A Beta Regression Model for Improved Solar Radiation Predictions

RANDALL MULLEN, LUCY MARSHALL, AND BRIAN MCGLYNN*

Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, Montana

(Manuscript received 27 January 2012, in final form 10 January 2013)

ABSTRACT

Predicting global solar radiation is an integral part of much environmental modeling. There are several approaches for predicting global solar radiation at a site where no instrumentation exists. One popular approach uses the difference between daily high and low temperature, typically using a nonlinear equation to express the relationship between change in temperature and estimated global solar radiation. Additional variables are usually included in successive steps creating a hierarchy of analysis. The authors propose an alternative beta regression approach to modeling global solar radiation, allowing for the inclusion of multiple environmental predictor variables and strata into one flexible model. The model is applied to several case studies, and results are compared with recently proposed empirical solar radiation models. Beta regression provides a robust, flexible modeling approach for predicting global solar radiation that allows for the addition and removal of independent variables as appropriate and can be interpreted using standard inferential statistics. In addition, the beta regression model provides estimates of uncertainty that can be incorporated into subsequent models and calculations.

1. Introduction

Predictions of solar radiation are a requisite to models of soil moisture (Spokas and Forcella 2006), carbon flux and plant growth (van Dijk et al. 2005), wildlife behavior (Keating et al. 2007), evapotranspiration (Hargreaves and Samani 1982), weed management (Spokas and Forcella 2006), hydrology (Lindsey and Farnsworth 1997), and others. Numerous models have been proposed to predict solar radiation at ungauged locations because of the frequent lack of instrumentation to directly measure it (Thornton and Running 1999). One common approach is to use the difference between the daily maximum and the daily minimum temperature ΔT at a location as a means to predict the fraction of solar radiation that reaches Earth's surface. To date, a wide variety of models have been implemented that predict solar radiation based on observations of ΔT. One of the earliest was proposed by Hargreaves and Samani (1982). Bristow and Campbell (1984) proposed a model where transmissivity is a function of ΔT smoothed across two days. Richardson (1985) proposed a simple model where ΔT is a function of two site-specific empirical parameters and extraterrestrial radiation. Liu and Scott (2001) compare nine models that predict solar radiation, three of which use only ΔT, two that use only precipitation, and four that use both. Thornton and Running (1999) proposed a ΔT method enhanced with precipitation and dewpoint data. Samani et al. (2011) propose a modified version of Allen (1997), a model self-calibrated by season and location. A nonlinear equation is used in each of these to model the relationship between ΔT and solar radiation.

Fodor and Mika (2011) compared a four-parameter "S shaped" function borrowed from soil science with Donatelli and Campbell (1998)'s function for predicting the fraction of solar radiation that hits Earth's surface. This fraction, called fraction of clear day (FCD), is expressed as the percentage of solar radiation that reaches Earth's surface on a clear day. This latter value is referred to as clear-sky transmissivity (CST) and is described in detail, along with FCD, in section 2. Fodor and Mika (2011) correctly noted that earlier models (Bristow and Campbell 1984; Donatelli and Campbell 1998) forced the relationship between ΔT and FCD through the origin; FCD cannot ever be zero (except

* Current affiliation: Nicholas School of the Environment, Duke University, Durham, North Carolina.

Corresponding author address: Randall Mullen, Montana State University, Dept. of Land Resources and Environmental Sciences, P.O. Box 173120, Bozeman, MT 59717-3120.
E-mail: rsmullen@uw.edu

DOI: 10.1175/JAMC-D-12-038.1

perhaps in the polar winters). Fodor and Mika (2011) then proposed a four-parameter sinusoidal curve and found it produces smaller prediction errors when compared with Donatelli and Campbell (1998).

All solar radiation models are limited by the available observations for model fitting. Gueymard and Myers (2009) described three levels of stations that collect solar radiation data: 1) *solar monitoring sites* use inexpensive and automated instrumentation to provide local data quickly for a minimal cost, 2) *conventional long-term measurements* use proven techniques and are generally operated by weather service agencies, and 3) *research sites* are typically developed by atmospheric physicists or climatologists to obtain the highest accuracy possible to detect trends or test theoretical solar radiation models. These research sites have higher levels of redundancy with respect to instrumentation and power supply. Typically, ΔT models are developed and tested on high-quality data collected at research sites. Spokas and Forcella (2006) used data from 16 research sites throughout North America, Sweden, and Australia. Thornton and Running (1999) and Fodor and Mika (2011) used data from the Solar and Meteorological Surface Observation Network (SAMSON) database that included up to 109 stations from around the United States. Liu and Scott (2001) used 39 research sites distributed throughout Australia. Bristow and Campbell (1984) developed their model at three different locations in the northern United States. Using (relatively) independent sites with high-quality data to formulate predictive equations provides a strong basis for model development and assessment. However, data from solar monitoring sites allow for the investigation of spatial characteristics not possible when data collection is limited to research sites. Thus there remains a need for flexible modeling frameworks that can be applied to all sites that collect solar radiation data. This requires a model that is robust when analyzing small datasets or allows for combining previously separate analyses.

In this study, we implement a beta regression model to facilitate prediction of incoming solar radiation at ungauged locations or to fill gaps due to power or equipment failure in existing datasets. The intent of this study is not to develop a widely transferable model with fixed parameters, but rather establish a flexible method that allows researchers to add or remove variables based on local availability and appropriateness. The model also provides valid estimates of uncertainty and relatively unbiased predictions. Confidence intervals and capture rates are reported to emphasize and illustrate estimates of uncertainty. Interpretable parameters are obtainable using beta regression; however, higher-order models and models with explanatory variables that display multicollinearity can lead to erroneous results. The emphasis herein is on predicting global solar radiation. We consider the application of beta regression in the context of solar monitoring networks. As with previous models, the beta regression model we propose does not directly model global solar radiation but rather FCD. Detailed discussion of ΔT models, the deconstruction of global solar radiation, and beta regression follows.

2. A review of ΔT models for solar radiation prediction

Global solar radiation (GSR) can be broken down into three components. Extraterrestrial radiation (ETR) is the amount of solar radiation that hits the outside of the atmosphere. CST is the amount of ETR that will reach Earth's surface on a clear day. FCD is the fraction of CST that hits Earth's surface on any given day. The ΔT models take advantage of this deconstruction and relate the difference of high and low daily temperature to FCD. The suite of current ΔT models (Fodor and Mika 2011; Bristow and Campbell 1984; Donatelli and Campbell 1998) for predicting FCD, and subsequently GSR, can largely be described by the following sequence of analysis:

1) Determine ETR at a given site using geographical location, time of day, and time of year (e.g., Gates 1980).
2) For each day of the year (denoted yearday), estimate CST. This can be predicted empirically using historical data or can be modeled (e.g., using Fourier series) with shorter historical datasets.
3) For each day in a given dataset, divide measured daily GSR by CST to determine FCD.
4) Calculate ΔT for each day in the given dataset. The simple calculation (Hargreaves and Samani 1982) is

$$\Delta T^i = T^i_{\max} - T^i_{\min}, \tag{1}$$

where i = the ith day of the dataset.

A smoothed calculation first proposed by Bristow and Campbell (1984) and used frequently is

$$\Delta T^i = T^i_{\max} - 0.5(T^i_{\min} + T^{i+1}_{\min}). \tag{2}$$

5) Plot FCD versus ΔT and fit a nonlinear curve.
6) For any day at this or any nearby location, if GSR is unknown and ΔT is known, then GSR can be predicted using the fitted value for FCD and the following relationship:

$$\widehat{\mathrm{GSR}} = \mathrm{ETR} \times \mathrm{CST} \times \mathrm{FCD}_{(\mathrm{fitted})}. \tag{3}$$

It is assumed that the procedures in step 1 are well established (Gates 1980). For step 2, if a sufficiently long

dataset exists then CST can be obtained empirically. Thornton and Running (1999) use a moving window that encompasses 7 days (3 before and 3 after) for each yearday to empirically derive CST while also proposing a way to derive CST with no solar radiation data from a site. Fodor and Mika (2011) suggest using a Fourier series to model CST using the maximum values for each yearday in a dataset. For step 3, the ΔT value [Eq. (2)] suggested by Bristow and Campbell (1984) has been used by several subsequent studies (Fodor and Mika 2011); however, Thornton and Running (1999) found that using nonsmoothed values [Eq. (1)] led to less error.

Step 5, modeling the relationship between FCD and ΔT, is the most contested aspect of the above algorithm, and comparisons of methods are typically done using root-mean-square error (RMSE), mean signed deviance (MSD), or mean absolute error (MAE) (Donatelli and Campbell 1998; Fodor and Mika 2011). The traditional justification for fitting ΔT models is that the model is useful for prediction purposes. Little effort is spent interpreting the fitted parameters in part because interpreting the coefficients would not yield better predictions of FCD. Additionally, interpreting parameters of these models is difficult or impossible. Fodor and Mika (2011) make no attempt to interpret parameters using a simplified soil water retention curve. The emphasis here will be on prediction as well. Typically, predicted FCD values are inputted at Eq. (3) for steps 5 and 6 with no regard for estimates of uncertainty in the predicted values. Resulting GSR predictions are then reported without prediction intervals. Attempts to spatially interpolate parameters and/or final GSR predictions (step 6) are done as if known measured values are being presented (Fodor and Mika 2011; Thornton and Running 1999; Thornton et al. 2000).

