	42.035	−75.803	475
MS	3160101	Bull Mountain Creek	33.489	−88.433	86
WV	5050006	Kanawha River	38.138	−81.214	820
IA-1	7080205	Cedar River	41.971	−91.667	323
SD	10140204	White River	43.748	−99.556	792
IA-2	10230003	Little Sioux River	42.472	−95.797	433
TX	12110106	Frio River	28.736	−99.144	337
AZ-1	15010010	Virgin River	36.892	−113.92	1676
AZ-2	15040004	San Francisco River	33.049	−109.3	2097
AZ-3	15060103	Salt River	33.619	−110.92	1887
NV	16040101	Humboldt River	40.607	−116.2	1966
MT	17010204	Clark Fork	47.302	−115.09	1664
OR	17100303	Umqua River	43.586	−123.55	756

† The station identifier used in the text (see Fig. 1 legend for state codes).
‡ The hydrologic unit code (HUC) of the watershed in which the streamflow gauge resides.
§ USGS water years start in October of the year prior to the one listed in the ''first-year'' column and end in September of the year listed in the ''last-year'' column. In a few cases, the length of the HCDN record listed is shorter than that indicated by the start and end dates listed because of discontinuities.
‖ The total drainage area of the watershed whose streamflow is measured by the HUC gauge.
¶ The mean annual streamflow of the watershed normalized to watershed area.
# The annual rainfall measured in the vicinity of the gauge, as reported in the HCDN.
†† The dominant VEMAP vegetation type in the watershed at the start of the experiment: 3 = maritime temperate coniferous forest, 4 = continental temperate coniferous forest, 5 = cool temperate mixed forest, 6 = warm temperate/subtropical mixed forest, 7 = temperate deciduous forest, 10 = temperate mixed xeromorphic woodland, 11 = temperate conifer xeromorphic woodland, 13 = temperate deciduous savanna, 14 = warm temperate/subtropical mixed savanna, 17 = C_3 grasslands, 18 = C_4 grasslands, 20 = temperate arid shrublands, and 21 = subtropical arid shrublands.

dynamic biogeographical processes including succession and fire simulation. In total there are 21 VEMAP vegetation types plus wetlands, though wetland processes are not simulated by these models.

The formulation of hydrologic processes varies among models. Unless otherwise specified, model soil depth was constrained by the VEMAP soils data set, which is spatially variable (Kittel et al. 1995). A summary of the attributes of the hydrologic models embedded within each of the terrestrial ecosystem models is presented in Table 1.

1) Biome-BGC (or BBGC in the figures and tables) uses a single ''bucket'' model in which inputs of precipitation are balanced with the outputs of evapotranspiration and runoff. The time step is daily. Evapotranspiration is calculated using the Penman-Monteith equation (Monteith 1973). There is a single soil layer and any soil water in excess of field capacity is routed to runoff, including that which flows out from the bottom of the soil profile. A snow routine accumulates snow below 0°C and initiates the melt process above that temperature.

2) Century (''CENT'' in figures and tables) and MC1 share the same water-balance components. Both operate on a monthly time step. Evapotranspiration is calculated using Linacre (1977). There are as many as 10 soil layers, each 15 cm deep up to a depth of 60 cm and 30 cm deep below that point. A fixed fraction of rainfall is immediately allocated to surface runoff. The remaining water travels through successive soil layers as field capacity is exceeded. Some of the water released by the deepest layer enters the groundwater as the baseflow component of runoff and some is redirected to surface runoff via stormflow. A snow routine accumulates snow below 0°C and initiates the melt process above that temperature.

3) GTEC hydrology is derived from the SUNDIAL model (Bradbury et al. 1993). There are 12 soil layers: 0–50 cm in 5 cm increments, 50–100 cm, and 100–200 cm. Leaching occurs as a ''piston flow'' process, water successively filling each layer down the profile, before draining to the layer below. Bypass flow, or runoff, occurs if rainfall in a given period exceeds a specified threshold value. Any water reaching the bottom of the soil profile is redirected to surface runoff. Soil water values are updated daily. There is neither an evaporation function nor a snow routine.

4) LPJ uses monthly forcing input, but interpolates the climate to a daily time step for all processes including its hydrologic model. Evapotranspiration is computed using Monteith (1995). The modified bucket model from Neilson (1995) contains two soil layers, the first of which is 50 cm and the second 50–150 cm. Water in excess of field capacity (i.e., surface runoff and deep drainage) is considered runoff. A threshold temperature of −2°C tested daily determines whether precipitation enters the soil directly or is stored as snow.

5) The water-balance model for TEM comes from Vörösmarty et al. (1989), with the exception that evapotranspiration is calculated after Jensen-Haise (1963). TEM operates on a monthly time step. It uses

TABLE 2. Extended.

Gauge HCDN record§			Drainage area (km²)‖	Mean annual streamflow (mm/yr)¶	Annual rainfall (mm)#	VEMAP vegetation type††
Length (yr)	First year	Last year				
75	1914	1988	5944	538.5	1041.4	5
64	1900	1979	11 549	502.5	1346.2	6
111	1878	1988	23 136	471.2	1117.6	7
86	1903	1988	17 581	176.7	795.02	13
60	1929	1988	25 682	18.5	431.8	17
61	1919	1988	7203	103.1	711.2	18
73	1916	1988	8904	14.5	584.2	14
59	1930	1988	15 467	14.0	406.4	11
63	1914	1988	7270	26.7	459.7	21
75	1914	1988	12 315	65.6	558.8	10
81	1903	1988	13 129	27.4	228.6	20
72	1912	1988	28 228	236.8	457.2	4
83	1906	1988	12 134	551.1	1193.8	3

a single bucket model with a single soil layer whose depth is determined by soil texture class. Runoff is generated from subsurface runoff pools when field capacity is exceeded. Snowpack accumulates whenever mean monthly temperature is below −1.0°C; snowmelt occurs above this temperature. TEM is unique among the six models in that CO_2 concentration and plant growth do not affect any water-balance calculations (i.e., neither runoff nor actual evapotranspiration respond to plant processes) because the hydrologic model is run offline and prior to the ecosystem model.

Runoff validation using USGS records

To validate VEMAP-simulated runoff, we compared model results to the Hydro-Climatic Data Network (HCDN; Slack and Landwehr 1992) for 13 watersheds. The HCDN consists of streamflow gauging stations with records retrieved from the U.S. Geological Survey's (USGS) National Water Storage and Retrieval System. Only stations relatively free of confounding anthropogenic influences that would significantly alter the ''natural'' streamflow—such as diversions, regulation of flow, or changes in watershed land use—were included in the database. The HCDN data set contains the mean daily discharge for 1571 sites across the continental United States. The data set contains streamflow records collected between 1874 and 1988, with an average station record length of approximately 48 years.

We initially reduced the universe of sites to just over 100 by selecting only those sites that satisfied the following criteria: (a) records of at least 50 years, (b) watersheds of at least 5200 km² whose boundaries were completely contained within the United States, and (c) watersheds whose noncontributing areas were less than 10%. (Non-contributing areas are closed basins within a watershed.) The number of potential records was further reduced by selecting watersheds that fell within a single VEMAP vegetation type and by trying to maximize the range of climates represented. The 13 watersheds selected for this study range in size from 5944 to 28 228 km² (Fig. 1). The watershed areas reported in the HCDN differ by <3% from that we determined by using ArcView (ESRI, Redlands, California, USA). Records of the 13 watersheds are 59 to 111 years long ($\bar{X}$ = 74 years) (Table 2). In several cases, record length is shorter than that indicated by the start and end dates because the record was not continuous; the gauge-record-length column of Table 2 is corrected for any missing years. Watersheds are identified by a state abbeviation. Also included in Table 2 are eight-digit Hydrologic Unit Codes (HUC) that the USGS uses to uniquely label watersheds within the United States Three watersheds in this study comprise solely one HUC; the remaining 10 watersheds are composed of two or more HUCs (see Fig. 1).

Because of differences in units and geographic formats between the simulated values and observed data we undertook two types of conversions. First we converted the HCDN streamflow data from cubic feet per second to depth values by normalizing the flow by the area of the watershed. Hence, both observed and modeled values are reported as millimeters per month per unit area and are referred to as ''runoff.'' Second, we used ArcView to superimpose HUCs and VEMAP grid cells (e.g., Fig. 1b). From this we computed the fraction of a watershed overlapping each VEMAP cell. We used these fractions to area-weight simulated runoff; total runoff for a watershed was the sum of weighted grid-cell values.

All runoff values evaluated came from a single simulation of the VEMAP models. The two scenarios of increasing (i.e., historical) and constant (i.e., 1895 value) CO_2 produced nearly identical runoff data sets so the analyses presented here are based on the increasing CO_2 simulations. This issue of runoff similarity is explored further in the *Discussion*, below.

Statistical analyses

Several statistical methods were used to gauge the ''goodness of fit'' between the HCDN and the simu-

TABLE 3. Pearson product-moment correlation coefficient *r*, by watershed, between each of the models and the USGS's HCDN (Hydro-Climatic Data Network) observations.

