Ecosystems (2000) 3: 396–411
DOI: 10.1007/s100210000035

ECOSYSTEMS

Constructive Contrasts Between Modeled and Measured Climate Responses Over a Regional Scale

Henrietta I. Jager,[1]* W. W. Hargrove,[1] C. C. Brandt,[1] A. W. King,[1] R. J. Olson,[1] J. M. O. Scurlock,[1] and K. A. Rose[2]

[1]*Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-3036, USA; and* [2]*Louisiana State University, Baton Rouge, Louisiana 70808, USA*

Abstract

Reducing uncertainty in predictions of regional-scale models depends on meaningful contrasts with field measurements. This paper introduces a two-stage process that works from the premise that an appropriate goal for regional models is to produce reasonable behavior over dominant environmental gradients. We demonstrate two techniques for contrasting models with data, one based on the shape of modeled relationships (functional contrasts) and the other based on an examination of the residuals (residual contrasts) between the model and an empirically derived surface fit to field data. Functional contrasts evaluated the differences between the response of simulated net primary production (NPP) to climate variables and the response observed in field measurements of NPP. Residual contrasts compared deviations of NPP from the empirical surface to identify groupings (for example, vegetation classes, geographic regions) with model deviations different from those of the field data. In all model–data contrasts, we assigned sample weights to field measurements to ensure unbiased representation of the region, and we included both constructive comparisons and formal statistical tests. In general, we learned more from constructive methods designed to reveal structure or pattern in discrepancy than we did from statistical tests designed to falsify models. Although our constructive methods were more subjective and less concise, they succeeded in revealing gaps in our understanding of regional-scale processes that can guide future efforts to reduce scientific uncertainty. This was best illustrated by NPP predictions from the Biome-BGC model, which showed a stronger response to precipitation than apparently operates in the field. In another case, differences revealed in savanna and dry woodlands had insufficient field-data support, suggesting a need for future field studies to improve understanding in this, and other, poorly studied ecosystems.

Key words: constructive model validation; residual analysis; regional analysis; regression; net primary productivity.

Introduction

Evaluating regional-scale models in a constructive way presents unique challenges for two reasons. First, regional-scale and local-scale models have different goals (for example, those described by Levins [1966]). Modeling goals are important to consider when designing meaningful comparisons with data. Second, assembling field measurements that can be used in such comparisons on a regional scale often presents practical difficulties. Generality is an important goal for ecological models used to address regional issues. Because they focus on important large-scale patterns, such models are expected to sacrifice local precision in favor of global adequacy. Therefore, it is more important for model predictions to reproduce regional patterns observed in

Received 24 September 1999; accepted 20 April 2000.
**Corresponding author; e-mail:* jagerhi@ornl.gov

nature than to reproduce site-specific field measurements.

Regional-scale contrasts between models and data typically require data collected from a variety of sources to achieve the needed spatial coverage. Overton (1990) used the term "found" data to describe measurements for which the probability of inclusion in a sample is unknown. It is difficult to discern how much weight each measurement should be given in extrapolating to a specified population in the absence of a regional context. Because some types of sites could be overrepresented and others underrepresented, appropriate weighting of each datum is an issue. Here, we adopted Overton's (1990) methodology for placing our found field data onto a common framework. We used a post-hoc sampling design to combine many disparate studies, much in the way a meta-analysis does. Hargrove and Pickering (1992) provide strong arguments that meta-analysis is needed in regional ecology to make progress in discovering regional patterns. Similarly, attention to defining a common regional framework is important for model–data contrasts at a regional scale. In summary, regional-scale contrasts between models and field data share two important concerns: (a) contrasts should measure model quality by the ability to represent large-scale regional patterns in response to underlying environmental gradients, and (b) contrasts should ensure unbiased representation for both model predictions and field measurements at the regional scale.

Our aim is to contrast model predictions against field data in a way that identifies specific model improvements or data needs, rather than merely falsifying and then rejecting a model. The ultimate objective is to reduce model (and scientific) uncertainty by focusing on functional areas of discrepancy. We define a constructive contrast as one that attempts to discover and describe meaningful patterns of differences between model predictions and field data. In this study, we demonstrate two constructive techniques, one based on the shape of modeled relationships (functional contrasts) and the other based on an examination of the residuals between the model and an empirically derived surface fit to field data (residual contrasts). In the functional contrasts, we compare field and model relationships between net primary production (NPP) and environmental gradients of temperature and precipitation. This is motivated, in part, by the assumption that NPP models rely on these two environmental variables as drivers and that feedback about these relationships will be useful to modelers. In the residual contrasts, we elicit constructive feedback through exploratory analysis of residuals. One way to judge the quality of a model is by the absence of pattern in residuals (Zeide 1991; Jager and Overton 1993). The goal of our residual analysis is to find patterns in model–field data fit by grouping or displaying residuals in informative ways.

Our approach builds on the evaluation of Vegetation/Ecosystem Modeling and Analysis Project (VEMAP) models by Schimel and others (1997). Schimel and others conducted both site-specific and regional comparisons with VEMAP models: (a) site-specific comparisons with OTTER transect and Konza and CPER Long-term Ecological Research sites, and (b) regional comparisons of NPP with the Normalized Difference Vegetation Index (iNDVI). These authors found that successful validation of the three VEMAP models at particular sites did not guarantee successful simulation of spatial variability when compared with regional iNDVI. Although site-specific validation was valuable for pinpointing mechanisms behind model–data discrepancies, comparison with comprehensive spatial data was needed to evaluate regional-scale patterns of spatial variation in NPP. They recommended using regional-scale spatial data, even when estimates of NPP required extrapolation or indirect measurement.

Net Primary Productivity Data and Models

VEMAP Models

We illustrate our methods with estimates of net primary productivity simulated in Phase I of the Vegetation/Ecosystem Modeling and Analysis Project (VEMAP) (VEMAP Members, 1995; *http://www.cgd.ucar.edu:80/vemap*). VEMAP compares simulations of biogeography and biogeochemistry models for the conterminous US under conditions of contemporary and future atmospheric CO_2 and climate. The VEMAP Phase I biogeochemical models, BIOME-BGC (Hunt and Running 1992; Running and Hunt 1993), CENTURY (Parton and others 1987; Parton 1996), and the Terrestrial Ecosystem Model (TEM) (Raich and others 1991; Melillo and others 1993), simulate cycles of carbon, nitrogen, and water in terrestrial ecosystems. The models represent the influence of climate and other environmental variables on the dynamics of these cycles. The potential vegetation of VEMAP is derived from Kuchler (1964; 1975) and is aggregated into coarser vegetation categories associated with each grid cell. All of the models simulate NPP as the difference between gross carbon uptake and plant respiration, but they represent these processes in different ways

Table 1. Sources of Measured NPP Data in the US

Dataset description	Sample size	Reference
Sites representing various biomes for the International Biome Program (IBP)	10	DeAngelis and others 1981
The Osnabruck collection	116	Esser and others 1997
Sites in the Superior National Forest, Minnesota	63	Hall and others 1992
OTTER transect sites in Oregon	7	Runyon and others 1994
Estimates for each Major Land Resource Area (MLRA)-State in rangeland and grassland	100	Sala and others 1988

with different levels of mechanistic or process-based detail. For example, the BIOME-BGC simulates carbon uptake by vegetation with a submodel of daily canopy photosynthesis based on leaf biochemistry; CENTURY simulates carbon uptake through environmental limitations on monthly maximum plant production. A detailed comparison of the model's representation of NPP can be found in VEMAP Members (1995).

The VEMAP models have continued to evolve since they were used to generate Phase I predictions in 1997 (W. J. Parton, D. McGuire, P. Thorton personal communication). Therefore, the results reported here do not apply to later versions of the models and are presented mainly as an illustration. Our contrasts are based on the sample of 3,168 VEMAP Phase I predictions taken at 0.5° VEMAP grid cell centers that span the conterminous US. The sample excludes grid cells centered on water bodies or wetlands. The VEMAP models make NPP predictions for the potential vegetation of each grid cell.

