Combining Snow Water Equivalent Data from Multiple Sources to Estimate Spatio-Temporal Trends and Compare Measurement Systems

Mary Kathryn Cowles[1], Dale L. Zimmerman, Aaron Christ, and David L. McGinnis

[1]Mary Kathryn Cowles is Assistant Professor; Dale L. Zimmerman is Professor; and Aaron Christ is Graduate Student, Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242. David L. McGinnis is Assistant Professor, Department of Geography, University of Iowa, Iowa City, Iowa 52242.

Abstract

Owing to the importance of snowfall to water supplies in the western United States, government agencies regularly collect data on snow water equivalent (the amount of water in snow) over this region. Several different measurement systems, of possibly different levels of accuracy and reliability, are in operation: snow courses, snow telemetry, aerial markers, and airborne gamma radiation. Data are available at more than 2000 distinct sites, dating back a variable number of years (in a few cases to 1910). Historically, these data have been used primarily to generate flood forecasts and short-term (intra-annual) predictions of streamflow and water supply. However, they also have potential for addressing the possible effects of long-term climate change on snowpack accumulations and seasonal water supplies. We present a Bayesian spatio-temporal analysis of the combined snow water equivalent (SWE) data from all four systems that allows for systematic differences in accuracy and reliability. The primary objectives of our analysis are (1) to estimate the long-term temporal trend in SWE over the western U.S. and characterize how this trend varies spatially, with quantifiable estimates of variability, and (2) to investigate whether there are systematic differences in the accuracy and reliability of the four measurement systems. We find substantial evidence of a decreasing temporal trend in SWE in the Pacific Northwest and northern Rockies, but no evidence of a trend in the intermountain region and southern Rockies. Our analysis also indicates that some of the systems differ significantly with respect to their accuracy and reliability.

Key Words: Bayesian analysis; Climate change; Markov Chain Monte Carlo; Spatial statistics; Spatio-temporal process.

1 INTRODUCTION

Environmental data often come from multiple, disparate sources or consist of subsets collected under markedly different conditions in space or time. It is often preferable (and sometimes necessary) to base statistical inference upon data combined from all these sources or subsets. Statistical methodology should account for possible systematic differences in the distributions of the response variable(s) across sources or subsets. Because most environmental data vary spatially and temporally, the methodology also should model spatial and temporal variation as flexibly as possible. This may require a model with a relatively large number of parameters. A Bayesian analysis, in which inference on model parameters is based on their posterior distributions given the data, offers the most hope for dealing with models sufficiently complex to be realistic.

The specific combined-data problem considered in this article is estimation of the amount of water in the western United States snowpack. Approximately 75% of annual discharge in western rivers begins as snowpack (Palmer, 1988). Thus, wise water management decisions in this region depend crucially on good estimation of the water supply in the snowpack throughout the year. Data that provide timely information are collected regularly by several U.S. government agencies. The specific quantity measured is snow water equivalent (SWE), which is the amount of water in the snow. This article considers SWE data from the eleven westernmost of the 48 conterminous United States, to which we refer as the "western U.S." Over this region, several different systems, constituting distinct sampling networks, are used to measure SWE either directly or indirectly. We consider the following four systems.

1. *Snow courses.* Snow courses are relatively straight transects, approximately 1000 meters in length, that are located in wind-protected open areas. A person takes a measurement, at each of five to ten sites along the transect, by driving a hollow tube through the snow to the ground below. The snow inside the tube is weighed and the weight is converted directly to SWE using a special scale. The average SWE over all sites sampled along the transect is reported as the SWE for the transect. Snow course SWE measurements have been taken in the western U.S. since about 1910. Currently, most measurements are taken at more-or-less monthly intervals (a few are taken biweekly) from January to June on more than 1700 snow courses. The Natural Resources Conser-

vation Service (NRCS) of the U.S. Department of Agriculture oversees the sampling program.