Model†	Watershed‡													Average *r*, by model
	OR	NY	MS	WV	MT	IA-1	IA-2	AZ-3	NV	AZ-2	SD	TX	AZ-1	
BBGC	0.92	0.74	0.88	0.84	0.37	0.67	0.70	0.65	0.40	0.54	0.72	0.53	0.46	0.65
CENT	0.94	0.74	0.87	0.85	0.45	0.60	0.61	0.66	0.47	0.56	0.68	0.55	0.52	0.65
GTEC	0.87	0.67	0.83	0.81	−0.08	0.49	0.44	0.45	0.17	0.43	0.64	0.50	0.40	0.51
LPJ	0.88	0.67	0.74	0.76	0.60	0.56	0.60	0.61	0.54	0.51	0.36	0.12	0.39	0.57
MC1	0.90	0.69	0.83	0.78	0.41	0.41	0.44	0.55	0.36	0.55	0.63	0.51	0.48	0.58
TEM	0.93	0.58	0.88	0.84	0.78	0.68	0.63	0.78	0.69	0.68	0.57	0.38	0.56	0.69
Average *r*, by model	0.91	0.68	0.84	0.81	0.42	0.57	0.57	0.62	0.44	0.55	0.60	0.43	0.47	0.61

† For model key, see Table 1.
‡ For watershed key, see Fig. 1 legend. Watersheds are ordered from wettest (at left) to driest (at right) based on mean annual HCDN streamflow.

lated values in each of the 13 watersheds. The Pearson product–moment correlation coefficient *r* (Sokal and Rohlf 1995) was calculated between observed monthly data and each simulation (six models) over the entire period of overlap between the two (a minimum of 50 years). (In the case of WV the period of overlap is less than the total HCDN record length because the start of record keeping precedes the date on which the simulations commence. See Table 2.) To investigate the ability of the models to reproduce seasonal (i.e., monthly) runoff patterns within watersheds, Kendall's coefficient of rank correlation τ (Kanji 1999) was calculated using the average runoff of each month derived from the two time series (again, a minimum of 50 years). These average values of monthly runoff ($n = 12$, one for each month) calculated from each of the observed and simulated data sets were independently ranked. By comparing the ranks assigned to successive pairs (observed$_i$, simulated$_i$) and (observed$_j$, simulated$_j$) of monthly averages (e.g., ranks of February observed and simulated runoff compared to January ranks) a correlation coefficient was calculated indicating to what degree the pattern of monthly simulated runoff mimicked that of the observed in a given watershed.

We chose to not report significance levels for the reasons that the monthly values of runoff exhibited some serial correlation and because the Pearson analyses were based on hundreds of observations; all the values are ''significant'' at probability values <0.001. We used the Pearson and Kendall statistics solely for descriptive purposes.

To evaluate the magnitude of differences between observed and simulated values, several methods were adopted to measure absolute and relative error. (1) One error metric adopted was mean absolute error (MAE), computed as

$$\text{MAE} = \frac{\sum_{j=1}^{N} |O_j - S_j|}{N}$$

where S_j and O_j are monthly simulated and observed values, respectively, and N is the number of monthly observations (Willmott 1984). (2) Another metric used was a bias estimator (BIAS), calculated as

$$\text{BIAS} = \bar{S} - \bar{O}$$

where $\bar{S}$ and $\bar{O}$ are mean simulated and mean observed values, respectively, as derived from the entire monthly data set (Watterson et al. 1999, Wolock and McCabe 1999). (3) The Nash-Sutcliffe coefficient of efficiency (NS; Nash and Sutcliffe 1970) is another widely used statistic for evaluating the performance of hydrologic models (e.g., Wilcox et al. 1990, Wolock and McCabe 1999, Peel et al. 2001, Sauquet and Leblois 2001). As the ratio of the mean square error to the variance in the measured data, subtracted from unity, it is computed as

$$\text{NS} = 1 - \frac{\sum_{j=1}^{N} (O_j - S_j)^2}{\sum_{j=1}^{N} (O_j - \bar{O})^2}.$$

NS ranges in value from minus infinity (characteristic of a poor model) to 1 (perfect model). If the variance of errors (numerator) is as large as the variance of the observations (denominator) then NS = 0; if the variance of errors exceeds observed variance then NS $<$ 0. If errors approach 0, then high and low flows are well reproduced, and NS approaches 1.

The ability of the models to reproduce observed, temporal runoff patterns in each watershed was ranked according to each metric. An overall rank was determined by using all metrics as an index. Where we ranked using the BIAS estimator, we used its absolute value. In all cases, the model assigned the rank ''1'' most closely matched observed data.

Long-term trends in the data were identified by calculating five-year moving averages. The moving average of a target year was determined by averaging a five-year span including the two calendar years before and after a target year. Because the time series of observed data for five of the watersheds were incomplete, the starting date of the moving averages in these watersheds was the

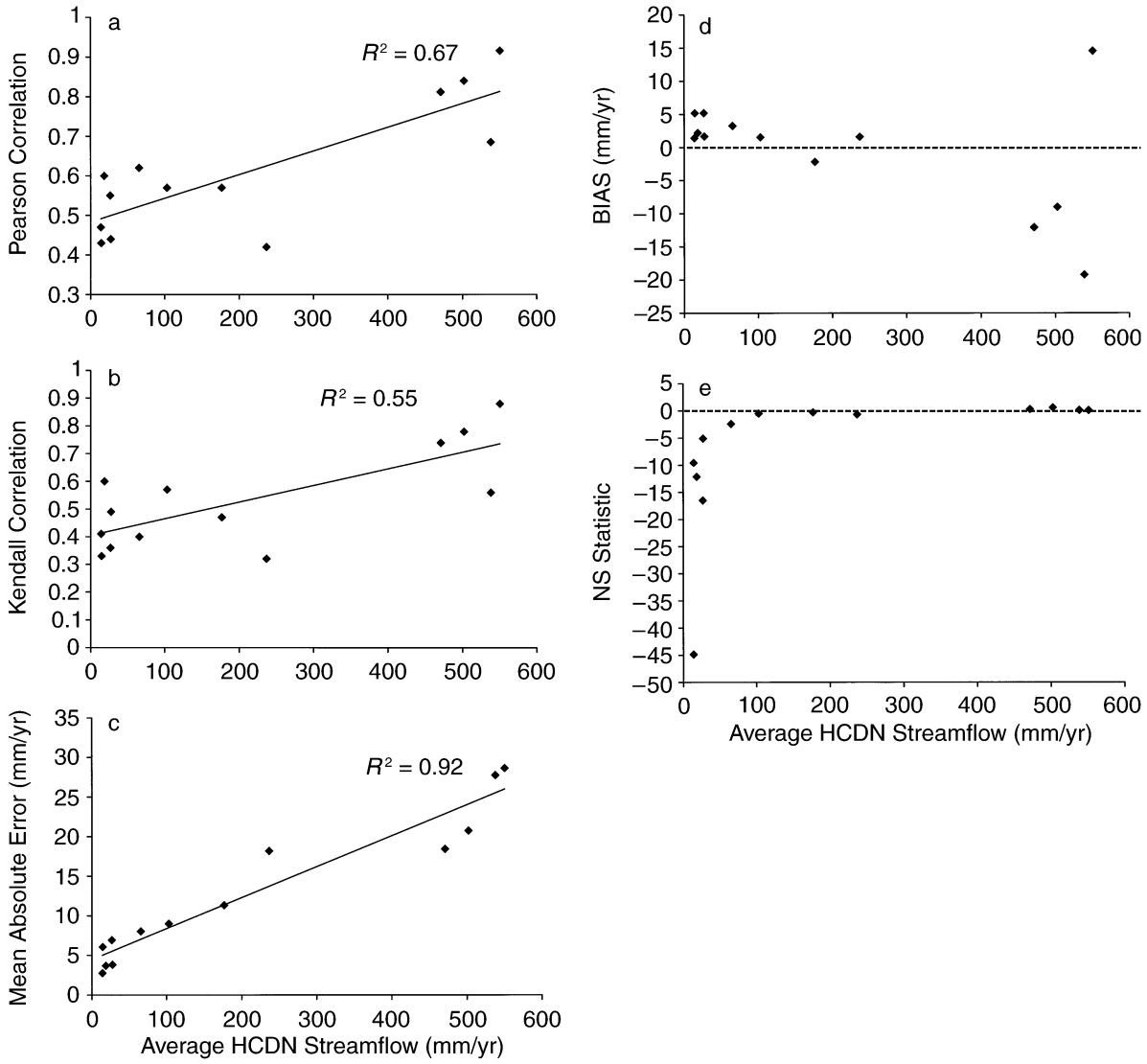

FIG. 2. Results of statistical tests used to evaluate the goodness of fit between the USGS's HCDN data and the simulated values in each of the 13 watersheds: (a) Pearson product–moment coefficient measuring correlation between monthly simulated and observed runoff, plotted as the average of the six models by watershed against mean annual streamflow from the HCDN (Hydro-Climatic Data Network) records; (b) Kendall's τ for assessing differences in the temporal pattern of monthly runoff vs. streamflow; (c) MAE (mean absolute error) vs. streamflow; (d) BIAS (a bias estimator) vs. streamflow; (e) NS (Nash-Sutcliffe coefficient of efficiency) vs. streamflow.

third calendar year after the data gap. We did not undertake detailed statistical analyses of these trends because analyses of runoff trends in the US have been reported elsewhere (e.g., Lettenmaier et al. 1994, Lins and Slack 1999, McCabe and Wolock 2002) and were not the focus of our present study. Other analyses of long-term runoff and actual evapotranspiration trends from the VEMAP Phase 2 experiments can be found in Gordon and Famiglietti (*in press*).

Results in Tables 3–7 are ordered from wettest watershed on the left (OR) to driest on the right (AZ-1) based on mean annual streamflow. Annual precipitation reported in Table 2 is for the location of the gauge and may not reflect the rainfall in the entire watershed. For example, the gauge located in watershed MT is the fourth driest of the 13 watersheds studied in terms of precipitation. Yet it is the fifth wettest in terms of the mean annual HCDN streamflow. For these reasons, we used mean annual HCDN streamflow (in millimeters per year to eliminate a watershed size bias) as a proxy for wetness or dryness of a watershed. Streamflow tends to be highly correlated with watershed precipitation (Dolph and Marks 1992, Wolock and McCabe 1999, Lewis et al. 2000).