Field Data

Our database of productivity measurements currently includes five datasets (Table 1) (Scurlock and others 1999). In this comparison, we used information on total NPP in units of g C m^{-2} y^{-1}. For the Major Land Resource Area-State dataset (Sala and others 1988), we converted values for aboveground NPP provided in g dry weight $m^{-2}y^{-1}$ to units of g C $m^{-2}y^{-1}$ by the factor 0.45 (Hall and Scurlock 1991). Many studies measured aboveground but not belowground NPP. A smaller number of studies measured belowground but not aboveground NPP. For sites that lacked total NPP, we imputed (estimated) total NPP based on linear relationships with no intercept that were derived from sites with both measurements. These relationships differed significantly between vegetation types dominated by grasses (Eq. [1]; $n = 17$) and those dominated by trees (Eq. [2]; $n = 29$). In these equations, the standard error on the slope estimate is shown in parentheses.

$$\text{Total NPP} = 3.02\ (\pm 0.19)\ \text{Aboveground NPP}$$

$$\text{Total NPP} = 1.44\ (\pm 0.04)\ \text{Belowground NPP} \quad (1)$$

$$\text{Total NPP} = 1.39\ (\pm 0.07)\ \text{Aboveground NPP}$$

$$\text{Total NPP} = 2.34\ (\pm 0.24)\ \text{Belowground NPP} \quad (2)$$

We included 296 NPP measurements in our contrasts. Most of these field measurements represented grassland, shrubland, and forest vegetation types. We excluded field measurements belonging to vegetation classes indicating an agricultural land use (for example, plantations) or wetlands because the models did not predict NPP for these vegetation classes.

To describe relationships between field-measured NPP and environmental gradients, we needed a consistent source of environmental data. For each field site, we interpolated values of three environmental variables, total annual precipitation, mean annual temperature, and vegetation class, from a 1×1 km grid. We obtained orographically corrected total annual precipitation from spatial data at a resolution of 4×4 km. Similarly, we obtained mean annual temperature, corrected for elevation, from US National 1961–91 Climate Normals measured at 4761 National Climatic Data Center meteorological stations. We compared estimates from the 1-km grid source to site-specific values and found reasonable agreement. We used national vegetation maps (Kuchler 1975) digitized to a 1-km resolution to assign a vegetation type to measurements lacking this information.

Table 2. Criteria for Distinguishing Constructive from Nonconstructive Methods for Contrasting Models against Field Data

Nonconstructive	Constructive
Test statistics and associated probabilities are the main results presented.	Visual patterns and relationships are the main results presented.
Model–data comparisons are not broken down into meaningful groups.	Model–data comparisons are broken down into meaningful groups to explore differences in model–data correspondence.
Choices of field data and model settings are restricted, and paired comparisons are used to improve power.	Choices of field data and model settings are broad to reveal behaviors under a wide range of conditions.

Comparability of Model Predictions and Field Data

Comparability of a collection of field data with model predictions is a nontrivial issue in contrasts at a regional scale. Several questions about the comparability of model predictions and field data emerged in this analysis. First, do model predictions apply to the same spatial and temporal scale as the field measurements? The VEMAP models considered here represent dynamics at a spatial scale comparable to that usually measured in the field (for example, the stand level). VEMAP I simulations were produced from temperatures for grid cell centers and cell-averaged precipitation. Field measurements were obtained at a small spatial scale.

Second, do field measurements of NPP from sites represent potential vegetation, which is what the VEMAP models predict? Because our approach does not involve matching sites geographically, the fact that a particular field site belongs to a different vegetation type from the VEMAP cell that it inhabits is not an issue. However, the concern that recently disturbed sites have higher productivity than more mature sites is valid. We excluded plantations and agricultural sites with obviously altered disturbance regimes and species composition. However, our data might include field measurements of NPP for natural vegetation on sites previously exposed to unnatural disturbance regimes caused by fire suppression, logging, or similar activities. VEMAP models calibrated against unusually productive research sites might share this bias (Schimel and others 1997).

Third, the field data were assigned climate estimates from a different source than the VEMAP predictions. For the field data, we associated the best local estimates of climate available to us. For the VEMAP predictions, we associated climate values used by the VEMAP models as input. This is the appropriate approach for functional contrasts because geographic consistency is less important than obtaining the best characterization of (model and field) relationships between NPP and climate variables. The two sources of climate data are highly correlated ($\geq$0.94).

Methods

We used two main methods to contrast VEMAP models with field data: (a) functional contrasts and (b) residual contrasts. For both approaches, we used sample weights to ensure that our results applied to the correct regional universe, and we used an empirical model to describe the relationships between NPP and two environmental gradients. We first describe our method for assigning sample weights. Next, we describe fitting of the empirical models involving climate variables to both VEMAP models and field data. Finally, we present the two contrasts. Because we intersperse less constructive tests with our constructive methods, please refer to Table 2 as a roadmap for determining whether a particular method is, or is not, constructive.

Regional-Scale Model–Data Contrasts

Defining a universe of inference is no less important in contrasting models with field data than it is in any other scientific endeavor. The universe of inference defines the population to which scientific results apply. Our goal was to place both the NPP field data and the VEMAP model NPP predictions onto a common universe. VEMAP model predictions were reported for 0.5° grid cells covering the conterminous US that do not fall in water bodies or wetlands. Because VEMAP predicts NPP for potential vegetation, the VEMAP universe is abstract. Still, it is well defined, whereas the field samples were collected for myriad purposes and not according to a sample design that would facilitate making regional inferences (Figure 1). We adopted the VEMAP sampling framework as our universe; therefore, our results apply to areas with potential vegetation in the conterminous US, excluding wetlands.

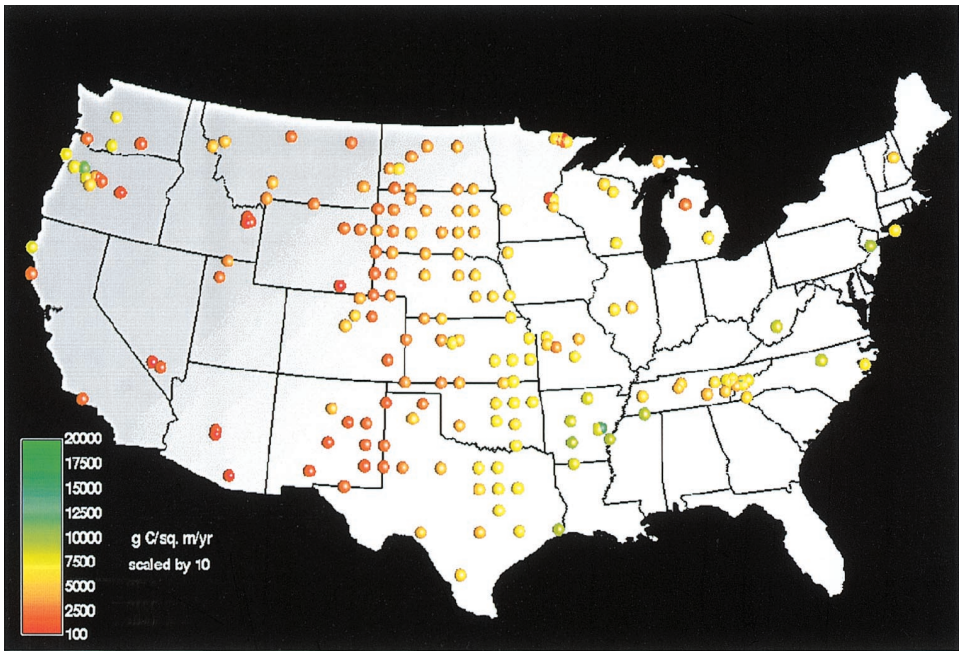

Figure 1. Map of the conterminous US showing locations of NPP measurements used as part of our field data and overlain by the VEMAP grid.

The VEMAP grid's climatic data provided information needed to tie our "found" NPP sample to the regular grid sample design used for model predictions. Overton (1990) and Overton and others (1993) developed a method for combining a found sample with a probability sample (that is, one with a specified universe of inference and known probabilities of inclusion for each measurement). Following this approach, we selected two poststratification variables: mean annual temperature and total annual precipitation. Univariate quintiles, calculated from the VEMAP grid (Table 3), defined 25 climatic strata. We assigned each field measurement a sample weight inversely proportional to its probability of inclusion in a given climate stratum. These sample weights (w) are the same for all measurements falling into a climate stratum. They correct the field sample for deviations from what we would expect to find in an equal-probability sample of the US. We calculated the sample weights as:

$$w = \frac{\%\ \text{grid cells in a stratum}}{\%\ \text{field observations in a stratum}} \qquad (3)$$

We used weighted least squares involving sample weights to describe NPP relationships with temperature and precipitation. This ensured that the estimated regression relationships were unbiased with respect to the intended universe of inference. The weighted least-squares estimator is often employed in linear regression using complex survey data to counteract the bias in ordinary least squares arising from informative sampling (Kott 1991; Korn and Graubard 1995; Magee 1998).