3. A review of beta regression

Beta regression provides a framework for modeling continuous variables constrained in the standard unit interval (0, 1) (Ferrari and Cribari-Neto 2004). A necessary assumption is that the response variable is beta distributed with a mean that can be related to a set of regressors with estimable coefficients and a link function. The beta distribution is a continuous probability distribution defined on the interval between 0 and 1 and its probability density function is traditionally expressed as

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1, \qquad (4)$$

with shape parameters p and $q > 0$, and where $\Gamma()$ is the gamma function. Ferrari and Cribari-Neto (2004) reparameterized the beta distribution by setting $\mu = p/(p+q)$ and $\varphi = p + q$. This yields

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma[(1-\mu)\phi]} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \qquad (5)$$

where $0 < \mu < 1$ and $\varphi > 0$. As in the original parameterization, $\Gamma()$ is the gamma function. The expected value of y is μ, or $E(y) = \mu$. The parameter φ is known as the precision parameter since for fixed μ, larger φ gives smaller variance for the distribution. A beta-distributed variable can be denoted as $y \sim \beta(\mu, \phi)$. In matrix notation, beta regression is then represented as

$$g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \eta_i, \qquad (6)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a $k \times 1$ vector of unknown regression parameters, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}}$ is a vector of k regressors, or independent variables, $g(\mu)$ is a link function (in this case the logit link), and η_i is a linear predictor. Since the variance of y is a function of μ, the regression model is naturally heteroscedastic with

$$\mathrm{Var}(y_i) = \frac{u_i(1-u_i)}{1+\phi}. \qquad (7)$$

Beta regression provides an effective framework for modeling bounded environmental variables, such as FCD, when standard regression techniques are likely inappropriate. Assumptions of normality are usually incorrect because truncation of the response value makes even an approximate normal distribution unlikely. Almost by definition they display a large amount of heteroscedasticity with more variation around the midpoint and less close to 0 or 1. Like most proportion data, FCD distributions tend to be asymmetric, which leads to issues with confidence intervals and hypothesis testing. Beta regression addresses all of these issues (Ferrari and Cribari-Neto 2004). Further, functions to perform beta regression are now readily available in popular software programs (Cribari-Neto and Zeileis 2010). The flexibility of beta regression is easily demonstrated by modeling predictions of FCD using a set of climate variables that are regularly collected at weather stations as regressors. Unlike previously proposed methods, beta regression is not limited to one independent variable and all standard regression inferences can be made when fitting FCD versus ΔT, or any combination of climate variables available to the researcher.

In this study, a new flexible ΔT model using beta regression is compared with the Fodor and Mika (2011) model using data from a network of solar monitoring sites throughout North and South Dakota. Points of analysis include

1) comparison of standard indicators of fit such as RMSE, MAE, and MSD;
2) comparison of ease of fitting and determination of robustness for both methods (i.e., do respective algorithms converge; are model parameters easily identifiable?);
3) comparison of reliability of prediction intervals for FCD and demonstration of how to estimate prediction error of GSR using the variance of predicted FCD;
4) discussion of interpretation of the model parameters;
5) determination of whether modifications to the standard design of the beta regression model are necessary for improved model predictions, including data stratification.

4. Materials and methods

a. Data and site description

The study area is composed of North and South Dakota in the north-central United States. These states have distinct continental climate with very cold winters and hot semihumid summers, although the western part of North Dakota is considered semiarid. The highest recorded temperature in either state is 49°C and the coldest is −51°C. The average annual precipitation ranges from 35 to 75 cm throughout the study area.

Data from 99 Automated Weather Data Network (AWDN) (Fig. 1) operated by the High Plains Regional Climate Center were inspected for quality and quantity (length of data series and amount of missing data). Standard weather variables collected at the AWDN sites include (but are not limited to) daily high temperature, daily low temperature, relative humidity, and precipitation. Seven sites had a substantial amount of missing data and were dropped from the analysis. Chosen for comparison were 92 sites (Table 1) inside of North and South Dakota and two in Montana very near the border of North Dakota. All data denoted as bad, missing, or imputed (You et al. 2008) were removed. Three sites were chosen to demonstrate a variety of attributes concerning the data, as well as analysis of results. These are the Redfield, Takini, and Brookings sites in South Dakota. The total area that can be reasonably inferred as coverage is approximately 382 843 km^2, yielding a density of 2.5×10^{-4} sites per kilometer squared. This coverage provides an opportunity to assess the model performance over a comparably dense monitoring network. Fodor and Mika (2011) inspected 109 sites spread across the contiguous United States and Hawaii (~8 311 200 km^2; 1.3×10^{-5} sites per kilometer squared). Bechini et al. (2000) inspected 29 stations in northern Italy (~100 408 km^2; 2.8×10^{-4} sites per kilometer squared). The density of coverage for this study is thus almost 20 times denser than the dataset used by previous studies in North America (Fodor and Mika 2011).

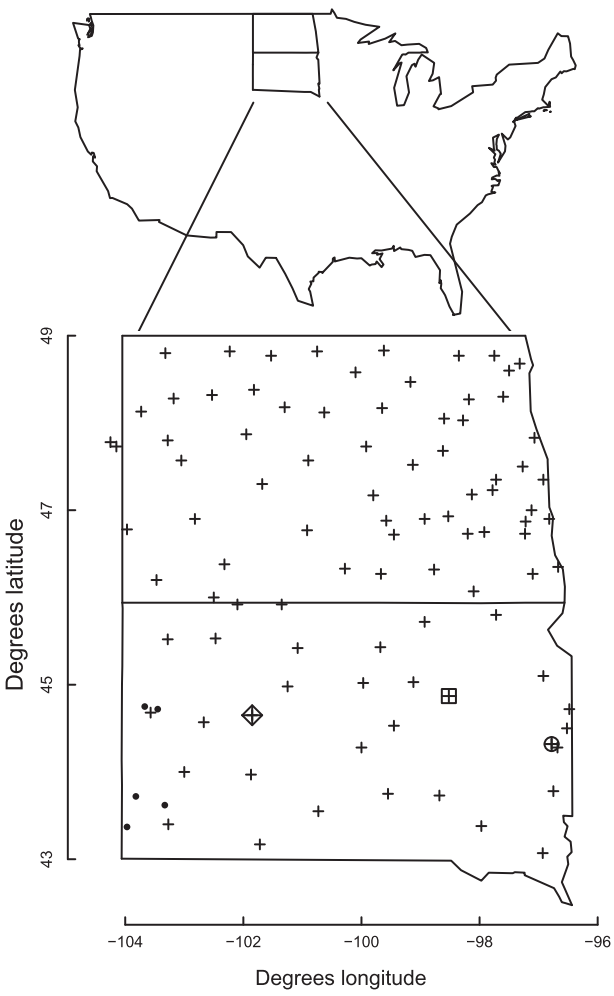

FIG. 1. The Montana, North Dakota, and South Dakota sites of the AWDN network. Three sites mentioned in the text, Redfield, Takini, and Brookings, are denoted with a square, a diamond, and a circle, respectively. The bulleted sites are the sites that did not have enough data to create valid CST Fourier series. Top figure shows the location of North and South Dakota in the United States.

b. Decomposing global solar radiation and model construction

Global solar radiation can be deconstructed into three elements: ETR, CST, and FCD. Doing so provides a simple approach for addressing seasonal cycles, effects

of elevation, and atmospheric attenuation independently. The historical context of this deconstruction is discussed in section 2, with details of how each component was calculated below. In this study, calculations for ETR and CST are essentially unchanged from past studies. The beta regression model we are proposing is intended to improve upon past methods for predicting FCD.

ETR was calculated using methods described in Gates (1980). In this method, day of year, latitude, distance to sun, and declination (derived using latitude) are determined for each site and for each day on which data are available. The calculation of ETR accounts for seasonal changes in the solar radiation. The solar constant is considered to be 1366 W m^{-2}.

The annual course of CST is typically cyclical with relatively small amplitude and asymmetrical peaks (Fodor and Mika 2011). Daily sky transmissivity (ST) values were determined for each data point (1 day^{-1}). Therefore, if a dataset is 10 years long, there will be 10 ST values for each yearday, or each unique date. Maximum ST values were extracted for each yearday using a 7-day moving window (Thornton and Running 1999) in case a reliable maximum cannot be captured in a relatively short dataset. These maximums were then fitted with the second-order Fourier series shown as Eq. (8) (Fodor and Mika 2011):

$$y = a + b\cos(x) + c\sin(x) + d\cos(2x) + e\sin(2x), \quad (8)$$

where $x = 2\pi(\text{yearday}/366)$ and a, b, c, d, and e are fitted constants.

This was done for each site individually, resulting in each site having an associated set of values for the Fourier series parameters. CST was modeled yearly, regardless of whether FCD was analyzed in seasonal strata or not. The effects of individual site characteristics on GSR are accounted for in the calculation of CST.