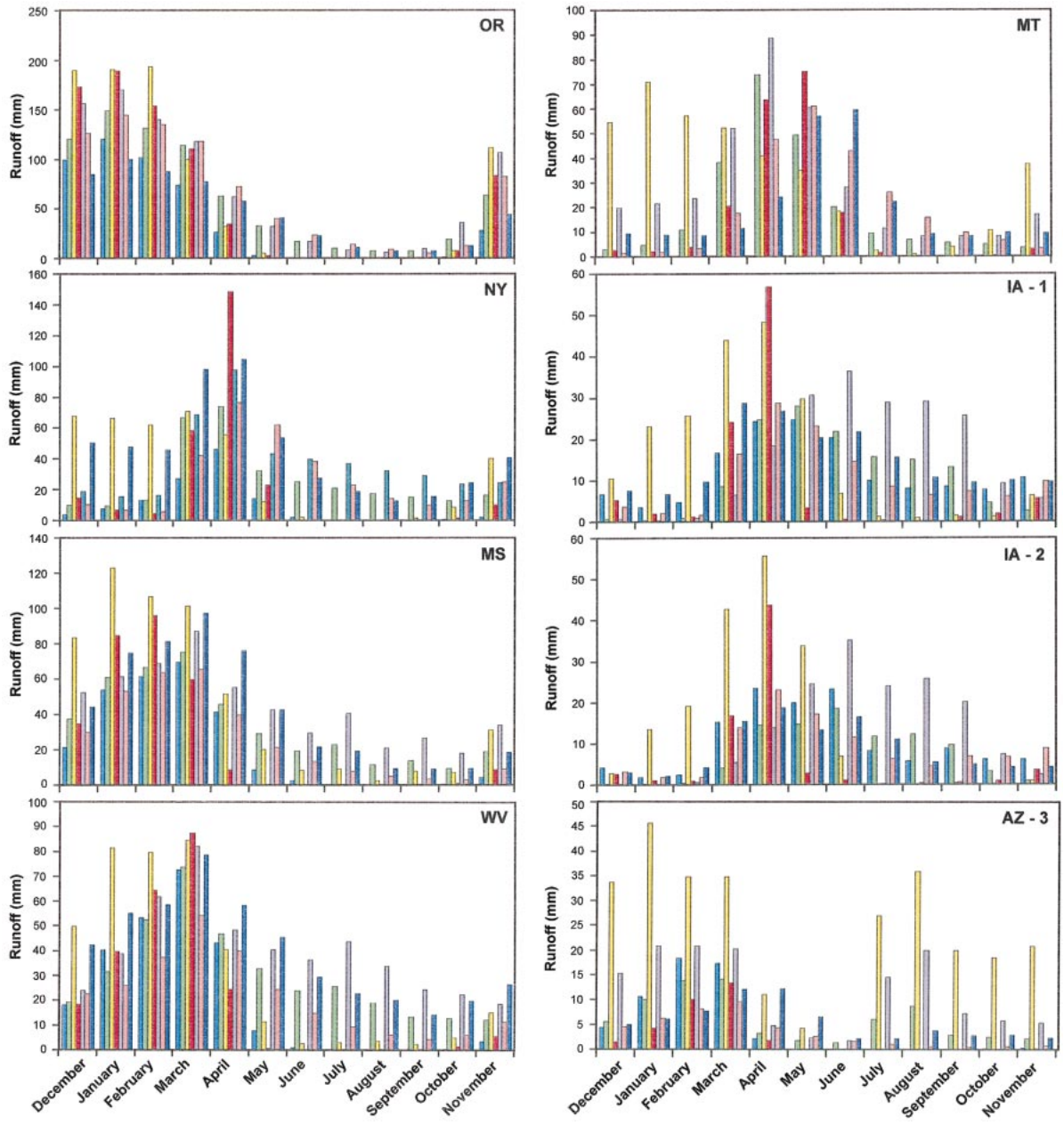

FIG. 3. Seasonal pattern of observed runoff (USGS's HCDN data) and simulated runoff (six VEMAP models), by watershed. Monthly values are averages over the entire time series. Panels are ordered by mean annual observed streamflow from wettest (OR) to driest (AZ-1), moving down the left-hand column and then down the right-hand column of each page. For watershed key, see Fig. 1 legend; for model key, see Table 1 footnote.

RESULTS

Monthly time-series correlation

The Pearson product–moment correlation coefficient r of the monthly values ranged from −0.08 for GTEC (Global Terrestrial Ecosystem Carbon model) in watershed MT to 0.94 for Century in watershed OR (Table 3). Average correlations by model, accounting for results from all 13 watersheds, ranged from a low of 0.51 for GTEC to a high of 0.69 for TEM (Terrestrial Ecosystem Model) (Table 3). Total variability was similar for four of the six models; GTEC and LPJ (Lund-Potsdam-Jena Dynamic Global Vegetation Model) exhibited the greatest variability in correlation coefficients across the 13 watersheds. The average of model correlations within a watershed ranged from a low of 0.42 in watershed MT to a high of 0.91 in watershed OR (Table 3). In some watersheds, the spread of correlation coefficients was narrow (e.g., OR and MS), while in others it was wide (e.g., MT and NV). There was a

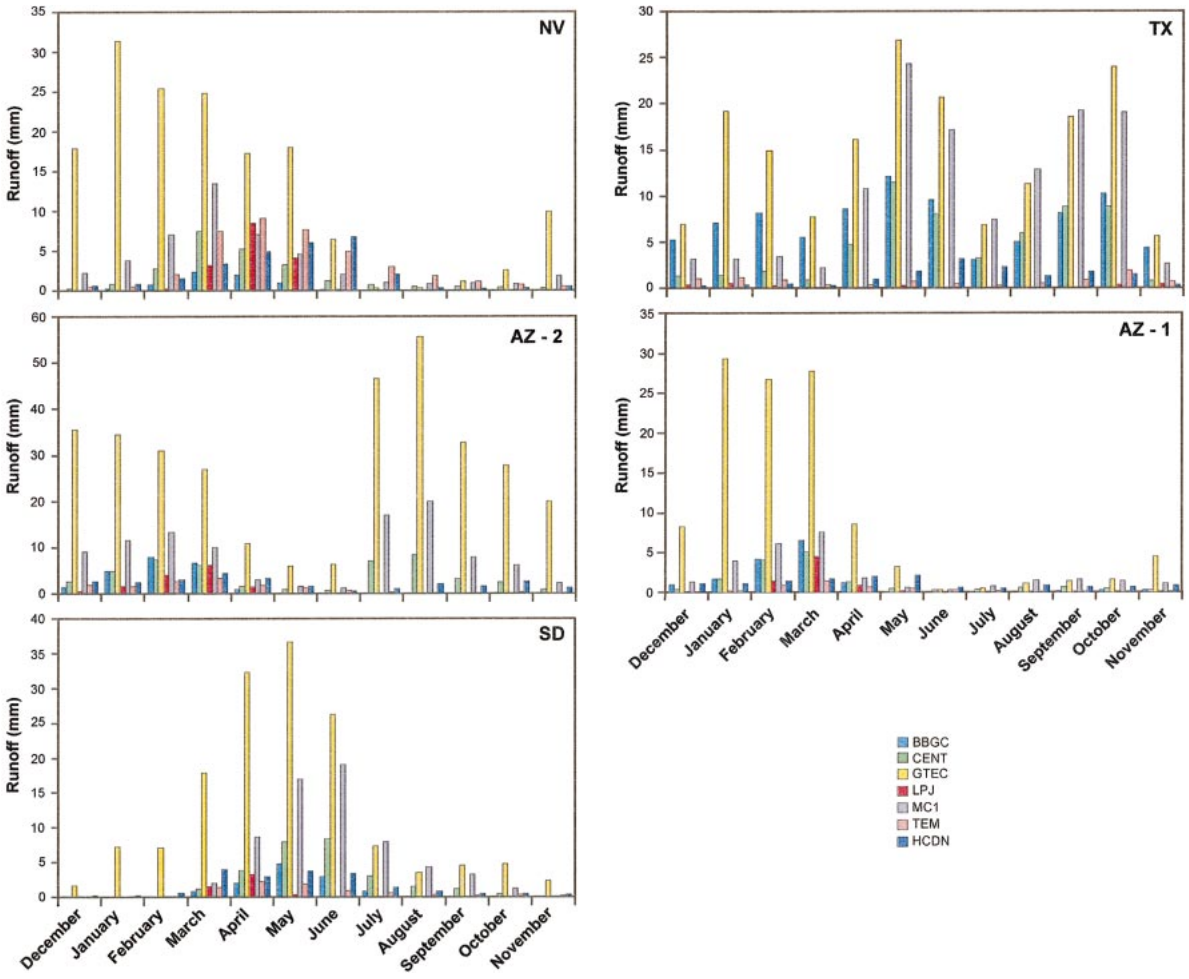

FIG. 3. Continued

positive, linear relationship between the average r of each watershed and the mean annual observed streamflow of that watershed (Fig. 2a; $R^2 = 0.67$). Fig. 3 depicts monthly runoff averaged over the entire historical data set for each of the observed and simulated values. This figure makes clear the close association between observed runoff (dark blue) and simulated runoff (other colors). No one model consistently performed the best or worst in every watershed.

Seasonal pattern correlation

The Kendall's τ analysis for seasonality yielded correlations that ranged from -0.03 for the GTEC model in watershed AZ-2 to 0.94 for both the Century and TEM models in watershed OR (Table 4). Biome-BGC correlations were the least variable, while GTEC and LPJ were the most. Average correlations by model, accounting for results from all 13 watersheds, ranged from a low of 0.39 for GTEC to a high of 0.67 for

TABLE 4. Kendall's correlation of rank order τ, by watershed, between each of the models and the USGS's HCDN observations.