Empirical Model

The Miami model form was first used as an empirical description of worldwide patterns in NPP by Lieth (1975). This empirical function serves two purposes in our study. First, our functional contrasts use this empirical form to describe the relationship between NPP and environmental gradients. Second, we use the same empirical model in our residual contrasts to generate a response surface of NPP. This response surface, referred to hereafter as the "field-data surface," stands in for field measurements when we compute residuals (model-predicted NPP minus field-data surface NPP). The field-data surface makes it possible to compare imputed data values with model predictions at all locations where model predictions are available, regardless of whether a location lacks field measurements or is no longer covered by natural vegetation. It gives us a functional way of comparing model predictions to field data with similar climate.

We adopted the following empirical model of mean annual temperature (T) and total annual precipitation (P):

$$NPP_T = \frac{N_{max}}{1 + e^{Y(T)}} \qquad (4)$$

$$NPP_p = N_{max}(1 - e^{Y(P)}) \qquad (5)$$

Table 3. Sample Weights in Each of the 25 Temperature and Precipitation Strata Formed by Quintiles of the Grid Population

		Temperature quintile (°C)				
		<5	5–7.1	7.1–9.9	9.9–14.2	>14.2
Precipitation quintile	<403	44.0	11.5	10.4	9.9	5.6
	403–571	33.2	11.8	7.8	11.3	5.0
	571–889	2.32	19.0	15.0	8.0	4.9
	889–1183	38.0	—	53.5	7.4	9.8
	>1183	2.75	—	4.1	2.8	20.6

Combinations lacking field data are indicated by dashes.

$$NPP = \min(NPP_T, NPP_P) \quad (6)$$

We adopted the Miami model form for three reasons. First, the Miami form assumes that one environmental factor, the "limiting factor," controls productivity. When temperature is low, temperature limits production and when precipitation is low, precipitation controls the rate of production. The Miami model takes the minimum of two NPP estimates: NPP_T is a nonlinear function of mean annual temperature and NPP_P is a nonlinear relationship with total annual precipitation.

Second, the asymptotic form of the Miami, which reaches a maximum NPP, N_{max} at high values of precipitation and temperature, is more appropriate than a linear model over a wide range of temperature and precipitation. Others have found a linear relationship between NPP and precipitation within particular regions or vegetation types (for example, Sala and others 1988). Our field data span a wider range of climatic conditions, although the relationships of subsets of data limited by one factor did not deviate much from linear.

Third, the Miami form gave better predictions of field NPP than alternative models that we considered. It explained 46% of variation, with much better predictions at precipitation-limited sites (>90%) than at temperature-limited sites. The Montreal model form (Lieth 1975), which describes NPP as a function of actual evapotranspiration (AET), explained 32% of the variation in field NPP when we used an average of VEMAP estimates of AET. The Montreal model form explained only 18% of variation in field NPP when we used AET from Ahn and Tateishi (1994).

Functional Contrasts

In our first analysis, we contrasted the relationship between NPP predicted by the VEMAP models and environmental variables with the relationship between field NPP and environmental variables. We used both formal tests (indicator analyses) and visual comparisons (residual analyses). Indicator analysis is a deductive method for testing the null hypothesis that the VEMAP models and field data share the same environmental relationships. Residual analysis is an inductive method that visually compares model and field relationships to find patterns in discrepancies. This latter is an example of a constructive contrast—one that identifies how the model (or data measurement) might be improved, rather than merely attempting to falsify the model.

Indicator analysis. Linear regression with an indicator or dummy variable (indicator analysis) allows different parameters to be estimated for different groups (Draper and Smith 1981). In our analysis, the indicator variable, I_v, refers to either the VEMAP observations or the field data Eq. (7).

$$I_v = \begin{cases} 1, & \text{when VEMAP} \\ 0, & \text{when field data} \end{cases} \quad (7)$$

We can formally test the hypothesis that models and data share the same environmental relationships by comparing slopes of a linear regression model that includes I_v.

$$Y(T) = \alpha_0 + \alpha_1 I_v + \alpha_2 T + \alpha_3 (I_v * T) \quad (8)$$

$$Y(P) = \beta_2 P + \beta_3 (I_v * P) \quad (9)$$

By substituting for I_v, the indicator model above reduces to the following models for field and VEMAP-predicted NPP:

$$Field\ Y(T) = \alpha_0 + \alpha_2 T \quad (10)$$

$$Field\ Y(P) = \beta_2 P \quad (11)$$

$$VEMAP\ Y(T) = (\alpha_0 + \alpha_1) + (\alpha_2 + \alpha_3) T \quad (12)$$

$$VEMAP\ Y(P) = (\beta_2+\beta_3)\mathrm{P} \qquad (13)$$

The technique of regression with an indicator variable requires balance in sample sizes between the field data and VEMAP predictions. To address this, we divide the VEMAP predictions of NPP randomly into 11 subsets, with each subset having roughly the same sample size as the field data. This ensures that the degrees of freedom correctly reflect the ability to test for significant differences between model- and field-based parameter estimates.

We use the following procedure to estimate parameter values: First, we split the field data into two parts, one limited by temperature and one limited by precipitation. Because the data must be split prior to estimating the parameter values, we use the parameter values originally estimated by Lieth (1975) to determine, for each observation (field or model), whether the temperature or precipitation estimate of NPP would be lower. Next, we adopt Lieth's value of maximum NPP, $N_{max} = 1350$ g C $m^{-2}y^{-1}$. Finally, we linearize the equations by applying the appropriate logit transformation (Atkinson 1985) to NPP (Eqs. [14] and [15]). This allows us to use linear regression to estimate the remaining parameters (αs, βs). We estimate parameters for Eq. (14) from the subset of data limited by temperature and for Eq. (15) from the subset of data limited by precipitation.

$$Logit(\mathrm{NPP_T}) = \log_e(N_{max}-NPP_T)-\log_e(NPP_T) = Y(T) \qquad (14)$$

$$Logit(\mathrm{NPP_P}) = \log_e(N_{max}-NPP_P)-\log_e(N_{max}) = Y(P) \qquad (15)$$

If the VEMAP model relationship between NPP and temperature differs from that observed in the field, then α_1 and α_3 should be significantly different from zero. Likewise, an estimate of β_3 significantly different from zero would indicate that the VEMAP model relationship with precipitation differs from that observed in the field data.

We present three kinds of information to compare the relationships. First, we present tests for significant differences between model and field regression coefficients. We report the average probability of rejecting a two-sided *t*-test of $\alpha_1 = 0$, $\alpha_3 = 0$, and $\beta_3 = 0$ over the 11 replicate subsets of VEMAP predictions. Second, we compare the curves estimated for each model and for the field data. This comparison shows how differences in parameter values can translate into differences in NPP that might not be apparent from looking at the parameter estimates themselves—it allows us to assess ecological, in addition to statistical, significance. Third, we graph the VEMAP NPP predictions along with the 75% and 95% prediction intervals from the field data relationships. Typically, these prediction intervals are used to indicate the degree of uncertainty associated with predictions at a particular value of x. Here, we wish to illustrate the variability in the field data around the field-data surface without confusing the graphs with too many overlaid points. The prediction intervals serve as reasonable reference lines: the 75% prediction interval encompasses 80% of the field data points, and the 95% prediction interval encompasses 94% of the field data.

Residual Contrasts

Our second approach seeks to uncover patterns in residuals to elucidate model–data discrepancies. The purpose of this analysis is to identify climatic conditions leading to better and worse agreement between model-predicted and measured NPP that can help to guide future research. A two-step process is required for this: (a) form residuals based on the field-data surface, and (b) compare model residuals with data residuals using covariates to help reveal structure in discrepancies.

Because we do not have paired comparisons between field measurements and model predictions, we need a way to compare field data with appropriate model predictions. Rather than use spatial proximity as the basis for matching points, we substitute a conditional expectation for the field NPP, given that the temperature and precipitation at each VEMAP grid point is known. In other words, we use our field-data surface to predict NPP based on temperature and precipitation provided with each of the VEMAP grid points. The field-data surface serves as a reference. The residuals obtained in this first step are differences between each VEMAP model prediction of NPP and the NPP estimated from climate by using the field-data surface.