Where GSR data are observed, FCD can be easily calculated by rearranging Eq. (3):

$$\text{FCD} = \text{GSR} \times \text{ETR}^{-1} \times \text{CST}^{-1}, \quad (9)$$

Where GSR is not observed, FCD can be predicted using temperature and other climate variables. For this step, the proposed beta regression model is implemented. For comparison, other methods are briefly presented here.

Bristow and Campbell (1984) suggested

$$\text{FCD} = a[1 - \exp(-b\Delta T^c], \quad (10)$$

where ΔT is calculated as shown in Eq. (2) and a, b, and c are fitted constants.

Fodor and Mika (2011) correctly point out that this and other previous models are inappropriately forced through the origin (Bristow and Campbell 1984; Donatelli and Marletto 1994; Donatelli and Campbell 1998), such that as ΔT approaches zero FCD approaches zero. They then propose a strictly monotonic equation that is not forced through the origin:

$$\text{FCD} = 1 - \frac{1 - a}{[1 + (b\Delta T)^c]^d}, \quad (11)$$

where a, b, c, and d are parameters that are empirically fitted for each site.

This model was found to produce smaller errors than a previous study by Donatelli and Campbell (1998). Error was further reduced when separate analyses were performed by season (winter, spring, summer, and fall) and precipitation (wet versus dry). Therefore, eight unique models were required for each site. In this study, FCD is predicted using beta regression. For model fitting, observed values of FCD are calculated using data from sites and days where GSR is measured [Eq. (9)]. Once fitted, the resulting regression equation can be used to predict FCD at locations and on days where explanatory variables are obtained but no measurement of GSR exists. This technique provides locally relevant parameter estimates such that a regression equation that has been fitted using nearby data can be used to predict FCD at a location that does not measure GSR.

There are multiple studies that review the implementation of beta regression models (Cribari-Neto and Zeileis 2010; Ferrari and Cribari-Neto 2004; Ospina et al. 2006; Rocha and Simas 2011; Simas et al. 2010; Smithson and Verkuilen 2006) and related model diagnostics (Chien 2010; Espinheira et al. 2008a,b). Interested readers are encouraged to consult these for further information on beta regression implementation. Here we construct an example of how the beta regression model may be applied to predictions of daily GSR using data from one station. These results will be compared with the Fodor and Mika model. Data collected at the Takini site from 2005 to 2009 are used to fit a beta regression model incorporating multiple climatic predictors The resulting parameters are used to make predictions for the 2010 Takini site data. We then extend the model to a solar monitoring network composed of 92 stations. GSR is calculated using FCD predictions from both the beta regression model and the Fodor and Mika (2011) model. We assess the performance of the beta regression model and its ease of use, and make recommendations regarding how it can be implemented. Because of the flexibility of the beta regression model, a binary variable for wet days was created as well as a continuous variable

TABLE 1. A complete list of data available for analysis. Sites shown in boldface were removed from the analysis because of insufficient data to fit a Fourier series for CST. Numbers shown in boldface indicate a season and precipitation grouping that had insufficient data (N/A) for testing of model fit.

				Fit		Dry days		Test				Fit		Wet days		Test		
Site	Total	Flagged	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall
Antelope Range	1461	907	260	231	240	254	66	59	59	78	**3**	33	35	19	10	33	32	12
Aurora	3287	1854	610	501	535	534	33	53	57	8	41	202	158	79	**5**	21	12	**0**
Baker	6574	233	1392	1076	1089	1342	88	52	59	86	65	446	450	192	**0**	40	33	**5**
Beach	6574	241	1423	1044	1121	1360	86	53	63	86	82	471	413	180	**3**	39	29	**5**
Beresford	8401	834	1661	1329	1476	1612	80	66	65	83	121	555	469	270	9	26	26	7
Berthold	3652	176	768	589	605	727	84	59	50	87	37	225	223	92	**1**	32	42	**4**
Bison	1826	1406	348	290	305	328	78	56	38	32	**14**	64	62	36	10	25	15	**0**
Bottineau	6574	25	1345	1034	1113	1332	87	62	52	87	73	438	435	187	**2**	30	39	**4**
Bowbells	3652	217	767	608	565	740	89	58	53	87	45	218	199	79	**0**	34	39	**4**
Bowdle	1826	508	310	267	249	256	76	60	58	77	**15**	87	85	30	9	30	14	10
Bowman	6574	60	1377	1021	1171	1372	83	54	64	87	91	476	372	170	**0**	38	28	**4**
Britton	4748	195	1019	757	796	943	76	60	61	84	53	313	297	146	12	30	31	7
Brookings	10 227	2572	1805	1563	1652	1720	34	51	56	9	118	631	547	277	**1**	19	12	**0**
Brorson	5844	10	1275	976	1049	1209	87	61	60	87	51	372	324	132	**2**	31	32	**4**
Cando	6209	174	1308	991	1019	1257	89	49	61	86	52	396	449	182	**0**	43	31	**5**
Caputa	6940	862	1360	1041	1208	1363	87	54	59	83	79	440	336	152	**2**	38	32	7
Carrington	8401	785	1766	1368	1391	1722	89	59	67	81	94	532	605	239	**0**	33	24	10
Cavalier	6574	242	1405	1061	1044	1305	86	56	58	84	72	418	479	227	**3**	36	34	7
Cottonwood	4748	28	1016	729	790	945	84	58	76	80	61	351	277	142	**5**	34	15	10
Crary	4383	363	927	689	691	850	88	55	64	86	55	293	308	130	**1**	37	28	**5**
Crosby	3287	131	682	521	545	635	89	55	54	87	36	206	190	91	**0**	36	38	**4**
Dazey	6574	59	1345	1073	1074	1343	82	60	64	84	91	448	479	195	7	32	28	7
Dell Rapids	3287	1087	655	487	540	578	85	56	54	72	47	229	176	122	**2**	36	37	18
Dickinson	7670	393	1616	1183	1349	1639	86	58	65	87	66	514	458	176	**3**	34	26	**4**
Dunn	730	149	87	70	66	78	87	56	65	87	**4**	**11**	26	**13**	**2**	36	26	**4**
Eagle Butte	2191	716	313	365	313	330	25	34	66	78	**15**	77	65	32	**0**	17	16	10
Edgeley	6574	303	1414	1060	1071	1335	85	51	67	84	91	453	449	193	**4**	35	25	7
Edgemont	**1461**	**1460**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Ekre	2191	301	418	328	327	378	89	56	58	82	34	131	132	69	**0**	36	34	8
Fargo	7670	174	1505	1229	1289	1529	89	63	63	86	220	550	536	281	**0**	29	29	**5**
Faulkton	1826	488	318	271	259	247	86	59	57	83	19	74	74	39	**3**	30	26	7
Fingal	3652	331	765	570	573	707	89	54	61	84	39	232	253	96	**0**	38	30	7
Forestriver	7305	235	1556	1224	1146	1466	88	62	65	84	88	438	555	247	**1**	30	27	7
Galesburg	5844	182	1211	926	956	1163	83	64	61	86	82	382	413	200	**6**	27	30	**5**
Gettysburg	10 227	2297	1837	1719	1786	1767	80	38	80	78	116	520	429	215	**4**	17	11	12
Grafton	1826	165	263	262	249	303	86	58	61	85	18	92	103	60	**3**	34	30	**6**
Grandforks	7670	173	1586	1239	1229	1540	80	57	59	81	109	490	587	268	9	35	33	10
Hamill	**1096**	**1096**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Harvey	6209	229	1315	1002	1029	1251	86	57	69	85	61	405	413	173	**3**	35	22	**6**

TABLE 1. (*Continued*)