2. *Snow telemetry (SNOTEL).* The SNOTEL system consists of a network of electronic devices set up in wind-protected open areas. Each device consists of several antifreeze-filled butyl rubber or stainless steel pillows exposed to the sky. When snow falls it accumulates on the pillows and increases their internal pressure by an amount proportional to the snow's weight. Adjacent instrumentation converts the pressure change to SWE and transmits it daily to a base station. Operated by NRCS, the SNOTEL network was introduced in the mid-1960's and has gradually grown in size. Currently there are about 550 sites in the western U.S.

3. *Aerial markers.* This system consists of depth markers read visually from low-flying aircraft. Many of the markers are at locations deemed too hazardous or costly to take measurements on the ground. The observed snow depth is converted to SWE using measurements of snow density taken at nearby snow course sites. The NRCS began collecting data of this type on a more-or-less monthly basis at a small number of sites in 1936. Gradually, more sites have been added to the network; currently, measurements are taken at about 60 sites in the western U.S.

4. *Airborne gamma radiation (AGR).* Along designated flight lines, an instrument in a low-flying aircraft measures the level of natural gamma radiation emanating from the soil and passing through the snow. The attenuation in gamma radiation from the level measured before the snow season, due to the water mass in the intervening snow, is converted electronically to SWE. The flight lines included in this study are about 300 m wide and range in length from 4.1 miles to 20.1 miles. A single measurement is an areal average over the length and breadth of the flight line. Flight lines are flown a few times between late January and mid-April. The system is operated by the National Operational Hydrologic Remote Sensing Center (NOHRSC) of the U.S. National Weather Service. Since its inception in the western U.S. in 1985, 422 distinct flight lines have been flown at least once.

Historically, SWE data produced by these four systems have been used primarily to generate flood forecasts and short-term (intra-annual) predictions of streamflow and water supply. These predictions nowadays can be made in near real-time (Hartman, Rost, and Anderson, 1995). However, with the effects of climate change (i.e., global warming) on snowpack accumulations and seasonal water supplies becoming a growing concern (Gleick, 1987; Lettenmaier and Sheer, 1991), the data's potential for also addressing questions of long-term climate change has begun to be realized; see, for example, Serreze et al. (1999).

Recent developments in the statistical analysis of spatially and temporally correlated data (Wikle, Berliner, and Cressie, 1998; Royle and Berliner, 1999) make it possible to address these questions. However, the fact that the SWE data come from multiple measurement systems complicates matters. To the extent that "more data" are better than "less data," it would be desirable to base an analysis on the data from all four systems. However, simply "lumping" all the data together may be unwise, for there may be differences in the accuracy and reliability of the four measurement systems. Although no comprehensive comparisons of the measurement systems appear to have been published, there is widespread agreement that such differences exist (though there is less agreement on the precise nature of the differences). For example, the NOHRSC website describing the AGR survey program, `www.nohrsc.nws.gov/98/html/gamma/gammapage.html`, states that the snow course data tend to underestimate actual SWE because they fail to account for water in ice at the bottom of the snow layer that the measuring tube cannot penetrate. Also, there is some evidence that variability in snow cover and forest biomass may cause AGR measurements to be systematically too low (Carroll and Carroll, 1989, 1990). Perhaps partly because of these perceived differences between measurement systems and uncertainty about how to properly account for them, previous statistical analyses of SWE addressing regional variation or climate change issues have utilized either the snow course data only (Aguado, 1990; Changnon, McKee, and Doesken, 1993; McCabe and Legates, 1995; Cayan, 1996) or the SNOTEL data only (McGinnis, 1997; Serreze et al., 1999). Other analyses (Carroll, 1995; Carroll et al., 1995; Huang and Cressie, 1996; Carroll and Cressie, 1996, 1997; Carroll, Carroll, and Poston, 1999), focussing on contemporaneous spatial prediction or very short-term forecasting, have utilized data combined from two or more systems, but they either do not account for differences among measurement systems or render the issue moot by standardizing the data (subtracting from each datum the average of all years' data at the datum's site and dividing by the standard deviation of all years' data at that site).