In the second step, we compare VEMAP residuals with residuals of the field data from the field-data surface. We obtain field-data residuals by subtracting NPP estimated from climate by using the field-data surface from field NPP. Residual contrasts are comparisons between the patterns of residual observed in the VEMAP model predictions with those in the field data. From a deductive point of view, this comparison indicates whether deviations of VEMAP predictions from the field–data surface are significant in light of uncertainty in the field–data relationship. From a constructive point of view, any differences between model and data residuals are of

potential interest. For example, smaller deviations in model than data might also provide insights.

First, we summarize residuals to test the overall ability of the field-data surface to predict VEMAP model results. This preliminary residual analysis evaluates overall fit. We evaluated the distribution of residuals, e_i for (a) bias (nonzero mean error) and (b) accuracy (nonzero mean absolute error).

$$mean\ error = \frac{1}{n}\sum_{i=1}^{n} e_i \qquad (16)$$

$$mean\ absolute\ error = \frac{1}{n}\sum_{i=1}^{n} |e_i| \qquad (17)$$

This analysis, by itself, is not likely to be particularly constructive because it does not provide information about when agreement is good and when it is poor. Therefore, our next analysis looks for patterns in the residuals that uncover the reasons for model–data discrepancies. To illustrate this approach, we organize the residuals by vegetation classes. It is conceivable that different parameterizations of the VEMAP models or different sampling methodologies in different vegetation classes could lead to model–field data differences. For each group, we compare the distribution of VEMAP model residuals with the distribution of field-data residuals.

Finally, we map the residuals to evaluate geographic patterns of fit between the VEMAP models and the field-data surface. The residuals are differences between each VEMAP model prediction of NPP and the NPP obtained using the field-data surface.

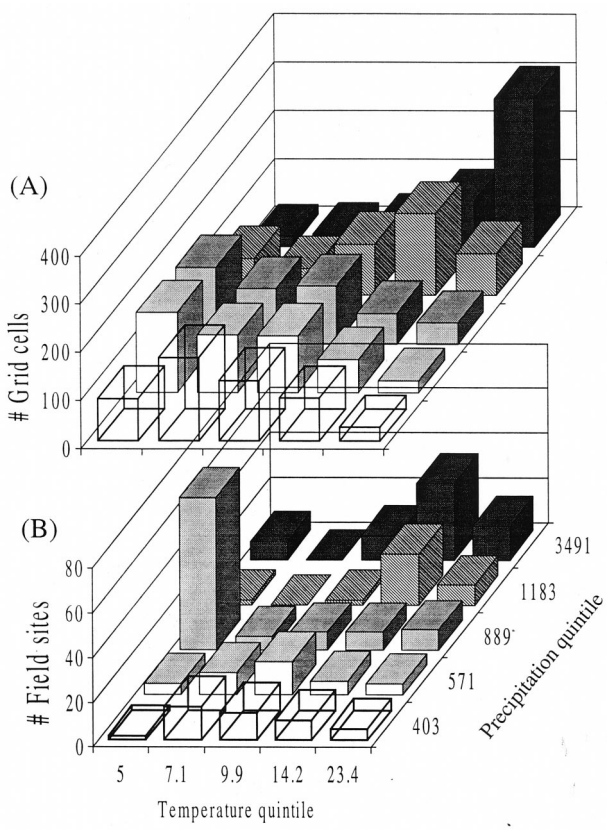

Figure 2. Frequencies of (**A**) grid cells and (**B**) field observations in each climate stratum defined by univariate quintiles of the temperature and precipitation distributions of VEMAP grid cells.

Results

Regional-Scale Model–Data Contrasts

The frequency counts of VEMAP grid points in environmental space showed that mean annual temperature and total annual precipitation were correlated in their spatial occurrence. Grid cells with high temperature and high precipitation were relatively common (Figure 2A), whereas those with low temperature and high precipitation or high temperature and low precipitation were rare. Combinations with low temperature or low precipitation were uncommon in the field data, while one combination that includes a number of measurements from the Superior National Forest stood out as being overrepresented (Figure 2B), prior to weighting. The sample weights are listed in Table 3. The weighted cumulative frequency distributions of temperature and precipitation (Figure 3) illustrate the improvement gained in representing US climate by applying sample weights. The weighted distributions of temperature and precipitation are closer to those of the VEMAP grid than are the unweighted distributions.

Functional Contrasts

The field data showed significant positive relationships between NPP and both temperature and precipitation (Figure 4). The Miami form of empirical model is visually consistent with the distribution of field data along the two gradients.

The results of our indicator regression showed differences in environmental relationships between the VEMAP Phase I model predictions and field measurements. Although we detected statistically significant differences ($P < 0.01$) in the NPP vs precipitation relationships for all three VEMAP models (Table 4), visual inspection suggests that the relationship between TEM's NPP and precipitation

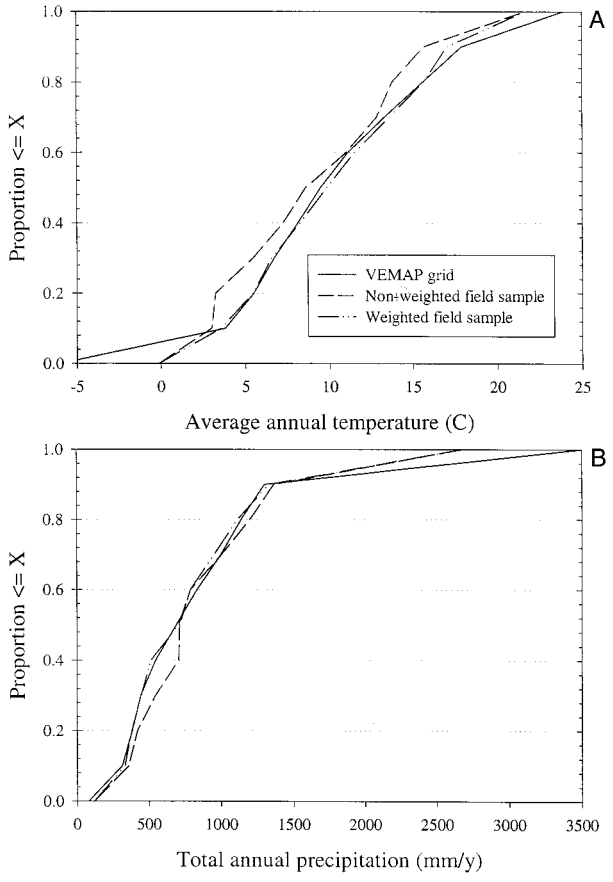

Figure 3. Comparison of the VEMAP grid distributions with weighted and unweighted field-sample distributions of (**A**) average annual temperature and (**B**) total annual precipitation.

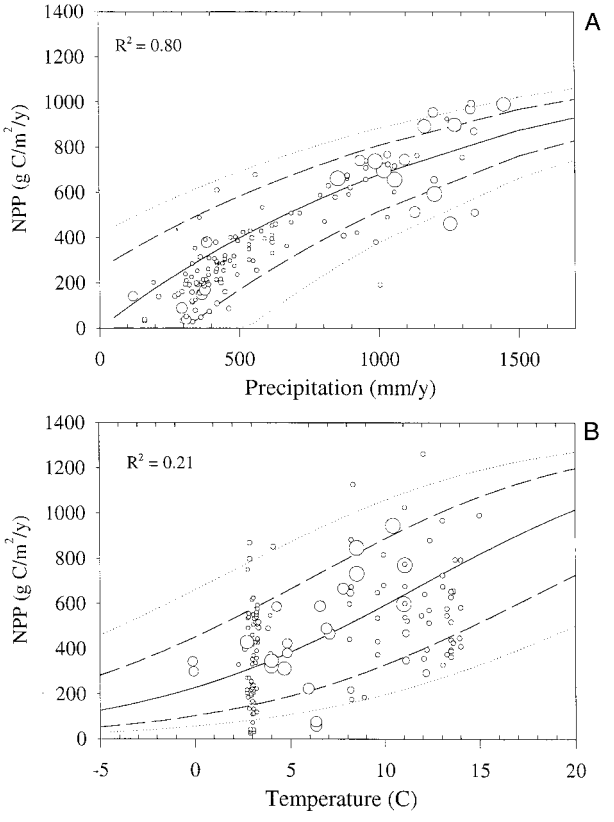

Figure 4. Field NPP measurements and their relationship with (**A**) total annual precipitation and (**B**) mean annual temperature. Three lines indicate the fitted empirical models (solid), the 75% prediction interval (dashed lines encompass 80% of the field data), and the 95% prediction interval (dotted lines encompass 94% of the field data). Each graph contains only the field data limited by the factor plotted on the x-axis. The size of each symbol reflects the sample weight.