Site	Total	Flagged	Winter	Fit		Dry days		Test		Fit		Wet days		Test			
				Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall
Hazen	6574	208	1429	1069	1130	1373	84	58	61	87	73	432	408	169	**5**	34	31	**4**
Hettinger	8401	728	1827	1374	1477	1788	89	56	63	85	96	564	504	198	**0**	36	28	**6**
Highmore	2191	648	326	303	350	348	84	59	53	84	**14**	125	96	55	**0**	33	38	**6**
Hillsboro	6574	404	1358	1067	1063	1319	89	58	58	85	78	378	479	224	**0**	34	33	**6**
Hofflund	4383	463	862	696	747	840	85	62	62	87	38	269	236	91	**4**	30	30	**4**
Inkster	730	191	83	73	65	79	85	57	61	72	**8**	**12**	24	**12**	**1**	34	26	**6**
Jamestown	8401	741	1805	1403	1396	1765	78	62	66	83	100	528	569	213	7	30	22	8
Jewel Cave	**3287**	**3282**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Karlsruhe	3652	235	771	595	596	721	88	60	59	86	41	231	218	96	**1**	32	32	**5**
Lake Cochrane	1826	333	336	261	266	288	76	64	44	7	20	95	101	66	8	28	12	**0**
Langdon	8401	39	1591	1205	1218	1568	86	58	55	83	91	525	610	267	**3**	34	37	8
Lemmon	2191	751	411	344	342	334	85	48	64	7	**15**	103	73	27	**4**	12	23	**0**
Leola	1826	825	227	250	236	187	71	59	52	79	**15**	85	68	35	9	31	26	7
Leonard	3287	320	679	549	518	616	89	63	59	86	42	185	212	97	**0**	29	33	**5**
Linton	6574	32	1419	1084	1123	1359	82	63	67	84	70	446	431	179	7	29	25	7
Lisbon	1826	136	340	253	253	308	77	59	63	83	22	98	115	56	12	33	29	8
Mandan	4383	190	944	705	736	889	84	59	60	87	49	285	241	112	**5**	33	32	**4**
Marion	1461	122	254	181	190	231	83	54	68	85	18	74	86	42	**6**	38	23	**6**
Martin	2191	666	388	290	297	320	81	58	72	78	25	158	68	51	7	34	18	11
Mayville	5844	171	1221	953	963	1152	85	58	64	85	71	357	415	206	**4**	32	28	**6**
McHenry	5844	277	1224	942	944	1159	84	57	65	81	76	391	416	190	**5**	35	27	10
McIntosh	2191	385	431	342	368	412	78	69	65	78	**15**	113	88	41	7	22	26	10
Michigan	2922	183	584	435	444	542	88	55	61	82	49	193	193	91	**1**	37	31	7
Minot	8401	264	1549	1232	1333	1583	87	63	60	86	111	492	453	203	**2**	29	31	**5**
Mohall	6574	246	1397	1090	1120	1361	88	52	57	87	72	422	419	173	**1**	27	35	**4**
Mott1n	3287	260	691	532	565	655	85	64	59	88	32	204	155	73	**4**	28	29	**3**
Newell AC Flume	**1096**	**1096**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Nisland	**4748**	**1135**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Oacoma	9862	2550	1984	1542	1750	1819	10	49	56	10	131	550	405	214	**1**	19	10	**0**
Oakes	7670	171	1609	1255	1287	1579	89	57	59	83	118	512	520	221	**0**	35	32	8
Oak Lake	2191	590	404	308	329	326	81	61	54	83	21	136	105	51	8	31	37	7
Oral	3287	2291	503	474	508	455	9	58	54	32	28	157	89	68	**0**	**0**	**0**	**0**
Orman Dam	**1096**	**1096**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Parkston	1826	936	225	220	234	168	80	62	36	**3**	23	102	76	28	**0**	30	21	**0**
Pierre	7305	1165	1529	1185	1201	1353	13	60	58	14	93	498	350	156	**3**	7	7	**0**
Pillsbury	3652	307	769	587	584	692	83	67	58	85	42	238	244	103	**6**	25	34	**6**
Plaza	3287	129	686	503	550	646	84	62	58	88	37	219	184	82	**5**	30	34	**3**
Prosper	7670	247	1601	1256	1261	1544	89	62	60	86	119	474	551	245	**0**	30	30	**5**
Redfield	10 227	1237	2063	1690	1651	1928	81	72	77	87	96	528	512	216	8	20	14	**3**
Robinson	6574	208	1424	1072	1094	1341	83	60	66	86	68	448	445	200	**6**	32	26	**5**
Rolla	5479	184	1146	871	907	1117	89	53	57	87	49	346	377	153	**0**	39	34	**4**

TABLE 1. (*Continued*)

Site	Total	Flagged	Winter	Fit		Dry days		Test				Fit		Wet days		Test		
				Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall
Ross	3287	152	686	498	535	647	88	59	56	89	32	198	194	76	**0**	33	36	**2**
Rugby	2922	288	599	458	478	557	87	55	63	85	33	185	162	63	**2**	37	29	**6**
Sidney	5844	83	1272	994	1024	1193	84	34	61	88	51	350	323	130	**3**	33	31	**3**
South Shore	8401	1693	1690	1322	1389	1527	16	50	57	10	99	613	507	212	7	23	14	**0**
Streeter	8401	716	1807	1391	1457	1739	81	59	66	87	89	542	550	237	8	33	26	**4**
St. Thomas	6209	201	1304	1008	1020	1236	86	56	61	85	77	376	447	213	**3**	35	31	**6**
Takini	2191	1931	174	299	308	128	10	46	53	**3**	**9**	79	47	**9**	**0**	18	7	**0**
Tappen	3287	177	691	512	522	640	83	57	67	84	32	207	206	87	**6**	35	25	**6**
Timber Lake	2191	1084	225	326	205	142	8	42	38	63	**12**	92	74	20	**0**	24	35	11
Turtle Lake	6574	43	1378	1079	1128	1370	84	59	69	86	81	428	424	169	**5**	33	22	**5**
Union Center	2191	557	351	288	349	404	76	65	72	79	**15**	120	100	50	**4**	27	18	9
Vale	**730**	**729**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
Wahpeton	3652	372	781	554	540	682	87	62	57	83	28	243	243	117	**2**	30	35	8
Watford City	6574	22	1418	1083	1140	1380	83	62	60	86	67	444	420	162	**6**	29	32	**5**
White Lake	1461	410	238	179	162	150	80	61	54	81	**14**	70	50	28	9	31	23	8
White River	1826	474	320	254	269	266	80	62	73	78	22	102	67	41	9	29	18	12
Williston	4748	8	1029	788	826	990	87	59	58	86	41	305	275	101	**2**	33	34	**5**
Wishek	3652	197	768	575	577	719	83	56	68	86	44	249	230	100	**6**	36	24	**5**
Wyndmere	7670	249	1605	1317	1254	1541	79	60	59	83	140	485	550	267	10	32	33	8

that is simply precipitation in millimeters. Model selection for the beta regression model was done using Akaike information criterion (AIC). Beta regression was implemented in R (R Development Core Team 2009) using the betareg package (Cribari-Neto and Zeileis 2010).

c. Prediction intervals for GSR

Since each predicted value of FCD is a beta distributed value, $y_i \sim \beta(u_i, \phi_i)$, then its distribution can be described with μ and φ. The 0.025 percentile can be considered the lower bound, and the 0.975 percentile can be considered the upper bound. These parameters, μ and φ, each have an associated uncertainty that is not incorporated into the uncertainty interval of FCD. The failure to account for this uncertainty is what distinguishes this estimate from a true prediction interval; however, it can be used similarly. The lower and upper bounds for the FCD uncertainty interval can be used to predict the upper and lower bounds for GSR.

True prediction intervals can be estimated. One could perform a simulation using predicted μ and φ for the new data, or a Bayesian approach could be used. Because of the high number of models run for this study, neither of these methods was used. However, they are an appropriate approach when analyzing one dataset using a single model.

d. Model comparisons

RMSE, MAE, and MSD (an indicator of bias), were used to compare each of the Fodor and Mika models—one for each season and precipitation (wet vs dry) combination—to a beta regression model using the same subsets of data. Several studies have shown that performing separate analyses for each subset reduced error and bias (Allen 1997; Fodor and Mika 2011; Samani et al. 2011). Ease of fit was determined by comparing the number of times computational efforts to fit each model failed to converge. This could manifest itself by not producing parameter estimates, or producing estimates that are essentially zero or infinite. For models that failed to converge, 3000 attempts were made fully encompassing all possible values based on Fodor and Mika (2011) and the analysis described in this paper.

To determine if subsetting was necessary for the beta regression model, all sites were analyzed with yearday converted into sine and cosine components. This eliminated the need for separate analyses for each season. Precipitation was entered as a continuous variable eliminating the need for separate analyses for wet and dry days. In this way, an entire dataset can be evaluated in one model. Total RMSE from the stratified models was compared to the RMSE for the combined model to determine if loss of information occurred.

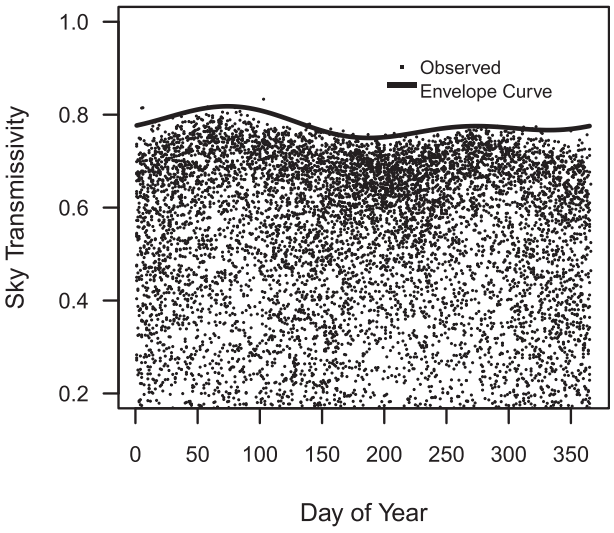

FIG. 2. Transmissivity plotted against day of year (yearday) for all available years. The Fourier series fitted envelope curves through the maximums for each yearday and is considered the fitted CST.

To test spatial interpolations of the fitted models, each site was analyzed using CST as well as the beta model fitted from the nearest site. The rate at which the observed value was captured by a 95% uncertainty interval was compared to capture rates obtained for that site using the site-specific CST and fitted model.