In this article we present a Bayesian spatio-temporal analysis of the combined SWE data from all four systems that allows for systematic differences in their accuracy and reliability. The methodology used is a major extension of that used by Isaacson and Zimmerman (2000) for combining temporally correlated data from multiple measurement systems at a single site. Here, our primary objectives are to estimate the temporal trend in SWE over the entire western U.S. and characterize how this trend varies spatially, with quantifiable estimates of variability, and to investigate whether there are systematic differences in the accuracy and reliability of the measurement systems. We accomplish these objectives using a model that allows for system-specific measurement-error biases and variances, includes effects for covariates such as latitude and elevation, and combines a conditional autoregressive (CAR) model for spatial correlations between regional values with a geostatistical model for spatial correlation among values measured at sites within regions. To the best of our knowledge, this two-stage modeling of spatial correlation has not been used previously. We also develop a modeling and computing strategy that explicitly accounts for the areal nature of the AGR measurements. We obtain estimates for all model unknowns using Markov chain Monte Carlo (MCMC) sampling. C programs for fitting our models are available from the first author's website, `www.stat.uiowa.edu/∼ kcowles`.

2 DATA

The data we analyze are SWE (in inches) on April 1, or on the closest measured date to April 1 within the two-week interval March 25 – April 8, from sites in the 11 contiguous states west of and including Montana, Wyoming, Colorado, and New Mexico (see Figure 1). Because snowpack

[Insert Figure 1 about here]

accumulations in the majority of high-elevation hydrologic basins in the western U.S. generally reach their peak by early April (Cayan, 1996), April 1 is generally regarded as a good date for indicating the available summer water supply. April 1 SWE (henceforth denoted as SWE) from all available years through 1998 (the earliest being 1910), all available sites, and all four measurement systems are included in the data set for analysis, except for a small proportion of observations that were deleted as a result of various data quality checks. The snow course, SNOTEL, and aerial marker data were obtained from the Water and Climate Center, NRCS, anonymous ftp site `ftp://wccdmp.wcc.nrcs.usda.gov/snow`, and the AGR data were obtained from the NOHRSC website `www.nohrsc.gov/98/html/gamma/gammanew.htm`. The result is a non-rectangular data set that includes $S = 2027$ sites and $T = 89$ years, with a total of $N = 70,745$ observations of SWE (57,330 snow course; 9,827 SNOTEL; 3,039 aerial marker; and 549 AGR).

Also included in our dataset are several covariates — latitude, longitude, and elevation — and the U.S. Geological Survey Hydrologic Unit Code (HUC) corresponding to each SWE observation. Latitude and longitude are measured in degrees and elevation is measured in thousands of feet. The HUC is based on a hierarchical spatial classification of watersheds containing the site where an observation was taken, with four levels of nesting ranging from HUC-2 (large regions such as California or the Pacific Northwest or major drainage basins such as the Missouri River basin) to HUC-8 (basins averaging approximately one one-hundredth the size of an HUC-2 unit).

3 MODEL

3.1 General Framework

Let $y_{s,t,m}$ denote the observed value of SWE measured in year t at site s by measurement method m. We assume that the observed SWE value is the sum of a true but unobservable value of SWE, denoted as $SWE_{s,t}$, and two additional components: systematic bias (if any) of the measurement method and random measurement error. Different measurement-error variances are assumed possible for the four measurement methods. Accordingly we write

$$y_{s,t,m} = SWE_{s,t} + \gamma_m + \epsilon_{s,t,m} \tag{1}$$

where γ_m is the systematic bias of measurement method m (m =1, 2, 3, and 4 for snow course, SNOTEL, aerial marker, and AGR data, respectively); and we suppose that the measurement errors $\{\epsilon_{s,t,m}\}$ are independently and identically distributed as $N(0, \sigma_m^2)$.