(Figure 5) is consistent with that of field measurements (Figure 4). The indicator test focuses on its certainty that the true coefficients of precipitation differ. The ecological relevance of the test is limited because it does not account for the considerable variation in NPP estimates. In essence, this describes the distinction between a confidence and prediction interval, where the latter has higher ecological relevance but is more difficult to summarize by a test.

The NPP–precipitation relationship produced by the BIOME-BGC and CENTURY models differed qualitatively from the field-data relationship (the heavy lines in Figures 6 and 7 show the empirical model fit to VEMAP predictions). For BIOME-BGC, the precipitation response of NPP is concave upward rather than convex (Figure 6). We observed a sharp increase in predicted NPP for sites with total annual precipitation higher than 800 mm/y that we did not see in the field data (Figure 4).

For CENTURY, the qualitative differences were subtler (Figure 7). The field data showed a random pattern surrounding the precipitation component of the field-data surface (Figure 4). In contrast, most CENTURY predictions seemed to follow a small number of curves, with very little variation surrounding each curve. Although the CENTURY model predictions generally fell within the bounds defined by the field data (Figure 7), the patterns shown by CENTURY model predictions seemed to represent at least two distinct functional responses to precipitation. One of these, a steeply inclining curve in response to precipitation, deviated significantly from the overall shape and magnitude of response followed by the field data. Upon further examination, it appeared that this response was typical of CENTURY model predictions for savanna grid cells (Figure 7A). Given the variation in the field data illustrated by the prediction intervals in Figure 5, the climatic relationships predicted by the

Table 4. Parameter Estimates for an Indicator Regression Model that Compares the Relationship of VEMAP NPP Predictions to Climatic Variables with Those of Field NPP Measurements

				Phase I VEMAP model		
			Field data	BIOME-BGC	CENTURY	TEM
Limiting environmental factor	Precipitation	Sample size	166	2160	2267	2267
		Coefficient	–0.0007	–0.0004	0.00009	0.00007
		$P > \|T\|$	<0.0001	<0.0001	0.0046	0.0013
		Sample size[a]	128	878	901	901
		Intercept	1.5959	–0.4179	–0.6699	0.0905
	Temperature	$P > \|T\|$	<0.0001	0.1358	0.0049	0.4943
		Coefficient	–0.1353	–0.0487	0.0476	–0.0364
		$P > \|T\|$	<0.0001	0.1915	0.1350	0.2582

*Three parameter estimates are listed: (a) the intercept of the temperature model (the I_v term in Eq. [8]), (b) the temperature coefficient (the I_v*T term in Eq. [8]), and (c) the precipitation coefficient (the I_v*P term in Eq. [9]). Probabilities ($P > |T|$) test for a parameter that is significantly different from zero, indicating that the model and field data have different relationships.*

[a]We excluded 23 temperature-limited BIOME-BGC predictions, two temperature-limited field measurements, and seven precipitation-limited BIOME-BGC predictions because they exceeded the maximum NPP value of 1350 g C/m^2/y used by the logit transform.

TEM model compared very well with those of the field data. In addition, TEM and the field data showed similar patterns of residual variation (Figures 4 and 5).

The CENTURY model response to temperature was weaker than that indicated by the field data (Figure 7). The significant interaction between the intercept parameter and I_v in Table 4 (average $P = 0.0049$) revealed a significant difference in the NPP–temperature relationship between the CENTURY model predictions and the field-data surface. We did not detect significant differences between temperature relationships of TEM or BIOME-BGC NPP predictions and that of the field-data surface NPP (Table 4).

Residual Contrasts

The average bias of each VEMAP model, the mean difference between VEMAP NPP and field-data surface NPP Eq. (16), was significantly different from the average bias of the field data ($P < 0.01$) (Table 5). Although the mean absolute errors were significantly different from zero for all VEMAP models, only the BIOME-BGC model had a significantly higher mean absolute error than the field data. The distribution of residuals illustrates that NPP predictions from the TEM model showed the best agreement with the field-data surface, but they were also less variable than the field data (Figure 8). The BIOME-BGC model predictions were unbiased relative to the field-data surface, but they deviated from the field-data surface more than the other VEMAP models or the field data (Figure 8). BIOME-BGC model predictions showed higher variability than did the field data. Because climatic variables driving the VEMAP models are averages for a grid cell, we would expect variation in VEMAP model NPP predictions to be lower than that observed. However, this is probably compensated for by numerous factors that influence individual sites but that are not represented in the VEMAP models.

Residuals structured by vegetation class. We continued with a structured residual analysis to identify patterns that might provide modelers and field ecologists with more information on where the largest discrepancies occurred. Table 6 lists the sample sizes associated with each vegetation type, both for the VEMAP grid and for the field data. The TEM model exhibited the smallest bias and absolute error (Figure 9). CENTURY predictions deviated most from the field-data surface in broadleaf forest, where they tended to be high. On average, BIOME-BGC predictions were high in forest classes, with a number of very high NPP predictions.

Residuals structured geographically. Maps of residuals showed distinct geographic patterns. In general, when predictions by the VEMAP models differed from the field-data surface, they tended to predict lower NPP than indicated by the field-data surface in western regions of the US and higher NPP than indicated by the field-data sur-

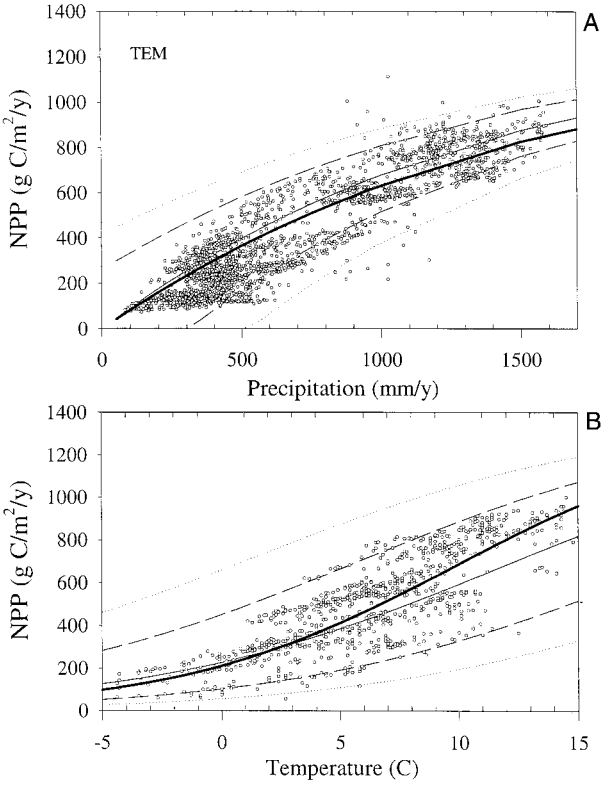

Figure 5. TEM model NPP predictions and their relationship with (**A**) total annual precipitation and (**B**) mean annual temperature. Three lines indicate the fitted empirical models Eqs. (10) and (11) (solid), the 75% (dashed), and the 95% (dotted) prediction interval. The heavy line follows Eqs. (12) and (13) fitted to TEM NPP predictions. Each graph contains only the grid points limited by the factor plotted on the x-axis.

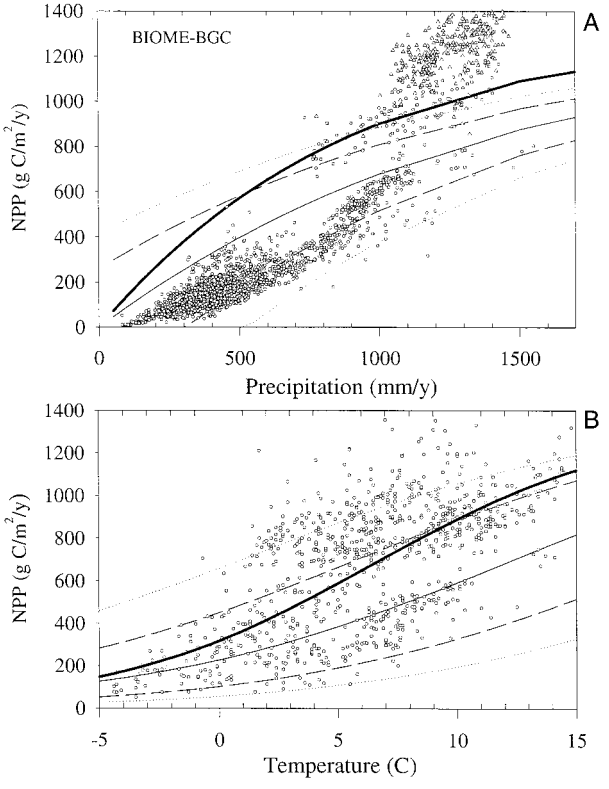

Figure 6. BIOME-BGC model NPP predictions and their relationship with (**A**) total annual precipitation and (**B**) mean annual temperature. Three lines indicate the fitted empirical models Eqs. (10) and (11) (solid), the 75% (dashed), and the 95% (dotted) prediction interval. The heavy line follows Eqs. (12) and (13) fitted to BIOME-BGC NPP predictions. Each graph contains only the VEMAP grid points limited by the factor plotted on the x-axis. Predictions for mixed forest are indicated in (**A**) by open triangles.