5. Results and discussion

CST was fitted for all stations [Eq. (8)]. As an example, an envelope curve for the Brookings weather station (Fig. 2) had the following fitted parameters; $a = 0.7789$, $b = 0.0130$, $c = 0.0193$, $d = -0.0157$, and $e = 0.0067$.

a. Fitting the Fodor and Mika model

The data were subset as recommended (Fodor and Mika 2011). For each season, wet and dry days were split into two groups. Each of the eight resulting groups was analyzed. For one of the eight models, wet winter days, at the Redfield site, the Fodor and Mika failed to converge. The total number of winter days with precipitation available for analysis was 96; not an uncommonly small sample size when considering sample sizes from the AWDN network (Fig. 3). Traditionally, it is thought that a sinusoidal curve best represents the relationship between change in temperature and FCD. Most analyses herein support this belief; however, this relationship is not ubiquitous. Close inspection of the wet winter days subset for the Redfield site lack this sinusoidal

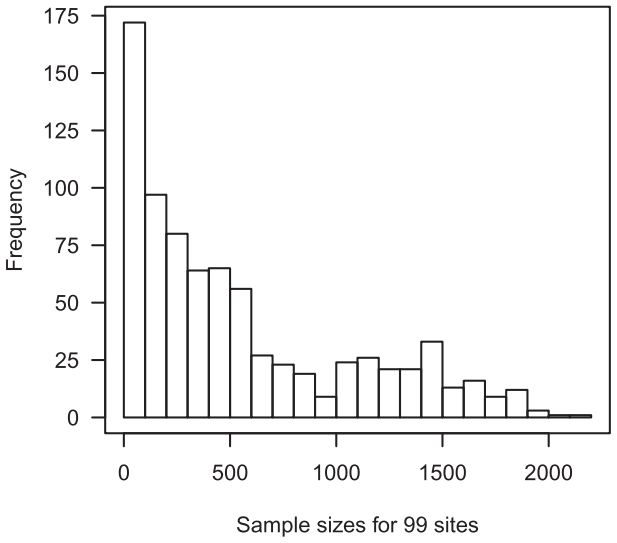

FIG. 3. A histogram of the sample sizes for the 99 sites. The strata are season and precipitation. Note the high frequency of relatively low sample sizes. This causes problems in fitting models that are limited only to one stratum at a time.

relationship (Fig. 4). This could be due to the sample size, a different relationship between these two variables at this site during the winter season, or a combination of both. Regardless, forcing a sinusoidal curve through the points shown in Fig. 4 leads to poor parameter identifiability. The entire dataset (92 sites, 4 seasons, and 2 strata for wet and dry days) was analyzed using the Fodor and Mika model. Of the 736 possible models, 236 (32%) failed to converge when using standard nonlinear regression techniques (Nelder and Mead 1965; Shanno 1970) implemented in R (Nash 1990; R Development Core Team 2009).

Fitting the beta model was not problematic. The beta regression model was far more robust to this nonidentifiability issue than the Fodor and Mika (2011) model. Additionally, theoretical principles are available that yield estimates of uncertainty surrounding predicted response values (prediction intervals). Estimates of uncertainty for nonlinear regression do exist but often rely on asymptotic estimates of variance for parameters (Goh and Pooi 1997).

b. Fitting the beta regression model at the Takini site

Following standard procedures for beta regression (Cribari-Neto and Zeileis 2010; Ferrari and Cribari-Neto 2004; Smithson and Verkuilen 2006), the 2005–09 dataset for dry spring days at the Takini station was analyzed. The resulting model parameters were used to construct predictions for the 2010 Takini dry spring days dataset. An initial inspection of the explanatory variables (Fig. 5) suggests there is notable correlation between relative humidity and both ΔT ($r = -0.70$) and adjusted ΔT ($r = -0.72$). This multicollinearity is a concern only if inferences regarding the estimates of coefficients in the final fitted model are desired. For prediction purposes, multicollinearity is of little concern.

To fit the sinusoidal relationship between ΔT and FCD, a squared and cubic ΔT term were added to the beta regression model. This is a standard approach for fitting nonlinear relationships in a linear model. Inspection of the correlation matrix (Fig. 5) suggests that FCD and subsequently solar radiation might also display

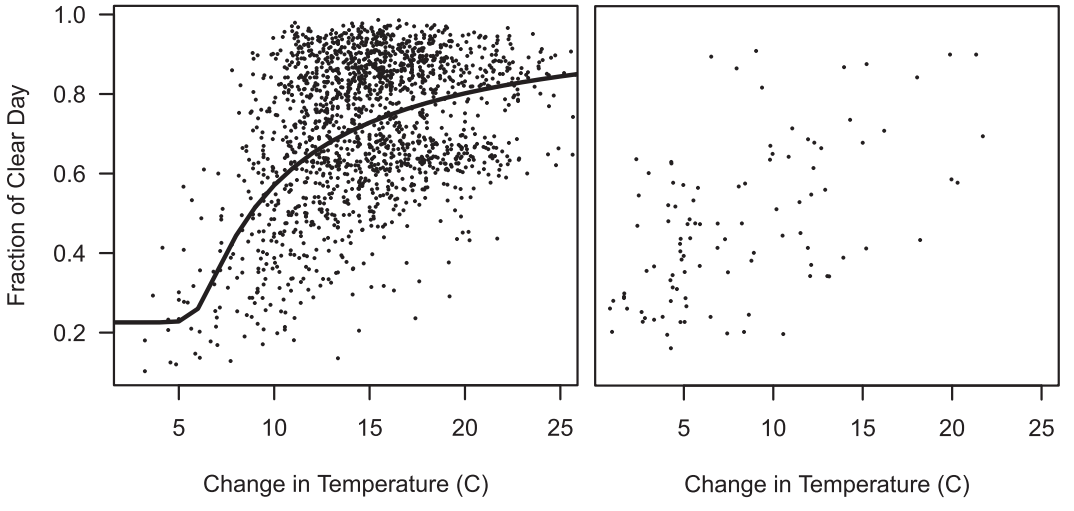

FIG. 4. (left) Data from dry summer days at Redfield. For this dataset, the sinusoidal curve is shown fitted to the data. (right) Data from wet winter days at the same site. Note the lack of sinusoidal structure to the data. Attempts to fit these data with a four-parameter sinusoidal curve led to a variety of possibilities. Nonidentifiability of model parameters was an issue.

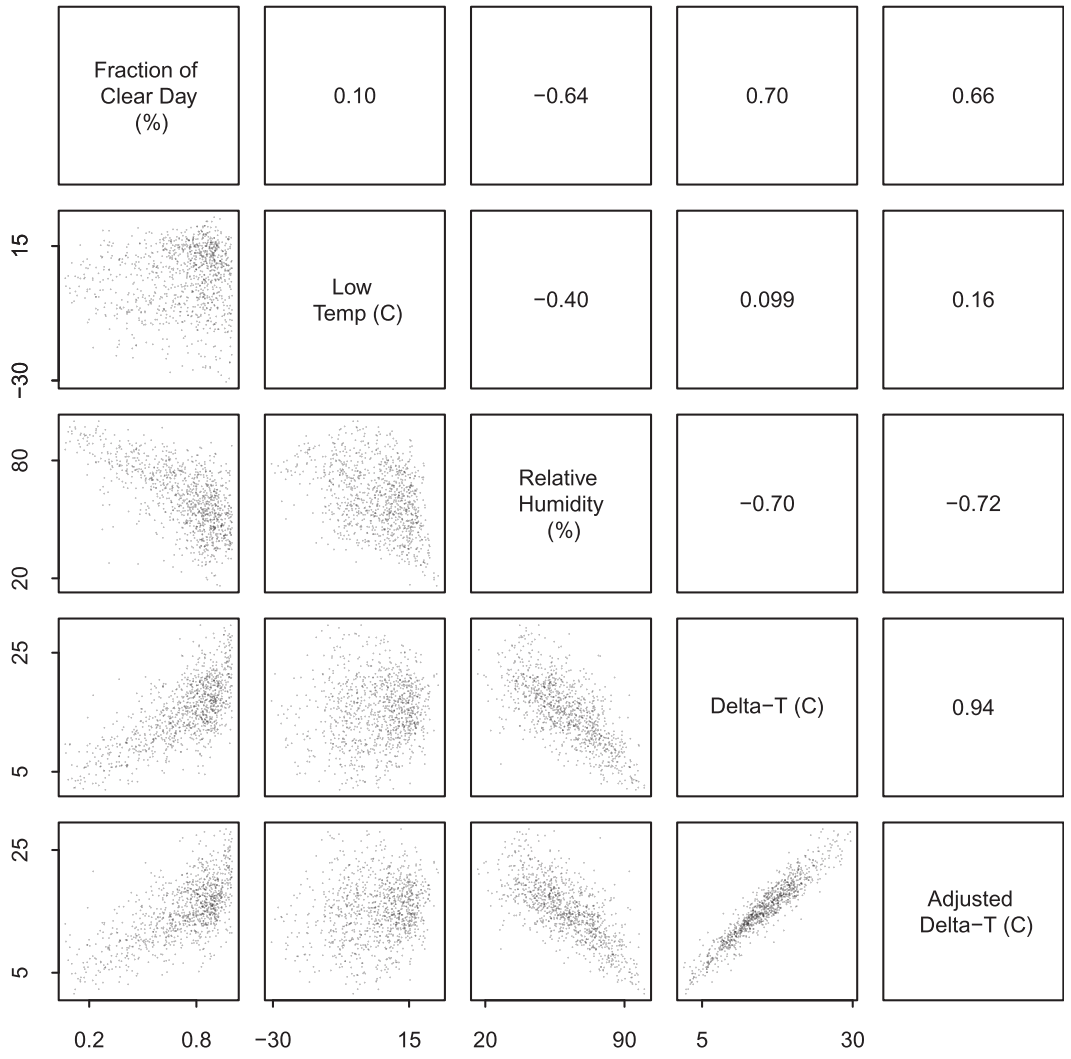

FIG. 5. A simple correlation matrix showing how FCD is correlated with the independent variables and how the independent variables are correlated with each other. Numbers in the panels are the correlation value ρ.