Model	Watershed													Average τ
	OR	NY	MS	WV	MT	IA-1	IA-2	AZ-3	NV	AZ-2	SD	TX	AZ-1	
BBGC	0.91	0.88	0.82	0.88	0.36	0.64	0.79	0.56	0.60	0.64	0.73	0.36	0.48	0.67
CENT	0.94	0.21	0.88	0.73	0.39	0.67	0.79	0.39	0.48	0.24	0.70	0.67	0.39	0.58
GTEC	0.82	0.70	0.64	0.64	−0.18	0.15	0.49	0.15	0.23	−0.03	0.61	0.39	0.52	0.39
LPJ	0.85	0.82	0.60	0.58	0.45	0.15	0.09	0.60	0.61	0.59	0.24	−0.32	0.50	0.49
MC1	0.82	0.18	0.85	0.61	0.27	0.52	0.64	0.24	0.52	0.18	0.64	0.67	0.33	0.50
TEM	0.94	0.33	0.91	0.97	0.61	0.67	0.61	0.48	0.52	0.55	0.70	−0.21	0.24	0.59
Average τ	0.88	0.56	0.78	0.74	0.32	0.47	0.57	0.40	0.49	0.36	0.60	0.33	0.41	0.54

Note: Format is as in Table 3.

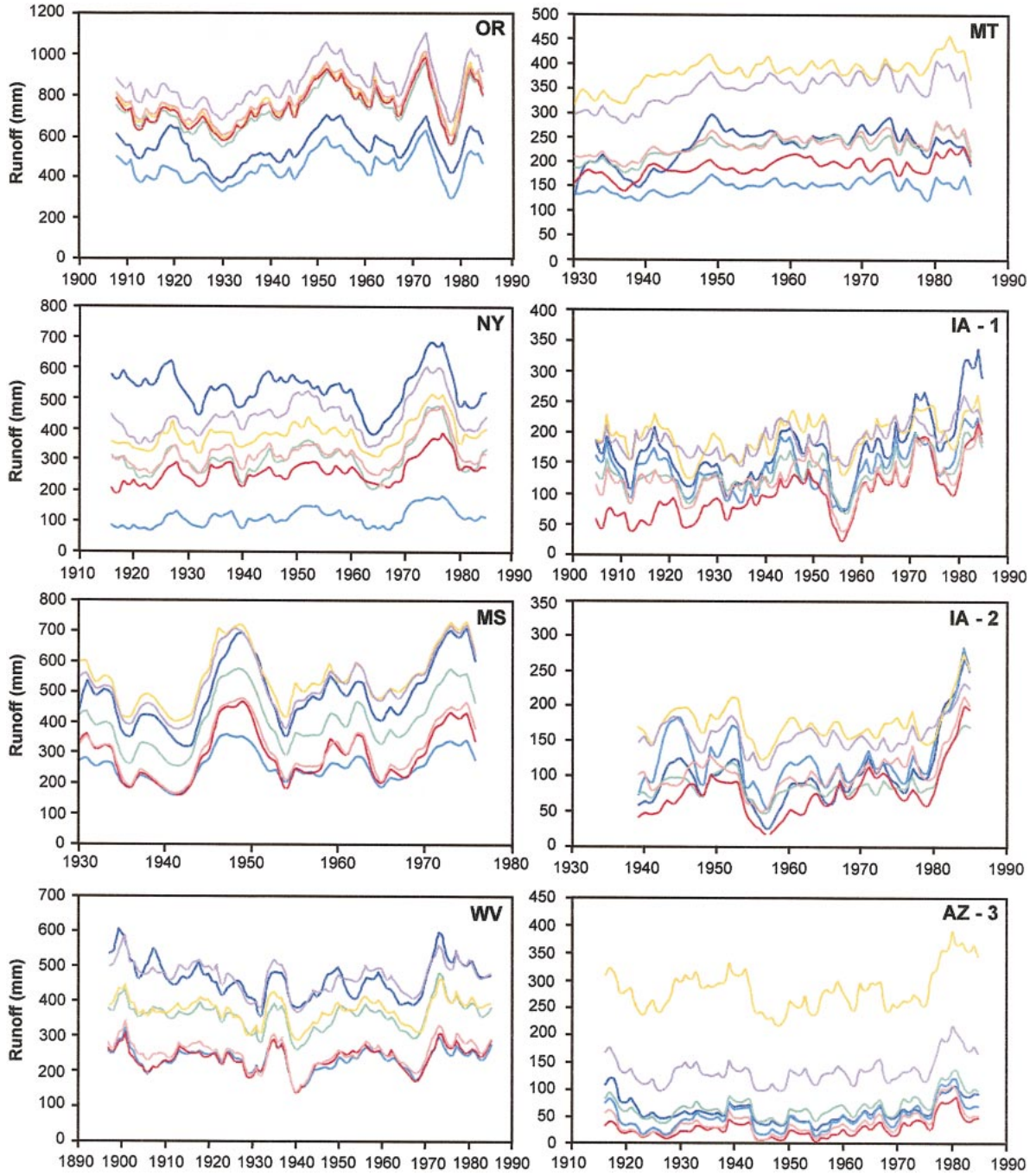

FIG. 4. Annual time series of runoff from each VEMAP model and observed HCDN (USGS's Hydro-Climatic Data Network) data set by watershed. Five-year moving averages were calculated to smooth the data sets and make the trends more apparent. The panels are ordered by mean annual HCDN streamflow from wettest (OR) to driest (AZ-1), moving down the left-hand column and then down the right-hand column of each page.

Biome-BGC. The average of model correlations within a watershed ranged from a low of 0.32 in watershed MT to a high of 0.88 in watershed OR. In a few watersheds, the spread of correlation coefficients was narrow (e.g., OR and MS), but overall the Kendall's τ results were highly variable from model to model. Correlations were negative in watershed TX (for LPJ and TEM) and watershed MT (GTEC), indicating a modeled seasonality pattern opposite that of the observed. There was a positive trend between the Kendall's τ and mean annual observed streamflow (Fig. 2b; $R^2 = 0.55$); the better fit between simulated and observed stream-

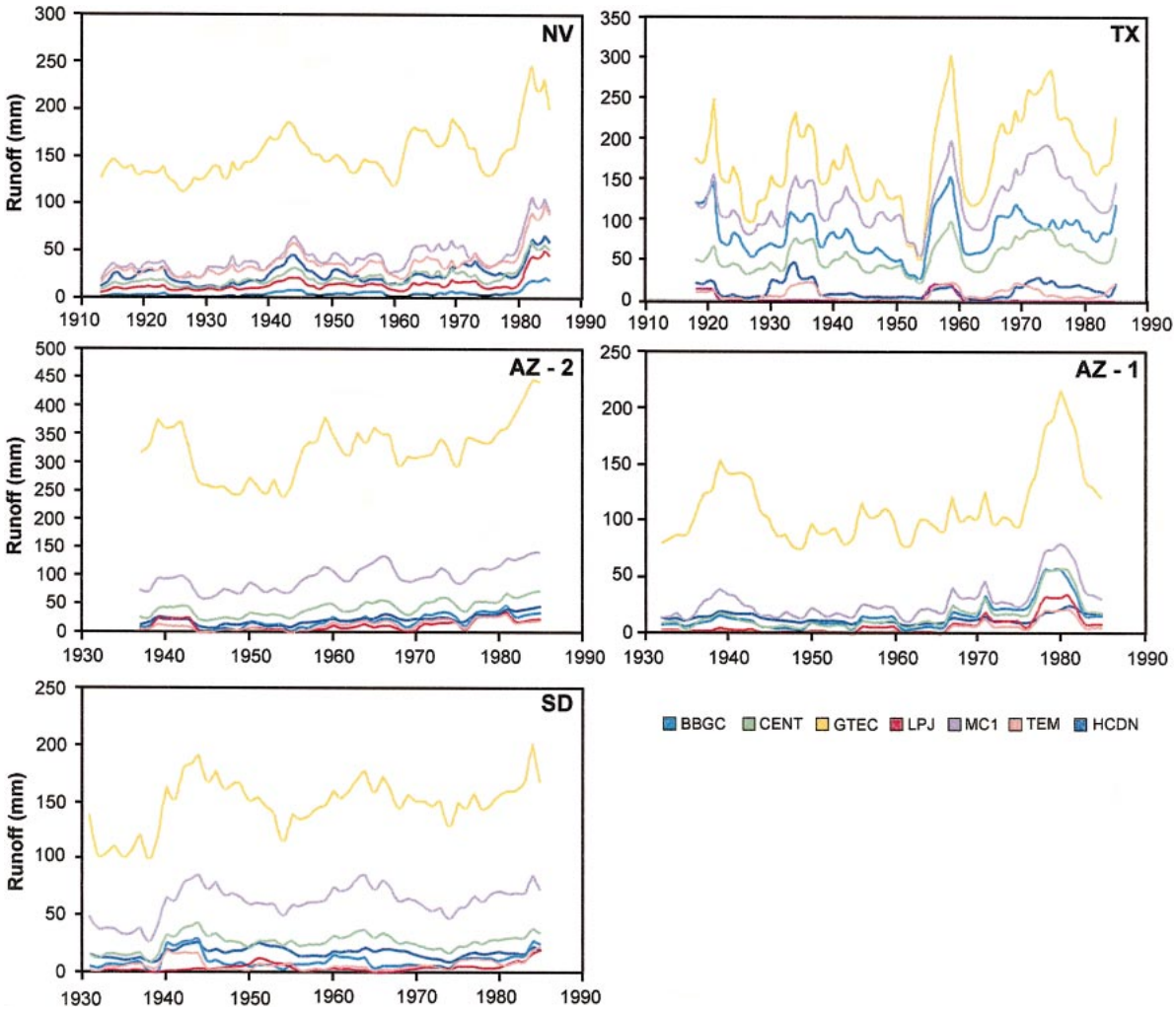

FIG. 4. Continued

flow in the wetter watersheds can also be seen in Fig. 3. Again, no one model consistently performed the best or worst in every watershed.