face in the northeastern and upper Midwest regions (Figure 10B–D). In the highly productive Pacific Northwest region and in northern Idaho, two VEMAP models estimated higher NPP than the field-data surface (Figure 10C and D, respectively). The BIOME-BGC model showed a striking pattern of predicting lower NPP than indicated by the field-data surface in most of the West and predicting higher NPP than indicated by the field-data surface in the eastern US relative to the field-data surface (Figure 10D). CENTURY model NPP compared well with the field-data NPP surface in the eastern US (Figure 10C). CENTURY predictions were generally lower than the field-data surface in the West, higher in the Midwest, and close to the surface in the Northeast. Of the three VEMAP models, the TEM model compared most favorably with NPP field-data surface (Figure 10B).

Discussion

Constructive Feedback

The main objective of constructive contrasting is to generate feedback that can be used to reduce model uncertainty through future data collection, experimentation, or model refinement. In this example, model–data contrasts suggested that the BIOME-BGC model predictions produced an exaggerated response to high precipitation. This exaggerated response was most clear in the graph showing a strong increase in predicted NPP along a precipitation gradient (Figure 6). It is also a likely explanation for the strong east–west geographic pattern that we observed in the residuals. Schimel and others (1997) also noted that the BIOME-BGC model produced substantial overestimates at the two wettest sites in their comparison with sites along the

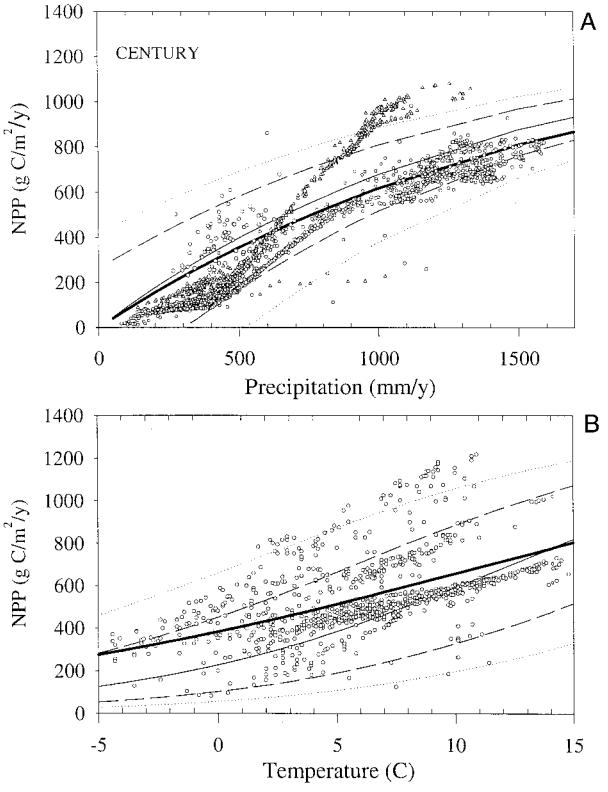

Figure 7. CENTURY model NPP predictions and their relationship with (**A**) total annual precipitation and (**B**) mean annual temperature. Three lines indicate the fitted empirical models Eqs. (10) and (11) (solid), the 75% (dashed), and the 95% (dotted) prediction interval. The heavy line follows Eqs. (12) and (13) fitted to CENTURY NPP predictions. Each graph contains only the grid points limited by the factor plotted on the x-axis. Predictions for savanna or dry woodlands are indicated by open triangles in (**A**).

OTTER transect. We observed that the highest predictions of BIOME-BGC occurred in forest vegetation. NPP predictions in grassland, shrubland, and savanna fell below the field-data surface, perhaps because of limitation of NPP by precipitation in the BIOME-BGC NPP model.

The CENTURY model results highlighted the constructive nature of the visual residual contrasts. Overall, the CENTURY model's NPP predictions fell within statistical bounds set by the field data—so, in a sense, the model predictions meet quality assurance specifications that might be defined by an engineer. Yet, visual inspection of the predicted relationship of NPP with precipitation revealed that the CENTURY model predictions were following a different relationship than the field observations (Figure 7). The CENTURY model seemed to predict an overresponse to precipitation in savanna and dry woodlands. We did not observe this relationship in field measurements from savanna or dry woodlands, but there were only three observations. This information suggests that NPP field estimates are needed in savanna and dry woodlands to determine whether the model's response mimics nature or whether it is an artifact of the CENTURY model. We also observed more subtle differences caused by a lack of response of NPP to temperature in temperature-limited sites. CENTURY model NPP predictions in very cold regions and in coniferous forests were higher than that observed in the field.

The overall comparison between TEM predictions and field data revealed remarkable agreement. In fact, our contrasts of the TEM model with field data failed to identify differences that could potentially lead to model revision or guidance for future field studies. We observed that TEM model predictions of NPP in northeastern broadleaf forests were higher than NPP observed in the field data, and those for grasslands tended to be lower. Given the uncertainties in the field data, the differences reported here were neither large enough nor certain enough to suggest modifications to the model. However, repeating this analysis with additional data for broadleaf forests (for example, from the US Forest Service's Forest Inventory and Analysis measurements) could help to interpret these differences.

Constructive Contrasts vs Formal Tests

In this study, we found that descriptive contrasts were more useful in providing constructive feedback than formal, deductive contrasts (that is, those focused on testing the hypothesis that models do not differ from field data). Statistical measures of goodness-of-fit (for example, *P* values in Tables 4 and 5) were not very useful in describing the potential for model improvement. These tests stopped at simple statements about model performance rather than providing meaningful feedback to the scientific process.

Constructive methods gave us a much better picture of how far the three models were from the field data, and, more importantly, why. In addition, constructive methods overcame one theoretical problem often attributed to model validation. Deductive tests never fail when evaluating models that generate highly variable model predictions, whereas constructive methods focus on outliers when they appear in a model and not in the data. Extreme predictions become a focus of constructive contrasts rather than a statistical smokescreen. This is not to say that deductive tests serve no useful purpose. The scientific community tends to be suspicious of constructive contrasts when conducted by modelers

Table 5. Overall Goodness-of-fit Comparison between Three Productivity Models and the Field-data Surface Fitted to NPP Measurements

VEMAP model	Sample size	Residual error Mean (standard error)	Absolute error Mean (standard error)
Field data	296	–6.8 (11.03)	136.7 (7.64)
BIOME-BGC	3168	45.1 (5.21)	240.7 (3.08)
CENTURY	3168	–41.7 (3.01)	145.5 (1.71)
TEM	3168	–32.6 (2.20)	105.2 (1.30)

Field data results are provided as a reference.

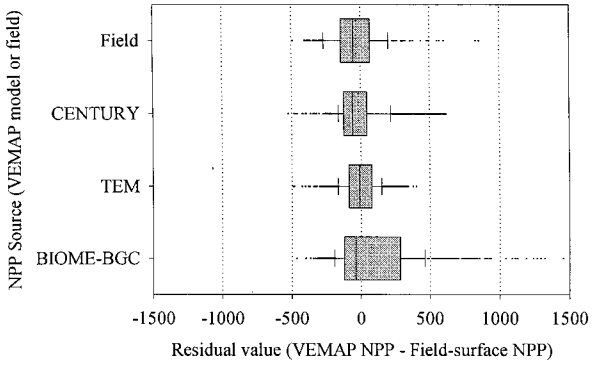

Figure 8. The distribution of residuals (model NPP–field-data surface NPP) is shown for each of the three VEMAP models and for the field data. Box whisker diagrams enclose the 25th and 75th percentiles within the distribution of residuals.

as a means of establishing the validity of their own models. In this situation, a formal test appears more objective because it is less amenable to selective presentation.