a nonlinear response to low temperature and relative humidity, therefore squared terms were added for each of those variables. The initial covariates in the beta regression model were ΔT, ΔT^2, ΔT^3, relative humidity (average of the day), relative humidity squared, daily low temperature, and daily low temperature squared. Two- and three-way interaction terms were allowed between ΔT, relative humidity, and low temperature. AIC values were calculated for each possible model that maintained the squared and cubic ΔT terms. The three models with the lowest AIC value were (from more to less complex) 1) the full model with all variables (AIC = −506.74); 2) the full model with the one three-way interaction term removed and the two-way interaction between ΔT and low temperature removed (AIC = −506.05); and 3) the model with all of the single covariates and only one interaction term, relative humidity and ΔT (AIC = −506.14). In addition, there were two other models that had AIC values that were within 3 of the best model (ΔAIC < 3). For the purpose of interpreting the estimates of the coefficients, model selection techniques can be used to determine the best model (Burnham and Anderson 2002), but for the purposes of prediction, any of these models may be assumed to work reasonably well. The precision parameter φ for the full model was 11.5 (SE = 0.9343). As an example of calculating an uncertainty interval, 22 April 2010, had a low temperature of 5.25°C, a ΔT of 12.00°C, and an average relative humidity of 72.76% at the Takini site. The observed FCD was 0.4104 and the predicted FCD is 0.6497. Predicted CST was 0.813, and ETR was 34.718. The observed GSR value was 11.589 MJ m^{-2} day^{-1}

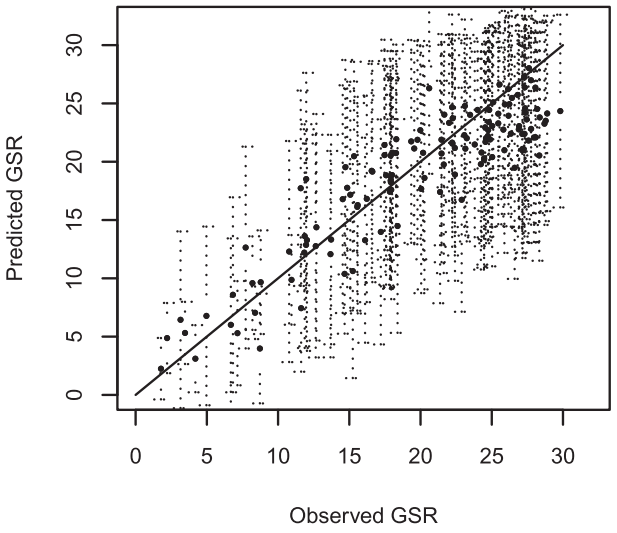

FIG. 6. Predicted GSR values plotted against observed GSR values with 95% prediction intervals shown. Data used were from dry spring days in Takini.

and the predicted GSR is 18.347 MJ m^{-2} day^{-1}. The uncertainty interval has lower and upper confidence bounds of 10.36 MJ m^{-2} day^{-1} and 18.34 MJ m^{-2} day^{-1} respectively, which capture the observed solar radiation value of 11.589 MJ m^{-2} day^{-1}.

This uncertainty could then be incorporated into all subsequent models that use estimated solar radiation as in input. In the previous example, CST is considered without error; however, it is a predicted rather than measured. We incorporated the uncertainty in CST and found a negligible change in the final estimate for $\widehat{GSR}$ and ultimately omitted it from the final analysis.

For this subset, the 95% uncertainty intervals for the Takini 2010 test dataset are shown in Fig. 6. These intervals captured the real value 100% of the time. This is not entirely surprising given the sample size of 46. However, for some purposes, a smaller prediction interval may be required. If smaller prediction intervals are desired, 90% intervals can be calculated by calculating the appropriate percentile from the resulting distribution.

c. *Capture rates for the beta regression model*

The full beta regression model was used to analyze the dry strata for 92 sites across the four seasons to determine what proportion of the observations were captured by the 95% prediction limits (referred to as the rate of capture). Subsets with less than 15 days available for fitting the model, or less than 7 days for testing the model were left out. There were 4 station–season combinations that were omitted for this reason. The average rate of capture of the true value was 93.05%, with a high of 100% and a low of 50%. In this latter case, there were only eight usable days from the dry strata, fall season, 2010 dataset (site = Aurora). Clearly, when dealing with networks of solar monitoring sites, there will be cases such as these that require individual attention. The average capture rates for winter, spring, summer, and fall were 97.81%, 95.86%, 94.50%, and 93.96%, respectively. To assess overall model fit, observed GSR was plotted against predicted GSR and the correlation was calculated (Fig. 7). This was done for all usable sites and subsets of data. The average correlations of observed GSR and predicted GSR on dry days for winter, spring, summer, and fall were 0.88, 0.78, 0.83, and 0.92, respectively. There were more station–season combinations that did not meet the minimal criteria for testing when inspecting wet days ($n = 142$). The overall average rate of capture of the true value for wet days was 89.89%, with a high of 100% and a low of 9.10%. In the latter case, there were 11 usable days in the 2010 test

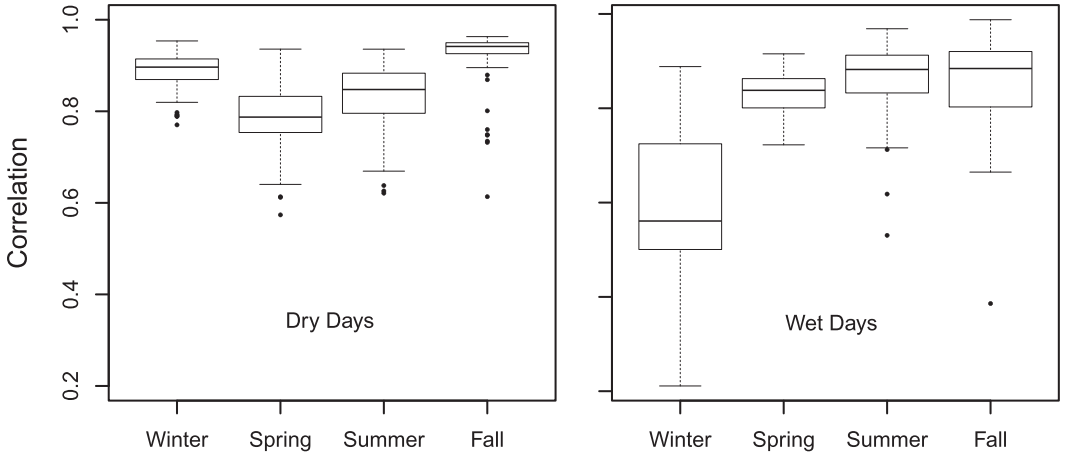

FIG. 7. Box plots showing the overall distributions of correlations between predicted GSR and observed GSR broken down into seasons and (left) dry and (right) wet days.