To model the true SWE, we consider several cases of the following general mixed linear model:

$$SWE_{s,t} = \sum_{k=0}^{k^*}(\beta_k + \alpha_{k,s})f_k(t) + \sum_{k=k^*+1}^{K} \beta_k f_k(lat_s,\ long_s,\ elev_s) + z_{s,t}. \tag{2}$$

Here $\beta_0, \ldots, \beta_K$ are unknown parameters; $\alpha_{0,s}, \ldots, \alpha_{k^*,s}$ are random offsets to the parameters describing temporal trend; the $f_k(\cdot)$'s are specified functions, (lat_s, $long_s$, $elev_s$) are the latitude, longitude, and elevation covariates for site s, and $z_{s,t}$ is an additive site- and time-specific "error" term that captures the effects of unobserved covariates.

Model (2) allows for an overall change in SWE over time (through $\sum_{k=0}^{k^*} \beta_k f_k(t)$) as well as local time trends (through $\sum_{k=0}^{k^*} \alpha_{k,s} f_k(t)$), while also accounting for covariates and for temporal and spatial correlation. Correlations are induced by allowing the $\alpha_{k,s}$'s and $z_{s,t}$'s to be spatially correlated and spatio-temporally correlated, respectively. A natural and interpretable way to model the correlation among the $\alpha_{k,s}$'s is by a geostatistical model, in which the correlation is a function of the distance between sites. However, such a model is not computationally feasible here; for example, estimation of spatially correlated site-specific trend parameters would require the inversion of one or more covariance matrices of dimension equal to the number of measurement sites (a 2027×2027 matrix) at each iteration of our MCMC sampler. The situation is no better for the $z_{s,t}$'s, for which we want to account for temporal as well as spatial correlation.

In an effort to model spatial and spatio-temporal correlation adequately while keeping computing time more reasonable, we modified model (2) to one in which spatial correlation operates at two levels or stages. The two levels correspond to correlation *between* subregions and correlation *within* subregions. We defined subregions for this purpose as the HUC-6 watersheds, except that nine of the watersheds were merged with their neighbors because they would have contained too few observations otherwise. The merging was determined by looking at which adjacent HUC-6 watershed had data locations closest to the points in the watershed to be merged. After merging, $R = 68$ subregions were defined, each containing from 7 to 99 observations.

The two-level model is then

$$SWE_{s,t} = \sum_{k=0}^{k^*} (\beta_k + \alpha_{k,r}) f_k(t) + \sum_{k=k^*+1}^{K} \beta_k f_k(lat_s,\ long_s,\ elev_s) + z_{s,t}. \tag{3}$$

where $\alpha_{0,r}, \ldots, \alpha_{k^*,r}$ are subregion-specific temporal trend parameters corresponding to subregion r, and r is the subregion containing s. To model spatial correlation among temporal trend parameters, we use conditional autoregressive (CAR) structures, one for each k. In these CAR models, neighbors are defined naturally as subregions sharing a common boundary. For the error terms $z_{s,t}$, we specify a within-subregion correlation structure that corresponds to a geostatistical model for the spatial component and an AR(1) model for the temporal component (described in more detail subsequently). Error terms corresponding to different subregions were regarded as independent.

Under this two-level representation of spatial correlation and the assumed independence among error terms from distinct subregions, the largest covariance matrix that must be inverted is of dimensions 99 by 99, which our linear-algebra routines can repeatedly invert sufficiently quickly. Thus model (3) is much more computationally feasible than Model (2).