When the goal of a model–data contrast is to reveal differences between model predictions and field data, the tools needed shift from those geared toward hypothesis testing to more creative and exploratory pattern-seeking methods (Romesburg 1981). Loehle (1997) gives the example of corroborating a predator–prey model by comparing its qualitative behavior to that observed in the field. He contrasts the creative solution, looking for the "donut shape" in predator–prey trajectories, with differencing of paired trajectories over time. By analogy, we contrasted model responses to environmental gradients with those observed in nature. A model might not show regional patterns of response that are observed in field data, or they might show patterns that are stronger than observed in the natural situations, where numerous other confounding and competing forces act (Kareiva 1990).

Regional-Scale Model–Data Contrasts

The methods used here are particularly well suited to regional-scale comparisons. Two scaling issues that arise when comparing models and data at the regional scale are: (a) unbiased representation of regional patterns (spatial extent) and (b) equivalent spatial support (spatial grain). We achieved unbiased representation by applying principles of sampling design. We defined a region of inference and assigned sample weights to ensure that the influence carried by individual field measurements was proportional to their representation of the specified region. Fortunately, our model predictions were drawn from a regular spatial sample, but model predictions can also be weighted if need be. Other poststratification factors, besides the two climate variables considered here, might be important. For example, Schimel and others (1997) suggested that regional patterns of disturbance create a great deal of spatial variation in NPP not accounted for in regional extrapolations of VEMAP model predictions. A logical next step would be to focus on disturbance history in a future contrast.

We did not specifically address the effects that different areas of spatial support might have on our results. In general, it is best to preserve spatial variation by making local model predictions at the same scale as the field measurements. However, if local NPP model predictions are aggregated to represent somewhat larger areas (for example, the half-degree cells in this case), then the increased spatial support of the estimates should merely reduce the variability of the estimates (Griffith 1988; Jager and others 1990). Our intuition is that functional contrasts should be relatively robust to such scale differences. On the other hand, coarser-scale predictions from nonlinear models that are obtained by

Table 6. Sample Sizes Associated with Each Vegetation Type for the VEMAP Grid and the Field Data

Kuchler vegetation class	Field sample size	VEMAP sample size
Grassland & shrubland	145 (49%)	1175 (37%)
Savanna & dry woodland	3 (1%)	512 (16%)
Subtropical forest	6 (2%)	164 (5%)
Mixed forest	19 (6%)	564 (18%)
Coniferous forest	89 (30%)	387 (12%)
Broadleaf forest	34 (12%)	366 (12%)

Percent composition of the two samples is provided in parentheses.

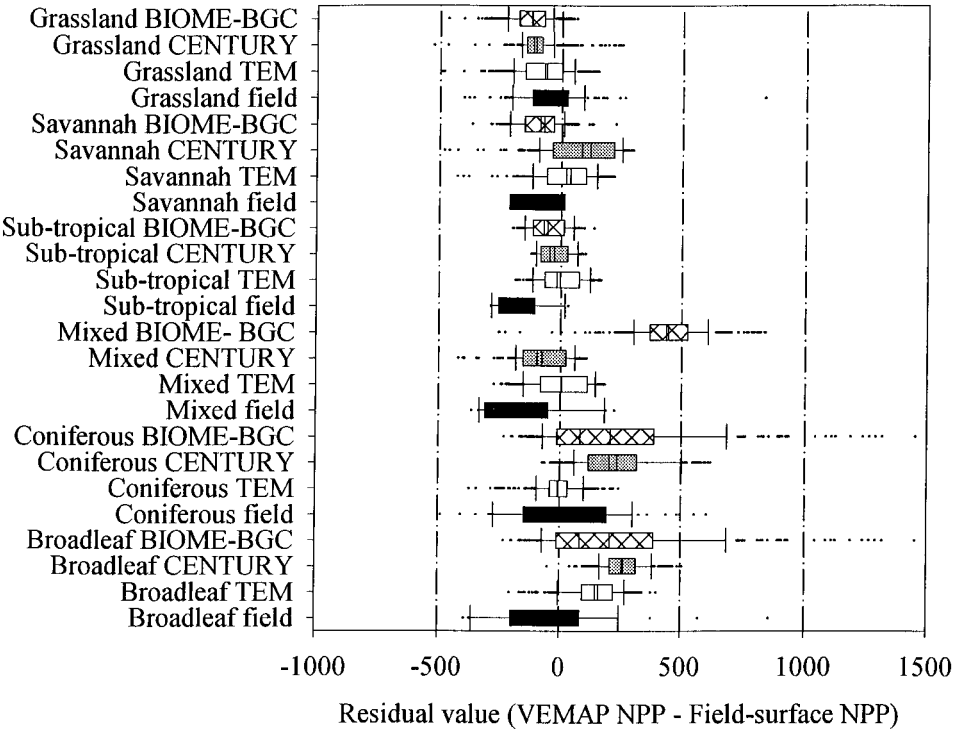

Figure 9. This series of histograms shows how the fit between model and field data varies with vegetation type for the field data and each of the three VEMAP models. Box whisker diagrams enclose the 25th and 75th percentiles within the box. Both the median (solid line) and mean (dotted line) are shown. Agreement is within the bounds of the uncertainty in the field data when the model distribution of residuals (model NPP–field-data surface NPP) falls within the range of the field distribution.

spatially averaging model inputs might be biased (King and others 1991; Rastetter and others 1992; Schimel and others 1997), potentially introducing discrepancies with the field data.

A strong practical advantage of these methods is that they do not require paired or colocated samples and model predictions. Substituting the empirical model adds robustness by removing the geographic context of the comparison and placing it in an abstract climatic space. By focusing on relationships between NPP and climate, we avoided the inevitable problems with inconsistent data sources and locational errors that arise in real-world applications.

ACKNOWLEDGMENTS

Support for this study was provided by the Terrestrial Ecology and Global Change program (NASA reference number TE/97-0026) funded by the US National Aeronautics and Space Administration, Office of Earth Science, Terrestrial Ecology Program under Interagency Agreement number 2013-K063-A1, under Lockheed Martin Energy Research Corporation contract DE-AC05-96OR22464 with the US Department of Energy. We thank the Climate System Modeling Program of the University Corporation for Atmospheric Research and the Ecosystem Dynamics and the Atmosphere Section of the Na-