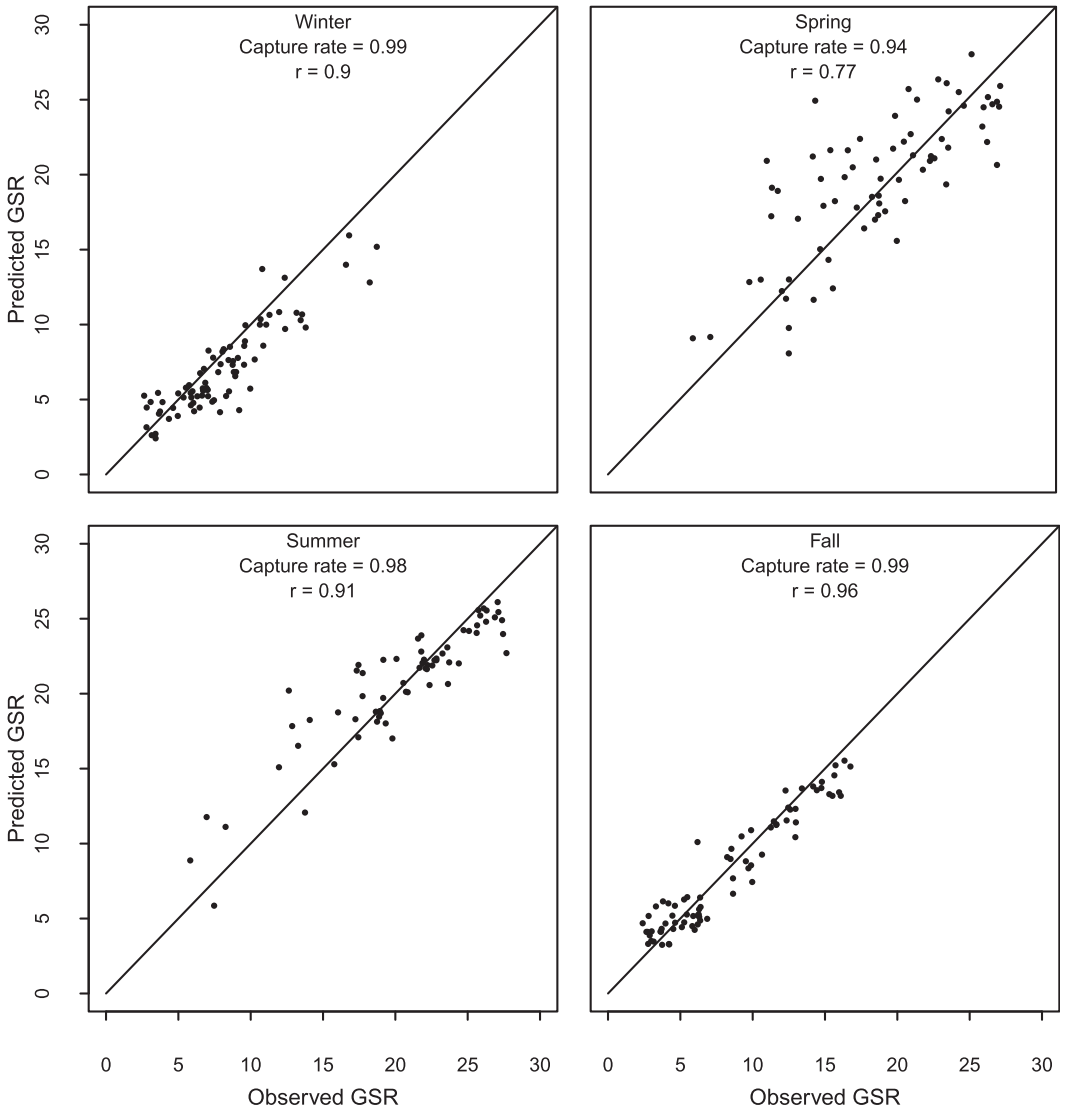

FIG. 8. Predicted GSR plotted against observed GSR for each of the four seasons using data from dry days. The tendency to underestimate days of high GSR is prevalent throughout all 92 sites, although in general, this is a bigger problem in the summer and spring and less of a problem in winter. Capture rates are the rates at which the 95% prediction interval captured the observed value.

dataset. The average capture rates for winter, spring, summer, and fall were 82.88%, 92.53%, 93.30%, and 77.30%, respectively. The correlations of observed GSR and predicted GSR on wet days for each season were 0.58, 0.84, 0.86, and 0.83, respectively. The beta regression model tended to underestimate high values of solar radiation (Fig. 8) and overestimate low values. Overall, this is a smaller problem in winter than in the other seasons and is possibly indicative of a missing variable in the model or a bias in instrumentation.

Note that the parameters a, b, c, and d in Fodor and Mika (2011) have very little interpretable value. A particularly high value of a does not tell researchers anything about the relationship between FCD and ΔT. In contrast, all inferential properties of generalized linear models apply to the beta regression model, as long as all standard regression diagnostic criteria are addressed. Standard methods of model selection can be applied to the beta regression approach and model inferences can be made.

d. Model comparison

Solar radiation predictions were made for all subsets of data that were successfully fitted using the Fodor and

TABLE 2. Comparisons of the Fodor and Mica model (F&M) and the beta regression model. Where N/A is shown, the Fodor and Mica model was unable to be fitted. In all cases the RMSE and MAE were lower for the beta regression model. In six cases, the bias (MSD) was lower for the F&M model but note the units are all in megajoules per meter squared per day and that the increase in bias is very small. This table uses data from the Redfield site.

		RMSE (MJ m^{-2} day^{-1})		MAE (MJ m^{-2} day^{-1})		MSD (MJ m^{-2} day^{-1})	
Season	Precipitation	Beta	F&M	Beta	F&M	Beta	F&M
Winter	Wet	24.791	N/A	6.025	N/A	−0.069	N/A
	Dry	121.160	128.242	6.840	7.664	0.010	0.037
Spring	Wet	125.006	134.704	28.831	33.478	−0.267	0.049
	Dry	207.640	222.446	24.427	28.035	−0.056	−0.015
Summer	Wet	108.674	122.772	22.538	28.765	−0.181	0.058
	Dry	179.974	196.199	18.701	22.225	0.090	0.042
Fall	Wet	44.805	48.143	8.766	10.121	−0.331	−0.036
	Dry	108.659	115.353	5.886	6.633	−0.011	−0.003

Mika (2011) model and compared to predictions estimated using the beta regression model (Table 2) for the Takini site. In each case, CST was derived using a Fourier series (Fodor and Mika 2011). For each dataset, the ΔT values were smoothed using Eq. (2). However, as was shown previously, using Eq. (1) led to better results (Thornton and Running 1999). Therefore, all models were run again using only the change in temperature for the day of interest. This yielded lower errors for all models and has the additional advantage of being less susceptible to erroneous values in the event of missing data (e.g., if day $i + 1$ is missing, then calculation for day i is not jeopardized). The RMSE, MAE, and MSD shown (Tables 2 and 3) are based on the residuals for actual versus predicted GSR. Similar results can be shown for actual FCD versus predicted FCD, but since the intent of these models is to ultimately predict solar radiation, results for GSR were compared. In all cases the RMSE and the MAE for the beta regression models were smaller than the Fodor and Mika model by an average of 10.28 MJ m^{-2} day^{-1}, with the lowest decrease being 3.34 MJ m^{-2} day^{-1}, and the largest decrease being 16.22 MJ m^{-2} day^{-1}. The MAE decreased for the beta regression model by an average of 3.00 MJ m^{-2} day^{-1}, with the lowest decrease being 0.75 MJ m^{-2} day^{-1}, and the largest decrease being 6.23 MJ m^{-2} day^{-1}. The mean signed deviance was larger for the beta regression model in 5 of the 7 cases. This increase in bias averaged 0.88 MJ m^{-2} day^{-1}, a full order of magnitude less than the decrease in RMSE. Therefore, the relatively small increase in bias is negated by the substantial decreases in RMSE and MAE in all 7 cases.

Each of the 92 sites was analyzed for each season and precipitation strata to determine if this pattern was consistent throughout the study area. The beta regression model outperformed the Fodor and Mika model with reduced RMSE and MAE (Table 3) for virtually every strata and every usable site. Overall, the RMSE was reduced an average of 17% and the MAE by 24%. The MSD was generally higher in the beta regression model but in every case by less than 0.25 MJ m^{-2} day^{-1}. This slight increase in bias should not be a problem for most analyses.

TABLE 3. Comparisons of the F&M model and the beta regression model. In all cases the RMSE and MAE were lower for the beta regression model. In five cases, the bias was lower for the F&M model but note the units are all in megajoules per meter squared per day and that the increase in bias is very small. This table uses all data from 92 sites.

		RMSE (MJ m^{-2} day^{-1})		MAE (MJ m^{-2} day^{-1})		MSD (MJ m^{-2} day^{-1})	
Season	Precipitation	Beta	F&M	Beta	F&M	Beta	F&M
Winter	Wet	26.34	31.66	4.72	6.82	0.094	0.012
	Dry	89.08	95.22	4.7	5.37	0.126	0.076
Spring	Wet	96.00	111.2	17.97	24.1	−0.196	0.044
	Dry	145.38	164.36	15.32	19.58	0.022	−0.052
Summer	Wet	86.63	109.3	12.87	20.49	−0.023	0.137
	Dry	110.84	124.57	9.36	11.82	0.06	0.074
Fall	Wet	34.38	43.33	4.19	6.66	−0.077	−0.027
	Dry	78.85	83.82	3.85	4.34	0.045	0.005

e. Combining subsets for the beta regression model

When inspecting data output from networks of solar monitoring sites, it is not unusual to have low sample sizes for numerous subsets of data (Fig. 3). This problem can be alleviated by combining groups. A single beta regression model was used to analyze the Redfield data to determine if seasonal (spring, summer, etc.) and climate (wet versus dry) grouping is necessary. The yearday variable was transformed to radians (as it is circular data) and the sine and cosine components were entered into the model as covariates. Precipitation was left in the model as a continuous variable. The resulting RMSE was 19.735, which is lower than the RMSE from each of the individual models run on separate strata (19.989). This indicates that indeed one model per site can outperform eight separate models for the same site. The beta regression approach allows for the introduction of numerous continuous variables and is the reason this reduction in subsetting without a loss of information is possible.

f. Interpolating between stations

There are advantages to using data from networks of solar monitoring sites despite less accurate solar radiation measurements. For instance, if site density is sufficient, Thiessen polygons (Brassel and Reif 1979) will suffice for spatial interpolation. An analysis was performed using the fitted CST Fourier series and the fitted beta regression model coefficients from the nearest site to test if Thiessen polygons were appropriate for spatial interpolation. All available data were used (up through 2010). Predictions intervals were calculated as previously described and capture rates were recorded. This was done for each site, for each season, and for dry and wet days. The overall mean capture rate was 92.89%. The average capture rate for dry days for winter, spring, summer, and fall were 94.14%, 93.11%, 92.07%, and 92.21%, respectively. The maximum rates were 98.81%, 98.84%, 98.03%, and 99.10% and the minimums were 80.54%, 80.80%, 71.04%, and 47.15%. For wet days, the overall mean capture rate was 81.27%. The average capture rate for dry days for winter, spring, summer, and fall were 69.83%, 87.46%, 89.57%, and 77.05%, respectively. The maximum rates were 96.44%, 97.61%, 99.46%, and 100% and the minimums were 69.81%, 87.41%, 89.53%, and 76.94%. These capture rates indicate that Thiessen polygons are sufficient for spatial interpolation of beta regression parameters within networks of solar monitoring sites.