For watersheds in which snow dynamics play an important role in the annual water balance, such as NY and MT, GTEC's lack of a snowpack accumulation and melt routine resulted in temporal errors in runoff (Fig. 3). GTEC overestimated runoff in the winter months and underestimated runoff during the spring. This pattern was in contrast to that observed in arid watersheds

TABLE 5. Mean absolute error (MAE) (in millimeters per year) calculated as the sum of the absolute differences between each pair of observed runoff values (USGS's HCDN data) and simulated runoff values (from one of the six models), divided by the total number of observations, for each of the 13 watersheds.

Model	Watershed													Average MAE
	OR	NY	MS	WV	MT	IA-1	IA-2	AZ-3	NV	AZ-2	SD	TX	AZ-1	
BBGC	19.52	35.28	21.52	20.89	18.33	8.69	7.44	5.48	1.93	2.41	1.38	6.30	1.52	11.59
CENT	20.66	23.44	16.47	12.82	15.79	10.15	6.96	4.82	1.96	2.90	2.04	4.05	1.48	9.50
GTEC	40.16	24.90	22.70	19.94	29.90	14.10	13.56	20.56	11.91	25.83	11.35	13.98	8.96	19.83
LPJ	37.04	32.91	25.31	23.55	16.03	12.98	8.17	4.58	1.93	2.03	1.41	1.24	1.33	12.96
MC1	32.02	23.42	20.25	14.76	19.30	14.29	11.80	9.27	3.26	6.90	4.76	9.57	2.32	13.22
TEM	22.41	26.62	18.28	18.64	9.78	7.70	6.06	3.45	1.98	1.50	1.25	1.24	1.02	9.23
Average MAE	28.64	27.76	20.76	18.44	18.19	11.32	9.00	8.03	3.83	6.93	3.70	6.06	2.77	12.72

Note: Format is as in Table 3.

TABLE 6. Bias estimator (BIAS) calculated as the mean simulated runoff (from one of the six models) minus the mean observed runoff (USGS's HCDN data), by watershed, with the 13 watersheds ordered left to right as wettest to driest; results are in millimeters per year.

	Watershed						
Model	OR	NY	MS	WV	MT	IA-1	IA-2
BBGC	−8.23	−35.20	−20.02	−19.30	−7.28	−2.57	1.86
CENT	15.22	−18.86	−7.69	−8.95	−0.59	−3.36	−1.07
GTEC	18.58	−12.69	4.01	−7.70	12.22	1.67	6.05
LPJ	16.77	−22.72	−17.46	−19.04	−3.91	−6.16	−2.41
MC1	25.72	−7.73	2.77	0.32	9.17	1.31	4.65
TEM	19.16	−17.84	−15.75	−17.88	−0.05	−4.04	0.18
Average BIAS	14.54	−19.17	−9.02	−12.09	1.59	−2.19	1.54

where GTEC overestimated runoff most months (e.g., TX).

Error analyses

Mean annual error (MAE) was proportional to the mean annual HCDN streamflow (Fig. 2c, slope = 0.045, $R^2 = 0.92$). Overall, the models performed similarly with the exception of GTEC, which produced the largest values of MAE. However, performance of the models was highly variable from watershed to watershed. The average of model errors within a watershed ranged from 9.23 to 19.83 mm/yr (Table 5). As would be expected, the MAE values were proportional to the relative wetness of the watersheds, with models producing large values of MAE in wet watersheds and small values in dry watersheds. Relative error averaged about 4.5% and was fairly even across all watersheds (Fig. 2c).

When results from all watersheds were considered together, the bias estimator (BIAS) was negative for four of the six models, indicating that they underestimated runoff (Table 6). The two models for which BIAS was positive across the watersheds, MC1 and GTEC, had negative values in watersheds NY (both models) and WV (GTEC). GTEC overestimated runoff by an order of magnitude in some watersheds, skewing the average results for those watersheds. BIAS was positive in the watersheds with annual streamflow of about 100 mm/yr or less; BIAS was negative in three out of the four watersheds with annual streamflow >400 mm/yr (Fig. 2d). In relative terms BIAS was within about ±5% in watersheds with annual streamflow >100 mm/yr, but exceeded +20% in some of the driest watersheds. BIAS grew in magnitude as streamflow increased. As with MAE, this was anticipated as the differences between the simulated and observed data sets were expected to grow larger (negative or positive) as runoff increased.

If MAE and BIAS had been similar in absolute magnitude, these metrics would have indicated that a model consistently under- or overestimated observed values. That some MAEs for individual models were larger in absolute magnitude than the corresponding BIAS values tells us that the models vacillated between under- and overestimates. For example, MC1's MAE in watershed IA-1 was 14.29 mm/yr (Table 5), an order of magnitude larger than its BIAS of 1.31 mm/yr (Table 6). The BIAS statistic alone in this case would have suggested the model simulated the observed values set well. Yet, the comparison here of MAE to BIAS suggests there were many errors in the simulated data that tended to cancel each other out with overestimates being nearly equal in magnitude to underestimates, yielding a BIAS much closer to 0 than to the MAE.

The Nash-Sutcliffe coefficient of efficiency, NS, ranged from 0.74 in watershed MS (representing a relatively wet watershed) for Century to −216.70 in watershed AZ-1 (representing a relatively dry watershed) for GTEC (Table 7). Among models, the NS of TEM was closest to 1.0 (0.23). The NS values of the other five models were all negative. When $0 \leq NS \leq 1$, the errors are no larger than the variance. If $NS < 0$, the errors are large relative to the variance. The curvilinear relationship between NS and mean annual HCDN streamflow (Fig. 2e) provides evidence for a threshold around 300–400 mm/yr of runoff. Watersheds whose runoff exceeded the threshold yielded a positive NS and those under the threshold yielded a negative NS. The greatest errors relative to observed variance occurred in the most arid watersheds, as might be expected.

The use of NS, MAE, and BIAS demonstrates the different components of error, but the relative rankings of the six models based on results averaged across all watersheds were relatively consistent (Table 8). TEM was the best performer, and GTEC the worst performer. Biome-BGC and Century were ranked either second or third, depending on which of the error measurements was used; LPJ and MC1 ranked fourth or fifth. Rankings of watersheds, based on averages across all models, were less consistent (Table 9). NS, a measure of relative error, was lower in wetter watersheds. Both MAE and BIAS, which measure absolute error, were lower in drier watersheds.

Long-term trends and interannual variability

The five-year moving average charts show that there was much variability from one watershed to the next

TABLE 6. Extended.

Watershed						
AZ-3	NV	AZ-2	SD	TX	AZ-1	Average BIAS
−1.02	−1.77	−0.38	−0.54	6.03	0.15	−6.79
0.56	−0.34	1.61	0.77	3.56	0.18	−1.46
18.69	10.64	25.58	11.06	13.67	8.27	8.47
−2.86	−0.95	−1.07	−1.08	−0.99	−0.57	−4.80
6.15	1.51	6.32	3.76	9.24	1.20	4.95
−2.07	0.99	−1.01	−0.84	−0.45	−0.78	−3.11
3.24	1.68	5.18	2.19	5.18	1.41	−0.46

during the 60- to 100-year period examined (Fig. 4). Runoff increased both in the actual measured annual runoff and in the simulated annual runoff in watersheds MS, IA-1, IA-2, AZ-2, and NV. The remaining watersheds did not show any long-term trends.

In the majority of watersheds, the models reproduced runoff trajectories qualitatively similar to those of the observed data (Fig. 4). The exceptions seemed to lie in the driest watersheds such as TX, AZ-1, SD, AZ-2, where the simulations produced greater interannual variability than was present in the observed data (these results can also be inferred from Table 7). Quantitative differences among the observed trend and the simulations were apparent in all watersheds. The models underestimated runoff in the NY, MS, WV, and IA-1 watersheds (consistent with Table 6). Overestimates occurred in the remaining watersheds.

DISCUSSION

Examination of the smoothed time series shows that correlations between observed and simulated runoff were similar for each of the models (Fig. 4). But even though the models were able to reproduce temporal variability at this scale, there were considerable differences in magnitude between observed values of runoff and those projected by the models in each watershed. There was also a lot of variability in these differences from one watershed to the next. Overall, the greatest differences in model performance occurred not in a single watershed but from watershed to watershed (Tables 3–7). This is because the models were as a class better able to reproduce observed runoff in wet watersheds than dry ones. While the analysis was limited to 13 watersheds, these watersheds represent a broad range of climatic zones and vegetation types. There was no relationship between watershed size and model performance (data not shown). There is no reason to believe our conclusions about model performance would differ if additional watersheds were examined based on the range of climate regimes represented by the watersheds evaluated.

The BIAS estimator averaged over the entire data set indicated that runoff was just slightly underestimated, even though in the majority of watersheds runoff was overestimated (Table 6). GTEC skewed the watershed results considerably, as is apparent from Figs. 3 and 4 and Table 6. By leaving surface evaporation out of its model, GTEC made more water available to runoff than should have been the case, and this was particularly important in dry areas where potential evapotranspiration would be expected to exceed actual evapotranspiration. Of the remaining five models only MC1 also overestimated runoff in most watersheds. A review of 11 land-surface models (Oki et al. 1999) found systematic underestimation of runoff. The authors primarily attributed this occurrence to the likelihood that rainfall gauges will underestimate precipitation, particularly under conditions of strong wind and during snowstorms. VEMAP's precipitation climatology no doubt reflects similar shortcomings.