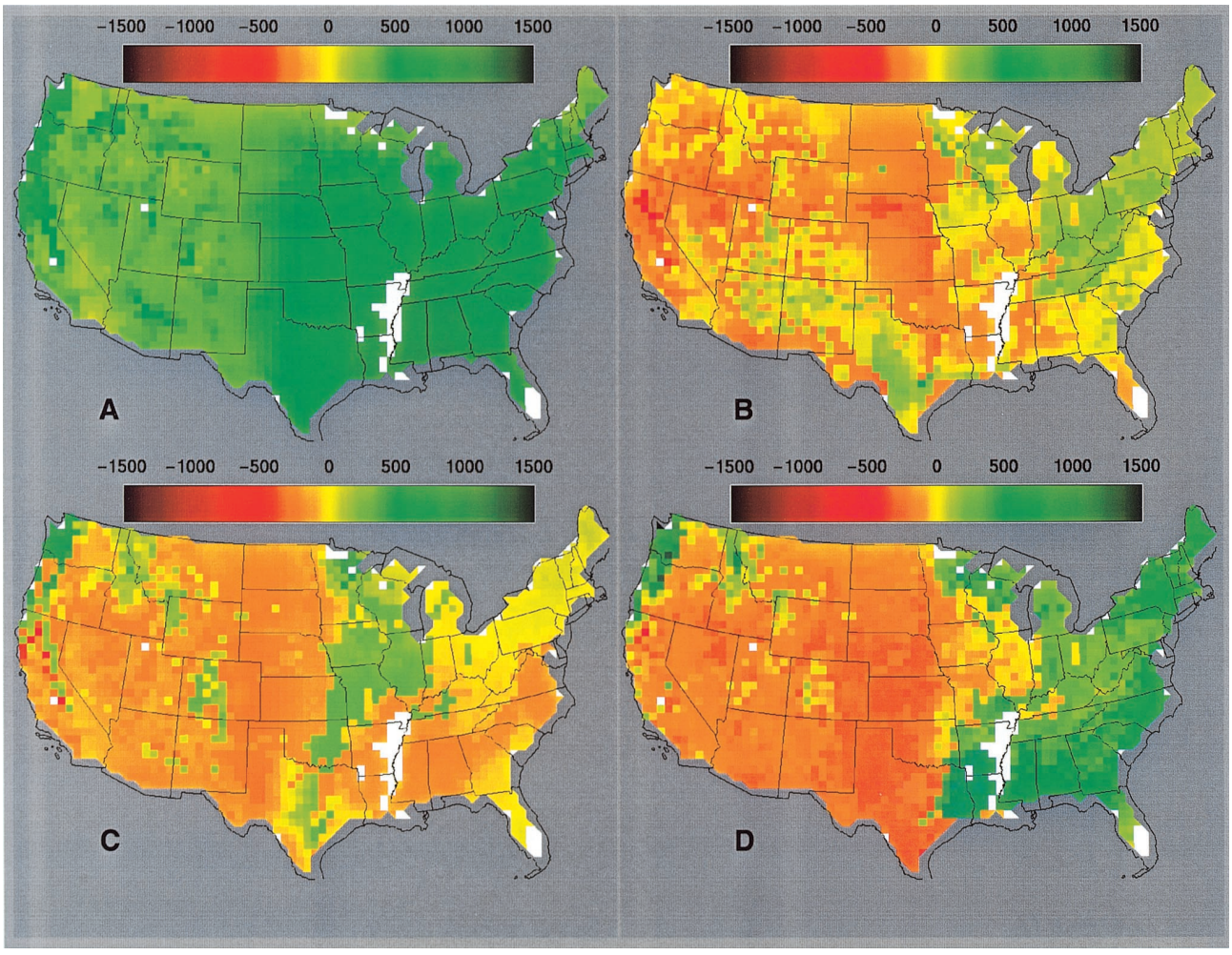

Figure 10. **A** The field data surface represents NPP as predicted by mean annual temperature and total annual precipitation. We compare geographic patterns of fit between each VEMAP model and the field-data surface by mapping residuals (model NPP–field-data surface NPP) in units of g C/m^2/y: **B** the TEM model NPP predictions, **C** the CENTURY model NPP predictions, and **D** the BIOME-BGC model NPP predictions.

tional Center for Atmospheric Research for providing us with the VEMAP Phase I data. Development of the VEMAP database was jointly supported by the NASA Mission to Planet Earth, the Electric Power Research Institute, the USDA Forest Service Southern Region Global Change Research Program, and the NSF-ATM Climate Dynamics Program through the University Corporation for Atmospheric Research's Climate System Modeling Program. We thank Dr. Eric Smith for comments on the statistical methodology. In addition, we are grateful for the insightful and timely reviews by Dr. Chuck Garten and two anonymous reviewers. Dr. Ed Rastetter contributed a great deal to improving the final product. This is ESD publication number 4966.

REFERENCES

Ahn CH, Tateishi R (1994) Development of a global 30-minute grid potential evapotranspiration data set. J Japan Soc Photogram Remote Sens 33:12–21

Atkinson AC (1985) Plots, transformations, and regression. Oxford: Clarendon Press

DeAngelis DL, Gardner RH, Shugart HH (1981) Productivity of forest ecosystems studied during the IBP: the woodlands data set. In: Reichle DE, editor. Dynamics of forest ecosystems. Cambridge: Cambridge University Press. p 567–672

Draper NR, Smith H (1981) Applied regression analysis. New York: Wiley

Esser G, Lieth HFH, Scurlock JMO, Olson RJ (1997) Worldwide estimates and bibliography of Net Primary Productivity derived from pre-1982 publications. ORNL/TM-13485. Oak Ridge, TN: Oak Ridge National Laboratory

Griffith DA (1988) Advanced spatial statistics: special topics in

the exploration of quantitative spatial data series. Boston: Kluwer Academic Publishers

Hall DO, Scurlock JMO (1991) Climate change and productivity of natural grasslands. Ann Bot 67 (Suppl):49–55

Hall FG, Huemmrich KF, Strebel DE, Goetz SJ, Mickeson JE, Woods KD (1992) Biophysical, morphological, canopy optical property, and productivity data from the Superior National Forest. 104568. Greenbelt, MD: Goddard Space Flight Center, National Aeronautics and Space Administration

Hargrove WW, Pickering J (1992) Pseudoreplication: a sine qua non for regional ecology. Landscape Ecol 6(4):251–258

Hunt ER, Running SW (1992) Simulated dry matter yields for aspen and spruce stands in the North American boreal forest. Can J Remote Sens 18:126–133

Jager HI, Overton WS (1993) Explanatory models for ecological response surfaces. In: Goodchild MF, Parks BO, Steyaert LT, editors. Environmental modeling with GIS. New York: Oxford University Press. p 422–431

Jager HK, Sale MJ, Schmoyer RL (1990) Cokriging to assess regional stream quality in the southern Blue Ridge province. Water Resour Res 26:1401–1412

Kareiva P (1990) Population dynamics in spatially complex environments: theory and data. Phil Trans R Soc London B 330:175–190

King AW, Johnson AR, O'Neill RV (1991) Transmutation and functional representation of heterogeneous landscapes. Landscape Ecol 5:239–253

Korn EL, Graubard BI (1995) Examples of differing weighted and unweighted estimates from a sample survey. Am Statis 49:291–295

Kott PS (1991) A model-based look at linear regression with survey data. Am Stat 45:107–112

Kuchler AW (1964) Potential natural vegetation of the conterminous United States, manual to accompany the map. New York: American Geographic Society

Kuchler AW (1975) Potential natural vegetation of the conterminous United States (map 1:3, 168,000). New York: American Geographic Society. p 143

Levins R (1966) The strategy of model building in population biology. Am Sci 54:421–431

Lieth H (1975) Modeling the primary productivity of the world. In: Lieth H, Whittaker RH, editors. Primary productivity of the biosphere. New York: Springer-Verlag

Loehle C (1997) A hypothesis testing framework for evaluating ecosystem model performance. Ecol Modelling 97:153–165

Magee L (1998) Improving survey-weighted least squares regression. J Roy Stat Soc B 60(P1), 115–126

Melillo JM, McGuire AD, Kicklighter DW, Moore B III, Vorosmarty CJ, Schloss AL (1993) Global climate change and terrestrial net primary production. Nature 363:234–240

Overton JM, Young TC, Overton WS (1993) Using "found" data to augment a probability sample: procedure and case study. Environ Monit Assess 26:65–83

Overton WS (1990) A strategy for use of found samples in a rigorous monitoring design. Technical report 119. Corvallis, OR: Department of Statistics, Oregon State University

Parton WJ (1996) The CENTURY model. In: Powlson DS, Smith P, Smith JU, editors. Evaluation of soil organic matter models using existing long-term datasets. Berlin: Springer-Verlag

Parton WJ, Schimel DS, Cole CV, Ojima DS (1987) Analysis of factors controlling soil organic matter levels in Great Plains grasslands. Soil Sci Soc Am J 51:1173–1179

Raich JW, Rastetter EB, Melillo JM, Kicklighter DW, Steudler PA, Peterson BJ, Grace AL, Moore B III, Vorosmarty CJ (1991) Potential net primary productivity in South America: application of a global model. Ecol Appl 1:399–429

Rastetter EB, King AW, Cosby BJ, Hornberger GM, O'Neill RV, Hobbie JE (1992) Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems. Ecol Appl 21:55–70

Romesburg HC (1981) Wildlife science: gaining reliable knowledge. J Wildlife Manage 45:293–313

Running SW, Hunt ER Jr. (1993) Generalization of a forest ecosystem process model for other biomes, BIOME-BGC, and an application for global-scale models. In: Scaling physiological processes: leaf to globe. New York: Academic Press. p 141–158

Runyon J, Waring RH, Goward SN, Welles JM (1994) Environmental limits on net primary production and light-use efficiency across the Oregon transect. Ecol Appl 4:226–237

Sala OE, Parton WJ, Joyce LA, Lauenroth WK (1988) Primary production of the central grassland region of the United States. Ecology 69:40–45

Schimel DS et al. (1997) Continental scale variability in ecosystem processes: models, data, and the role of disturbance. Ecol Monogr 67:251–271

Scurlock JMO, Cramer W, Olson RJ, Parton WJ, Prince SD (1999) Terrestrial NPP: towards a consistent data set for global model evaluation. Ecol Appl 9:913–919

VEMAP Members (1995) Vegetation/ecosystem modeling and analysis project: comparing biogeography and biogeochemistry models in a continental-scale study of terrestrial ecosystem responses to climate change and CO_2 doubling. Global Biogeochem Cycl 9:407–437

Zeide B (1991) Quality as a characteristic of ecological models. Ecol Model 55:161–174