Upon combining (1) with (3), we obtain the following model for our observed data:

$$y_{s,t,m} = \sum_{k=0}^{k^*}(\beta_k + \alpha_{k,r})f_k(t) + \sum_{k=k^*+1}^{K} \beta_k f_k(lat_s,\ long_s,\ elev_s) + z_{s,t} + \gamma_m + \epsilon_{s,t,m}. \tag{4}$$

We fitted specific cases of this model in which the functional dependence of $y_{s,t,m}$ on year, latitude, and longitude was linear or quadratic. For example, the linear-quadratic model (linear in year and quadratic in latitude and longitude) was

$$\begin{aligned} y_{s,t,m} \quad &= \quad (\beta_0 + \alpha_{0,r}) + (\beta_1 + \alpha_{1,r}) \cdot t + \beta_2 \cdot lat_s + \beta_3 \cdot long_s + \beta_4 \cdot lat_s^2 + \beta_5 \cdot long_s^2 \\ &\quad\quad + \beta_6 \cdot lat_s \cdot long_s + \beta_7 \cdot elev_s + z_{s,t} + \gamma_m + \epsilon_{s,t,m}. \end{aligned}$$

Note that in each of these cases, the data cannot inform about β_0 apart from the sums $\beta_0 + \gamma_m$ $(m = 1, 2, 3, 4)$. Therefore, to make all coefficients identifiable, henceforth we omit β_0 and redefine each γ_m as the sum of the overall intercept and the bias for measurement method m.

3.2 The Likelihood

Although $y_{s,t,m}$ is a well-defined notation, it is not the most useful one because the data are not rectangular, i.e. *SWE* was not measured by all measurement methods at all sites in all years. Thus we adopt the following notation, which is more efficient for nonrectangular data.

Denote by $\mathbf{Y}$ the vector of length $N = 70,745$ which is a concatenation of all observed values of the process. For observed value $y_i,\ i = 1, \ldots, N$, let m_i denote the method by which it was measured, s_i the spatial site at which it was measured, r_i the subregion within which that site is located, and t_i the index of the year in which it was measured.

Conditionally on the vector $\boldsymbol{\beta} = (\beta_k, \gamma_m)$ of overall effects, the latent vector $\mathbf{Z} = (z_{s_i,t_i})$ of spatiotemporally-correlated errors, and the vector $\boldsymbol{\alpha} = ((\alpha_{0,r}, \ldots, \alpha_{k^*,r})')$ of spatially-varying random coefficients, the likelihood may be written in the following compact matrix form:

$$
\begin{aligned}
\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{Z} &= [\mathbf{X}_O \mid \mathbf{X}_S \mid \mathbf{K}_{S,T}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ \mathbf{Z} \end{bmatrix} + \boldsymbol{\epsilon}, \\
\boldsymbol{\epsilon} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon).
\end{aligned}
$$

Here $\mathbf{X}_O$ is the design matrix for overall effects, $\mathbf{X}_S$ is the design matrix for spatially-varying random effects, and $\mathbf{K}_{S,T}$ is a matrix of zeroes and ones that matches z_{s_i,t_i} with y_i. $\boldsymbol{\Sigma}_\epsilon$ is a diagonal matrix with ith diagonal entry equal to $\sigma^2_{m_i}$, the measurement-error variance associated with measurement method m_i.

3.3 Second Stage

At the second stage of the model, all random coefficients are assumed to have zero means; their overall mean levels are absorbed into the corresponding fixed-effect coefficients. As noted previously, we use conditional autoregressive (CAR) priors to model the spatial correlations among the subregion-specific random intercept, slope, and quadratic coefficient offsets. We specify independent multivariate normal priors on the R-vectors $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_2$, respectively the random offsets to the intercept, slope, and quadratic coefficient on year for all subregions:

$$
\begin{aligned}
\boldsymbol{\alpha}_0|\tau_0^2, \phi_0 &\sim N(\mathbf{0}, \tau_0^2\boldsymbol{\Sigma}(\phi_0)), \\
\boldsymbol{\alpha}_1|\tau_1^2, \phi_1 &\sim N(\mathbf{0}, \tau_1^2\boldsymbol{\Sigma}(\phi_1)), \\
\boldsymbol{\alpha}_2|\tau_2^2, \phi_2 &\sim N(\mathbf{0}, \tau_2^2\boldsymbol{\Sigma}(\phi_2)).
\end{aligned}
$$