There are other reasons for the models to underestimate runoff. For optimal carbon fixation the models maximize leaf area for a given climate, resulting in increases in transpiration and reductions in runoff. Furthermore, the model grid cells lack topography, and topographic relief increases runoff. Review of the literature showed that where available moisture allows actual evapotranspiration rates to approach potential evapotranspiration rates, the methods used to calculate evapotranspiration are more prone to overestimate evapotranspiration than they are under water-limited conditions (Vörösmarty et al. 1998). Finally, in wet watersheds, both saturation excess runoff, and saturated/unsaturated subsurface flow are important components of runoff generation (Atkinson et al. 2002). The VEMAP models do not include these processes, so they may not generate enough runoff.

The tendency for runoff to be overestimated in the dry watersheds may best be explained by the difficulties posed in modeling hydrologic processes in arid and semi-arid regions. In these regions the hydraulic conductivity of soil varies by orders of magnitude as a function of soil moisture. Prediction of hydraulic conductivity, a major control of infiltration capacity, is particularly challenging at the dry end of the soil moisture range. At the wet end, hydraulic conductivity is bounded by a single parameter, its saturation value. At the dry end, however, estimating hydraulic conductiv-

TABLE 7. Nash-Sutcliffe coefficient of efficiency, NS, for the six models and 13 watersheds.

	Watershed						
Model	OR	NY	MS	WV	MT	IA-1	IA-2
BBGC	0.69	−0.32	0.58	0.17	−0.57	0.19	−0.27
CENT	0.50	0.33	0.74	0.62	−0.40	0.01	0.19
GTEC	−0.72	0.27	0.48	0.24	−2.34	−0.98	−2.69
LPJ	−0.30	−0.13	0.38	−0.06	−0.19	−0.84	−0.56
MC1	−0.13	0.41	0.68	0.55	−1.04	−1.04	−1.40
TEM	0.47	0.13	0.65	0.29	0.53	0.32	0.27
Average NS	0.09	0.12	0.59	0.30	−0.67	−0.27	−0.49

Notes: Nash-Sutcliffe (Nash and Sutcliffe 1970) is a dimensionless metric used to evaluate the performance of the models by watershed. It is the ratio of the mean absolute error to the variance in the measured data, subtracted from unity. NS ranges in value from minus infinity (characteristic of a poor model) to 1 (perfect model). The results are ordered by mean annual USGS HCDN streamflow, starting with the wettest watershed on the left and ending with the driest on the right. For key to models, see Table 1.

ity requires knowledge of additional parameters (e.g., Brooks and Corey 1964, van Genuchten 1980) in addition to the soil water content, and hence its simulation is subject to considerable uncertainty. If in the dry watersheds the models underestimated hydraulic conductivities and hence infiltration capacities, too much runoff would have resulted. This tendency could have been exacerbated by the intense precipitation that falls in many arid systems; a single storm can represent a large proportion of annual rainfall. Subgrid variations in moisture content present in natural systems but absent in the VEMAP models may have also contributed to the models' inability to simulate runoff well in dry regions.

The validation of VEMAP-simulated runoff using observed streamflow records is subject to error. None of the models' vegetation maps accounted for 20th century land-use and land-cover changes. We tried to minimize any error this may have introduced into our validations by selecting relatively unimpacted watersheds as the basis of comparison. Nonetheless, ''unimpacted'' watersheds are still likely to have been affected by some human disturbances that would alter hydrologic regimes compared to natural, undisturbed conditions. Whether human disturbance has increased or decreased runoff is unclear. While we assumed the observed streamflow data set was error free, this was likely not to be the case. These data sets may include recording errors from instrumentation, calibration errors, transcription errors, and the like. Moreover, streamflow gauges are notorious for their inaccuracies during times of flooding or drought. Measuring streamflow in mountainous terrain is also error prone as these areas are characterized by high spatial variability. Hence, it is difficult to place error bounds around the observed data set.

We analyzed only the increasing CO_2 scenario because, as mentioned in *Methods*, the two scenarios of increasing and constant CO_2 produced nearly identical runoff results. While CO_2 did increase 20% over the time frame of the historical scenario and we might have expected to see an increase in water-use efficiency, there was little effect of the CO_2 on the water budget produced by the models for several reasons. One reason is that temperature was increasing as well, which serves to temper the water-use efficiency gains present in these models at higher levels of CO_2 (Pan et al. 1998). A 20% increase may not have been sufficiently large to generate a detectable signal; model responses to increasing CO_2 in general are stronger as levels continue to increase (Gordon and Famiglietti, *in press*). For TEM, the hydrologic subroutine is decoupled from the ecosystem model. For the other models, internal feedbacks lead to a reduced response of the water budget to increased CO_2, because as soil water increases due to reduced transpiration, NPP (net primary production) is enhanced (VEMAP Members 1995, Pan et al. 1998).

While it is useful to compare and contrast the functional behavior of the contributing terrestrial ecosystem models, we have not focused on the detailed parameterizations and formulations of the processes being modeled. The formulations of these processes are highly diverse and complex, and a review of the efficacy of these parameterizations is beyond the scope of this paper. Key model differences and limitations are: (1) only TEM, Century, and MC1 account for contributions of rainfall to groundwater, and they do so in a simplistic manner; (2) GTEC overestimates runoff in most watersheds, likely because of the absence of an evaporation function in the model; (3) the absence of snow accumulation and melt processes in GTEC results in temporal errors in runoff; (4) TEM's hydrologic model runs independently of the ecosystem model; and (5) limited representation of belowground hydrologic processes in the VEMAP models probably plays a role in model shortcomings. To expand on this last point, rooting depth, for instance, tends to be shallow in these models with only a limited soil profile available to the plants for water extraction. In studying the effect of rooting depth on simulations of the hydrologic cycle in a tropical catchment, Hagemann and Kleidon (1999) found that use of the deepest rooting depths produced

TABLE 7. Extended.

Watershed						
AZ-3	NV	AZ-2	SD	TX	AZ-1	Average NS
−0.55	−0.09	−0.86	−0.32	−5.48	−14.43	−1.64
−0.03	−0.22	−1.16	−1.45	−2.54	−9.27	−0.98
−11.84	−27.34	−87.33	−58.57	−33.85	−216.70	−33.95
0.07	−0.14	−0.23	−0.11	−0.09	−5.37	−0.58
−2.59	−2.56	−9.74	−12.50	−15.41	−22.80	−5.20
0.49	−0.15	0.37	0.12	0.11	−0.56	0.23
−2.41	−5.08	−16.49	−12.14	−9.54	−44.86	−7.02

simulated values that most closely matched observations.

The three biogeochemistry models that relied on prescribed vegetation, TEM, Biome-BGC, and Century, performed somewhat better than the two biogeography models that generated their own vegetation, LPJ and MC1 (Table 8). The difference is even more compelling when one considers that Century and MC1 share the same hydrologic model. The additional uncertainty introduced by the use of dynamic global vegetation models (DGVMs) could be expected to reduce their accuracy. In a recent study of six DGVMs (Cramer et al. 2001), LPJ simulated a present-day distribution of vegetation types that was not as rich in vegetation classes as a satellite-derived map of contemporary natural vegetation types. Moreover, using VEMAP data, LPJ and MC1 produced present-day vegetation maps differing from one another (National Assessment Synthesis Team 2001). The erroneous placement of vegetation could have profound, localized effects on the water balance. For example, afforestation generally reduces runoff due to increases in evapotranspiration (Bosch and Hewitt 1982). While the static vegetation models may have done a better job of simulating current runoff, as climate and ecological conditions continue to change DGVMs should produce far fewer errors relative to their static vegetation counterparts in modeling future ecosystem and hydrologic regimes.

While the conclusions drawn about relative performance of the six models were not much affected by the measure of performance used (Table 8), differences in model performance from one watershed to the next did depend upon the measure used (Table 9). In particular, models in wet watersheds generated accurate predictions of runoff by three of the five measures of performance used, but MAE and absolute BIAS were greatest in those watersheds. In dry watersheds, the large, negative values of the Nash-Sutcliffe coefficient of efficiency illustrated the difficulties the models had in reproducing observed variance. The tendency of hydrologic models to produce more accurate results in wet watersheds than in dry ones over a range of timescales has been demonstrated elsewhere (Atkinson et al. 2002). It appears that accurate measurements of soil properties are a necessity in dry watersheds, but play a less integral role in wet watersheds.

The trend analysis demonstrated the ability of the models to simulate runoff patterns correctly over the long term even as our other analyses showed the models producing month-to-month errors. Results of our trend analysis are similar to those reported elsewhere in the literature. Hubbard et al. (1997), working with HCDN annual records, reported increases in runoff for 16 of the 20 U.S. Geological Survey-defined water resources regions from 1948 to 1988, with the largest increases occurring in the southwest. Several of the watersheds

TABLE 8. Rankings of each model's relative performance based on overall results from all 13 watersheds, using two metrics and three measures of error.

Model	Pearson *r*	Kendall's τ	MAE	BIAS	NS	Rank by *r*, τ, and MAE	Rank by *r*, τ, and BIAS	Rank by *r*, τ, and NS
BBGC	2	1	3	5	4	2	3	2
CENT	2	3	2	1	3	3	2	3
GTEC	6	6	6	6	6	6	6	6
LPJ	5	5	4	3	2	5	5	4
MC1	4	4	5	4	5	4	4	5
TEM	1	2	1	2	1	1	1	1

Notes: Metrics are identified in the column heads. In ranking, ''1'' corresponds to the best-performing model. Pearson *r* measures the correlation between the observed and simulated runoff. Kendall's τ is a rank correlation used here to assess the degree of congruence in seasonal patterns between observed and simulated values. The measures of error (Nash-Sutcliffe [NS; Nash and Sutcliffe 1970], mean absolute error [MAE], and bias [BIAS]) were used to assess relative and absolute errors (see *Methods: Statistical analyses*). The models were ranked (the last three columns) by averaging performance based on *r*, τ, and either MAE, BIAS, or NS.