6. Conclusions

We applied a beta regression model to predict global solar radiation and compared results to recently proposed empirical solar radiation (ΔT) models. The beta regression method resulted in a lower RMSE and MAE than recently proposed models (Fodor and Mika 2011) that have outperformed historical models (Bristow and Campbell 1984; Donatelli and Marletto 1994; Donatelli and Campbell 1998). Beta regression can be easily implemented in free software (R Development Core Team 2009) using the betareg package (Cribari-Neto and Zeileis 2010). This allows for a more robust and simpler model fitting method than previously proposed nonlinear methods. The parameters obtained using beta regressions are easily interpreted, if all diagnostic criteria (Chien 2010) are addressed. For example, certain regions, climate types, or strata may show common tendencies toward models with or without certain predictors (relative humidity, low temperature, etc.). Lower and upper bounds for estimates of FCD can be used to predict upper and lower bounds for GSR. This is helpful not only as a measurement of uncertainty for GSR, but also for subsequent models that incorporate GSR. This uncertainty is reflected in the prediction intervals provided, which can be larger than desired for some applications. While estimates of uncertainty appear large, we contend that previous methods of estimating GSR also possess large amounts of uncertainty, but the uncertainty was never estimated (Fodor and Mika 2011) and is consequently unavailable for inspection or use in subsequent models.

The beta regression method is flexible: it can be expanded if additional meteorological variables are available at a specific location, or it can be reduced if some variables are shown to be insignificant or unavailable. Because beta regression allows for a multiple regression analysis, variables such as time and precipitation that have been previously analyzed by subsetting the data can be incorporated into one model, which allows site–season combinations that previously had too few data points to analyze to be analyzed. The distribution parameters that accompany the predictions of a beta regression model can be used to estimate uncertainty in the final prediction of global solar radiation. To determine how well these models could be used at locations where no GSR data exist, each site was analyzed using the nearest neighbor. Predictions made using Thiessen polygons and beta regression parameters have slightly lower capture rates (mean of 93.16%) of the observed value using a 95% prediction interval. We have outlined a flexible modeling approach that allows for the addition and removal of independent variables as appropriate, accompanying measures of uncertainty, and ease of operation.

Acknowledgments. The authors thank the Natural Resources Conservation Service for funding this project and the High Plains Regional Climate Center for collecting these data and making them available for purchase.

REFERENCES

Allen, R. G., 1997: Self-calibrating method for estimating solar radiation from air temperature. *J. Hydrol. Eng.,* **2,** 56–67.

Bechini, L., G. Ducco, M. Donatelli, and A. Stein, 2000: Modelling, interpolation and stochastic simulation in space and time of global solar radiation. *Agric. Ecosyst. Environ.,* **81,** 29–42.

Brassel, K. E., and D. Reif, 1979: A procedure to generate Thiessen polygons. *Geogr. Anal.,* **11,** 289–303.

Bristow, K. L., and G. S. Campbell, 1984: On the relationship between incoming solar radiation and daily maximum and minimum temperature. *Agric. For. Meteor.,* **31,** 159–166.

Burnham, K. P., and D. R. Anderson, 2002: *Introduction to Model Selection and Multimodel Inference.* Springer, 488 pp.

Chien, L.-C., 2010: Diagnostic plots in beta-regression models. *J. Appl. Stat.,* **38,** 1607–1622.

Cribari-Neto, F., and A. Zeileis, 2010: Beta regression in R. *J. Stat. Software,* **34,** 1–24.

Donatelli, M., and V. Marletto, 1994: Estimating surface solar radiation by means of air temperature. *Proc. Third Annual ESA Congress,* Abano, Italy, European Society for Agronomy, 352–353.

——, and G. S. Campbell, 1998: A simple model to estimate global solar radiation. *Proc. Fifth European Society of Agronomy Congress,* Vol. 2, Nitra, Slovak Republic, 133–134.

Espinheira, P. L., S. L. P. Ferrari, and F. Cribari-Neto, 2008a: Influence diagnostics in beta regression. *Comput. Stat. Data Anal.,* **52,** 4417–4431.

——, ——, and ——, 2008b: On beta regression residuals. *J. Appl. Stat.,* **35,** 407–419.

Ferrari, S., and F. Cribari-Neto, 2004: Beta regression for modelling rates and proportions. *J. Appl. Stat.,* **31,** 799–815.

Fodor, N., and J. Mika, 2011: Using analogies from soil science for estimating solar radiation. *Agric. For. Meteor.,* **151,** 78–86.

Gates, D. M., 1980: *Biophysical Ecology.* Springer-Verlag, 611 pp.

Goh, K. L., and A. H. Pooi, 1997: Adjustment of prediction intervals in nonlinear regression. *Biom. J.,* **39,** 719–731.

Gueymard, C. A., and D. R. Myers, 2009: Evaluation of conventional and high-performance routine solar radiation measurements for improved solar resource, climatological trends, and radiative modeling. *Sol. Energy,* **83,** 171–185.

Hargreaves, G. H., and Z. A. Samani, 1982: Estimating potential evapotranspiration. *J. Irrig. Drain. Div.,* **108,** 225–230.

Keating, K. A., P. J. P. Gogan, J. M. Vore, and L. R. Irby, 2007: A simple solar radiation index for wildlife habitat studies. *J. Wildl. Manage.,* **71,** 1344–1348.

Lindsey, S. D, and R. K. Farnsworth, 1997: Sources of solar radiation estimates and their effect on daily potential evaporation for use in streamflow modeling. *J. Hydrol.,* **201,** 348–366.

Liu, D. L., and B. J. Scott, 2001: Estimation of solar radiation in Australia from rainfall and temperature observations. *Agric. For. Meteor.,* **106,** 41–59.

Nash, J. C., 1990: *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* Hilger, 278 pp.

Nelder, J. A., and R. Mead, 1965: A simplex method for function minimization. *Comput. J.,* **7,** 308–313.

Ospina, R., F. Cribari-Neto, and K. L. P. Vasconcellos, 2006: Improved point and interval estimation for a beta regression model. *Comput. Stat. Data Anal.,* **51,** 960–981.

R Development Core Team, cited 2009: The R project for statistical computing. [Available online at http://www.R-project.org.]

Richardson, C. W., 1985: Weather simulation for crop management models. *Trans. Amer. Soc. Agric. Eng.,* **28,** 1602–1606.

Rocha, A., and A. Simas, 2011: Influence diagnostics in a general class of beta regression models. *Test,* **20,** 95–119.

Samani, Z. A., G. H. Hargreaves, V. D. Tran, and A. S. Bawazir, 2011: Estimating solar radiation from temperature with spatial and temporal calibration. *J. Irrig. Drain. Eng.,* **137,** 692–696.

Shanno, D. F., 1970: Conditioning of quasi-Newton methods for function minimization. *Math. Comput.,* **24,** 647–656.

Simas, A. B., W. Barreto-Souza, and A. V. Rocha, 2010: Improved estimators for a general class of beta regression models. *Comput. Stat. Data Anal.,* **54,** 348–366.

Smithson, M., and J. Verkuilen, 2006: A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods,* **11,** 54–71.

Spokas, K., and F. Forcella, 2006: Estimating hourly incoming solar radiation from limited meteorological data. *Weed Sci.,* **54,** 182–189.

Thornton, P. E., and S. W. Running, 1999: An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. *Agric. For. Meteor.,* **93,** 211–228.

——, H. Hasenauer, and M. A. White, 2000: Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: An application over complex terrain in Austria. *Agric. For. Meteor.,* **104,** 255–271.

van Dijk, A. I. J. M., A. J. Dolman, and E.-D. Schulze, 2005: Radiation, temperature, and leaf area explain ecosystem carbon fluxes in boreal and temperate European forests. *Global Biogeochem. Cycles,* **19,** GB2029, doi:10.1029/2004GB002417.

You, J., K. G. Hubbard, and S. Goddard, 2008: Comparison of methods for spatially estimating station temperatures in a quality control system. *Int. J. Climatol.,* **28,** 777–787.