TABLE 9. Rankings of watersheds based on ability of the ensemble of six models to simulate runoff.

Watershed	r	τ	NS	MAE	BIAS	Rank by r, τ, and NS	Rank by r, τ, and MAE	Rank by r, τ, and BIAS	Annual HCDN
OR	1	1	4	13	12	2	2	1	1
NY	4	6	3	12	13	4	9	10	2
MS	2	2	1	11	10	1	2	1	3
WV	3	3	2	10	11	3	4	5	4
MT	13	13	7	9	3	12	13	12	5
IA-1	7	8	5	8	5	6	10	6	6
IA-2	7	5	6	7	2	5	5	1	7
AZ-3	5	10	8	6	7	8	7	8	8
NV	11	7	9	3	4	9	7	8	9
AZ-2	9	11	12	5	8	10	11	11	10
SD	6	4	11	2	5	7	1	4	11
TX	12	12	10	4	8	13	12	13	12
AZ-1	10	9	13	1	1	10	6	6	13

Notes: Metrics are identified in the column heads. All rankings are derived from averages of the six models. A ''1'' corresponds to the watershed in which the ensemble of models was best able to reproduce observed patterns of runoff with the smallest errors. Pearson r measures the correlation between the observed and simulated values in each watershed. Kendall's τ is a rank correlation used here to assess the degree of congruence in seasonal patterns between observed and simulated values. The measures of error (Nash-Sutcliffe [NS; Nash and Sutcliffe 1970], mean absolute error [MAE], and bias [BIAS]) were used to assess relative and absolute errors (see *Methods: Statistical analyses*). The watersheds were ranked by averaging model performance based on r, τ, and either MAE, BIAS, or NS. A ranking of the watersheds based on mean annual USGS HCDN streamflow (with ''1'' being the wettest) is provided for reference. For watershed key, see Fig. 1.

examined in this study showed increases in runoff over the time series. In addition, several more exhibited such a trend after 1950. The increases in runoff reported here and elsewhere are consistent with the observation that over the past century there has been a steady increase in the frequency of days with precipitation and in the magnitude of extreme one-day precipitation events (Karl et al. 1996). However, runoff has not increased in western portions of the United States.

Summary

In comparing simulated runoff from six models participating in VEMAP Phase 2 to observed streamflow records from the USGS's Hydro-Climatic Data Network (HCDN) we found small differences in performance among the six models. However, the models yielded highly divergent results depending upon the watershed analyzed. The models exhibited their worst performance in simulating runoff in arid and semi-arid areas and came closest to reproducing the observed data in the wettest regions. The models were able to qualitatively simulate the observed temporal patterns of annual runoff in each of the watersheds over the period of analysis, though the absolute quantity of runoff produced by the models varied widely by model. Using monthly data, we concluded that the three models relying on prescribed vegetation (Biome-BGC, Century, and TEM) outperformed their two DGVMs counterparts (LPJ and MC1), and that GTEC gave the poorest fit to the observations due to the absence of an evaporation function and a snow routine. TEM was the overall best performer, no doubt in large part because its hydrologic model has been independently validated and it ran offline from the ecosystem model. The validation results presented here suggest that improvements in the simulation of hydrologic processes in land-surface models will come, in part, from a more realistic representation of subgrid scale soil moisture and from a more detailed understanding and representation of subsurface processes (e.g., Entekhabi and Eagleson 1989, Famiglietti and Wood 1991, 1994, Hagemann and Kleidon 1999). Moreover, watershed-scale streamflow routing, a topic not considered in this study, may improve the timing of runoff and runoff delivery at watershed outlets (Oki 1999, Olivera et al. 2001).

ACKNOWLEDGMENTS

We appreciate the contributions of the modelers: J.M. Melillo, D.W. Kicklighter, A.D. McGuire, and H. Tian (TEM); D.S. Ojima, W.J. Parton, and R. McKeown (Century); R. Neilson, D. Bachelet, J. Lenihan, and R. Drapek (MC1); M. Sykes, I.C. Prentice, B. Smith, and S. Sitch (LPJ); S. Running and P. Thornton (Biome-BGC); and M. Post and A. King (GTEC). Access to the VEMAP Phase 2 data set was provided by the VEMAP data group within the Climate and Global Dynamics Division, National Center for Atmospheric Research. Development of the VEMAP database was supported by NASA Mission to Planet Earth, Electric Power Research Institute (EPRI), and USDA Forest Service Global Change Research Program. Special thanks are due to Jesse Gordon for programming assistance, David Maidment for help with ArcView, and Ron Neilson, Nan Rosenbloom, and Cristina Kaufman for patiently answering many questions. We gratefully acknowledge the comments of two anonymous review-

ers. This research was made possible by a U.S. Department of Energy Global Change Education Program Graduate Research Environmental Fellowship to W. Gordon and administered by the Oak Ridge Institute for Science and Education, as well as fellowship support from the University of Texas at Austin.

Literature Cited

Abdulla, F. A., D. P. Lettenmaier, 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. Journal of Geophysical Research **10**(D3):7449–7459.

Arora, V. K., F. S. H. Chiew, and R. B. Grayson. 2000. The use of river runoff to test CSIRO9 land surface scheme in the Amazon and Mississippi river basins. International Journal of Climatology **20**:1077–1096.

Atkinson, S. E., R. A. Woods, and M. Sivapalan. 2002. Climate and landscape controls on water balance model complexity over changing timescales. Water Resources Research **38**, 1314, doi:10.1029/2002WR001487.

Bonan, G. B. 1998. The land surface climatology of the NCAR land surface model coupled to the NCAR community climate model. Journal of Climate **11**:1307–1326.

Bosch, J. M., and J. D. Hewitt. 1982. A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. Journal of Hydrology **55**:3–23.

Bradbury, N. J., A. P. Whitmore, P. B. S. Hart, and D. S. Jenkinson. 1993. Modelling the fate of nitrogen in crop and soil in the years following application of ^{15}N-labelled fertilizer to winter wheat. Journal of Agricultural Science **121**:363–379.

Brooks, R. H., and A. T. Corey. 1964. Hydraulic properties of porous media. Hydrology Paper number 3. Colorado State University, Fort Collins, Colorado, USA.

Claussen, M., C. Kubatzki, V. Brovkin, and A. Ganopolski. 1999. Simulation of an abrupt change in Saharian vegetation in the mid-Holocene. Geophysical Research Letters **26**:2037–2040.

Coe, M. T., and G. B. Bonan. 1997. Feedbacks between climate and surface waters in northern Africa during the middle Holocene. Journal of Geophysical Research **102**: 11 087–11 101.

Cramer, W., et al. 2001. Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. Global Change Biology **7**:357–373.

Cramer, W., D. W. Kicklighter, A. Boneau, B. Moore, III, G. Churkina, B. Nemry, A. Ruimy, A. L. Schloss, and the Participants of the Potsdam NPP Model Intercomparison. 1999. Comparing global models of terrestrial net primary productivity (NPP): overview and key results. Global Change Biology **5**(Supplement 1):1–15.

Daly, C., D. Bachelet, J. M. Lenihan, R. P. Neilson, W. Parton, and D. Ojima. 2000. Dynamic simulations of tree–grass interactions for global change studies. Ecological Applications **10**:449–469.

Daly, C., R. P. Neilson, and D. L. Phillips. 1994. A statistical–topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology **33**:140–158.

Dolph, J., and D. Marks. 1992. Characterizing the distribution of observed precipitation and runoff over the continental United States. Climatic Change **22**:99–119.

Entekhabi, D., and P. S. Eagleson. 1989. Land surface hydrology parameterization for atmospheric general-circulation models including subgrid-scale spatial variability. Journal of Climate **2**:816–831.

Famiglietti, J. S., and E. F. Wood. 1991. Evapotranspiration and runoff from large land areas: land surface hydrology for atmospheric general circulation models. Surveys in Geophysics **12**:179–204.

Famiglietti, J. S., and E. F. Wood. 1994. Multi-scale modeling of spatially-variable water and energy balance processes. Water Resources Research **30**:3061–3078.

Gordon, W.S., and J. S. Famiglietti. *In Press.* Trends in VEMAP Phase 2 model simulations of water balance in the 20th and 21st centuries. Global Biogeochemical Cycles.

Hagemann, S., and A. Kleidon. 1999. The influence of rooting depth on the simulated hydrological cycle of a GCM. Physical Chemistry of the Earth (B) **24**:775–779.

Haxeltine, A., and I. C. Prentice. 1996. BIOME3: an equilibrium biosphere model based on ecophysiological constraints, resouce availability and competition among plant functional types. Global Biogeochemical Cycles **10**:693–709.

Hubbard, E. F., J. M. Landwehr, and A. R. Barker. 1997. Temporal variability in the hydrologic regimes of the United States. Pages 97–103 *in* A. Gustard, editor. FRIENDS '97—regional hydrology: concepts and models for sustainable water resource management. International Association of Hydrologic Sciences Press, Wallingford, Oxfordshire, UK.

Hunt, E. R., and S. W. Running. 1992. Simulated dry matter yields for aspen and spruce stands in the North American boreal forest. Canadian Journal of Remote Sensing **18**:126–133.

Jensen, M. E., and H. R. Haise. 1963. Estimating evapotranspiration from solar radiation. Journal of the Irrigation and Drainage Division of the American Society of Civil Engineering **4**:15–41.

Kanji, G. K. 1999. 100 statistical tests. SAGE Publications, London, UK.

Karl, T. R., R. W. Knight, D. R. Easterling, and R. G. Quayle. 1996. Indices of climate change for the United States. Bulletin of the American Meteorological Society **77**:279–292.

Kittel, T. G. F., N. A. Rosenbloom, C. Kaufman, J. A. Royle, C. Daly, H. H. Fisher, W. P. Gibson, S. Aulenbach, R. McKeown, D. S. Schimel, and VEMAP2 participants. 2000. VEMAP Phase 2 historical and future scenario climate database for the conterminous US. [Online: ⟨http://www-eosdis.ornl.gov⟩.]

Kittel, T. G. F., N. A. Rosenbloom, T. H. Painter, D. S. Schimel, and VEMAP participants. 1995. The VEMAP integrated database for modeling United States ecosystem/vegetation sensitivity to climate change. Journal of Biogeography **22**:857–862.

Kittel, T. G. F., J. A. Royle, C. Daly, N. A. Rosenbloom, W. P. Gibson, H. H. Fisher, D. S. Schimel, L. M. Berliner, and VEMAP2 participants. 1997. A gridded historical (1895–1993) bioclimate dataset for the conterminous United States. Pages 219–222 *in* Proceedings of the 10th Conference on Applied Climatology, 20–24 October 1997, Reno, Nevado, USA. American Meteorological Society, Boston, Massachusetts, USA.

Koster, R. D., T. Oki, and M. Suarez. 1999. The offline validation of land surface models: assessing success at the annual timescale. Journal of the Meteorological Society of Japan **77**:257–263.

Küchler, A. W. 1975. Potential natural vegetation of the United States. Second edition. Map 1:3,168,000. American Geographical Society, New York, New York, USA.

Lettenmaier, D. P., E. F. Wood, and J. R. Wallis. 1994. Hydroclimatic trends in the continental United States, 1948–88. Journal of Climate **7**:586–607.

Lewis, D., M. J. Singer, R. A. Dahlgren, and K. W. Tate. 2000. Hydrology in a California oak woodland watershed: a 17-year study. Journal of Hydrology **240**:106–117.

Linacre, E. T. 1977. A simple formula for estimating evapotranspiration rates in various climates, using temperature data alone. Agricultural Meteorology **18**:409–424.

Lins, H. F., and J. R. Slack. 1999. Streamflow trends in the United States. Geophysical Research Letters **26**:227–230.

McCabe, G. J., and D. M. Wolock. 2002. A step increase in streamflow in the conterminous United States. Geophysical Research Letters **29**:2185–2188.

McGuire, A. D., J. M. Melillo, L. A. Joyce, D. W. Kicklighter, A. L. Grace, B. Moore, III, and C. J. Vörösmarty. 1992. Interactions between carbon and nitrogen dynamics in estimating net primary productivity for potential vegetation in North America. Global Biogeochemical Cycles **6**:101–124.

Melillo, J. M., A. D. McGuire, D. W. Kicklighter, B. Moore, C. J. Vörösmarty, and A. L. Schloss. 1993. Global climate change and terrestrial net primary production. Nature **363**: 234–240.

Monteith, J. L. 1973. Principles of environmental physics. Elsevier, New York, New York, USA.

Monteith, J. L. 1995. Accommodation between transpiring vegetation and the convective boundary layer. Journal of Hydrology **166**:251–263.

Nash, J. E., and J. V. Sutcliffe. 1970. River flow forecasting through conceptual models. Part I. A discussion of principles. Journal of Hydrology **10**:282–290.

National Assessment Synthesis Team. 2001. Climate change impacts on the United States: the potential consequences of climate variability and change. Report for the US Global Change Research Program. Cambridge University Press, Cambridge, UK.

Neilson, R. P. 1995. A model for predicting continental-scale vegetation distribution and water balance. Ecological Applications **5**:362–385.

Oki, T., T. Nishimura, and P. Dirmeyer. 1999. Assessment of annual runoff from land surface models using total runoff integrating pathways. Journal of the Meteorological Society of Japan **77**:235–255.

Olivera, F., J. S. Famiglietti, and K. Asante. 2001. Global-scale flow routing using a source-to-sink algorithm. Water Resources Research **36**:2197–2207.

Pan, Y., J. M. Melillo, A. D. McGuire, D. W. Kicklighter, L. F. Pitelka, K. Hibbard, L. L. Pierce, S. W. Running, D. S. Ojima, W. J. Parton, D. S. Schimel, and other VEMAP members. 1998. Modeled responses of terrestrial ecosystems to elevated atmospheric CO_2: a comparison of simulations by the biogeochemistry models of the Vegetation/Ecosystem Modeling and Analysis Project (VEMAP). Oecologia **114**:389–404.

Parton, W. J., D. S. Schimel, C. V. Cole, and D. S. Ojima. 1987. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. Soil Science Society of America Journal **51**:1173–1179.

Parton, W. J., J. M. O. Scurlock, D. S. Ojima, T. G. Gilmanov, R. J. Scholes, D. S. Kammalrut, and J. I. Kinyamario. 1993. Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide. Global Biogeochemical Cycles **7**:785–809.

Parton, W. J., J. W. B. Stewart, and C. V. Cole. 1988. Dynamics of C, N, P and S in grassland soils: a model. Biogeochemistry **5**:109–131.

Peel, M. C., T. A. McMahon, B. L. Finlayson, and F. G. R. Watson. 2001. Identification and explanation of continental differences in the variability of annual runoff. Journal of Hydrology **250**:224–240.

Post, W. M., A. W. King, and S. D. Wullschleger. 1997. Historical variations in terrestrial biospheric carbon storage. Global Biogeochemical Cycles **11**:99–110.

Rastetter, E. B. 1996. Validating models of ecosystem response to global change. BioScience **46**:190–198.

Richardson, C. W. 1981. Stochastic simulation of daily precipitation, temperature and solar radiation. Water Resources Research **17**:182–190.

Richardson, C. W., and D. A. Wright. 1984. WGEN: a model for generating daily weather variables. U.S. Department of Agriculture, Agricultural Research Service, ARS-8, Washington, D.C., USA.

Running, S. W., and E. R. Hunt. 1993. Generalization of a forest ecosystem process model for other biomes, BIOME-BGC, and an application for global-scale models. Pages 141–158 *in* J. R. Ehleringer and C. B. Field, editors. Scaling physiological processes: leaf to globe. Academic Press, San Diego, California, USA.

Sauquet, E., and E. Leblois. 2001. Discharge analysis and runoff mapping applied to the evaluation of model performance. Physics and Chemistry of the Earth **26**:473–478.

Schimel, D., et al. 2000. Contribution of increasing CO_2 and climate to carbon storage by ecosystems in the United States. Science **287**:2004–2006.

Scurlock, J. M. O., W. Cramer, R. J. Olson, W. J. Parton, and S. D. Prince. 1999. Terrestrial NPP: toward a consistent data set for global model evaluation. Ecological Applications **9**:913–919.

Sitch, S., B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. O. Kaplan, S. Levis, W. Lucht, M. T. Sykes, K. Thonicke, and S. Venevsky. 2003. Evaluation of ecosystem dynamics, plant geography, and terrestrial carbon cycling in the LPJ Dynamic Global Vegetation Model. Global Change Biology **9**:161–185.

Slack, J. R., and J. M. Landwehr. 1992. Hydro-climatic data network (HCDN): a U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988. Open-file report **92-129**, U.S. Geological Survey, Reston, Virginia, USA.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. Third edition. W. H. Freeman and Company, New York, New York, USA.

Tian, H., J. M. Melillo, D. W. Kicklighter, A. D. McGuire, J. Helfrich, III, B. Moore, III, and C. J. Vörösmarty. 2000. Climatic and biotic controls on annual carbon storage in Amazonian ecosystems. Global Ecology and Biogeography **9**:315–327.

van Genuchten, M. Th. 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Science Society of America Journal **44**:892–898.

VEMAP Members. 1995. Vegetation/Ecosystem Modeling and Analysis Project (VEMAP): comparing biogeography and biogeochemistry models in a continental-scale study of terretrial ecosystem responses to climate change and CO_2 doubling. Global Biogeochemical Cycles **9**:407–437.

Vörösmarty, C. J., C. A. Federer, and A. L. Schloss. 1998. Potential evapotranspiration functions compared on US watersheds: possible implications for global-scale water balance and terrestrial ecosystem modeling. Journal of Hydrology **207**:147–169.

Vörösmarty, C. J., and B. Moore, III. 1991. Modeling basin-scale hydrology in support of physical climate and global biogeochemical studies: an example using the Zambezi River. Studies in Geophysics **12**:271–311.

Vörösmarty, C. J., B. Moore, III, A. L. Grace, M. P. Gildea, J. M. Melillo, B. J. Peterson, E. B. Rastetter, and P. A. Steudler. 1989. Continental-scale models of water balance and fluvial transport: an application to South America. Global Biogeochemical Cycles **3**:241–265.

Watterson, I. G., M. R. Dix, and R. A. Colman. 1999. A comparison of present and doubled CO_2 climates and feedbacks simulated by three general-circulation models. Journal of Geophysical Research **104**:1943–1956.

Wilcox, B. P., W. J. Rawis, D. L. Brakensiek, and J. R. Wright.

1990. Predicting runoff from rangeland catchments: a comparison of two models. Water Resources Research **26**:2401–2410.

Willmott, C. J. 1984. On the evaluation of model performance in physical geography. Pages 443–460 *in* G. L. Gaile and C. J. Willmott, editors. Spatial statistics and models. D. Reidel, Dordrecht, The Netherlands.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe. 1985. Statistics for the evaluation and comparison of models. Journal of Geophysical Research **90**:8995–9005.

Wolock, D. M., and G. J. McCabe. 1999. Explaining spatial variability in mean annual runoff in the conterminous United States. Climate Research **11**:149–159.

Wood, E. F., et al. 1998. The project for intercomparison of land-surface parameterization schemes (PILPS) Phase 2(c) Red–Arkansas river basin experiment. Global and Planetary Change **19**:115–135.