Here $\boldsymbol{\Sigma}(\phi_0) = (\mathbf{I}_R - \phi_0\mathbf{A})^{-1}$, where $\mathbf{I}_R$ denotes the identity matrix of dimension R and $\mathbf{A}$ is an $R \times R$ "adjacency matrix" — i.e. $A_{ij} = 1$ if subregions i and j are neighbors and 0 otherwise. Pairs of subregions are defined as neighbors if they share a common boundary. Thus ϕ_0 models spatial correlations among subregion-specific intercepts and τ_0^2 is a scale parameter. The covariance structure of the subregion-specific slopes and quadratic coefficients on year are defined analogously.

The distinct parameters ϕ_0, ϕ_1, and ϕ_2 allow for the possibility of different degrees of spatial correlation among the intercepts, slopes, and quadratic coefficients.

Let $\mathbf{Z}_r$ be the subvector of $\mathbf{Z}$ corresponding to true values of SWE within subregion r. Let T_r denote the number of years spanned from the earliest to the latest SWE measurement in subregion r and S_r the total number of distinct sites in subregion r at which SWE was measured in any year. We assume that the correlation structure of $\mathbf{Z}_r$ is separable into a spatial component and a temporal component. Unlike the random coefficients, the elements of $\mathbf{Z}_r$ are estimated at each individual point site or, in the case of AGR flight lines, at the midpoint of each flight line, rather than over regions. (The possible effects of regarding the AGR data as point rather than areal data are investigated in Section 4.6.) Geostatistical models, in which spatial correlation is modeled as a function of the distances between point sites, are better-suited than CAR models to such data, particularly when sites are as irregularly spaced as the SWE measurement locations.

Accordingly, we specify the following second-stage prior on each $\mathbf{Z}_r$:

$$\mathbf{Z}_r|\tau_Z^2, \rho_Z, \phi_Z \sim N\left(\mathbf{0}, \frac{\tau_Z^2}{(1-\rho_Z^2)}\mathbf{\Sigma}_r(\rho_Z) \otimes \mathbf{\Sigma}_r(\phi_Z)\right).$$

Here $\tau_Z^2/(1-\rho_Z^2)$ is the overall variance of the $z_{s,t}$'s; $\mathbf{\Sigma}_r(\rho_Z)$ is a $T_r \times T_r$ AR(1) correlation matrix capturing temporal correlations between $z_{s,t}$'s at the same site across time; and $\mathbf{\Sigma}_r(\phi_Z)$ is an $S_r \times S_r$ correlation matrix expressing spatial correlation among $z_{s,t}$'s measured at the same time across sites. We used the isotropic spherical correlation function — i.e. if sites j and k, $j \neq k$, are separated by distance d, then the j,k-th element of $\mathbf{\Sigma}_r(\phi_Z)$ is $\frac{1}{2}(\phi^3 d^3 - 3\phi d + 2)$ if $d \leq \frac{1}{\phi}$, and 0 otherwise.

Our model is exactly equivalent to a model in which $\alpha_{0,r}$ is omitted from the linear predictor for y_i and $\mathbf{Z}_r$ has second-stage prior $\mathbf{Z}_r|\tau_Z^2, \rho_Z, \phi_Z \sim N\left(\mathbf{1}\alpha_{0,r}, \frac{\tau_Z^2}{(1-\rho_Z^2)}\mathbf{\Sigma}_r(\rho_Z) \otimes \mathbf{\Sigma}_r(\phi_Z)\right)$. Thus, correlations between z's in different subregions are modeled indirectly through the correlations between the $\alpha_{0,r}$'s.

3.4 Priors

We specify semiconjugate priors on the regression coefficients and variances as follows:

$$\tau_Z^2 \sim IG(553.0, 553.0),$$

$$
\begin{aligned}
\tau_0^2 &\sim IG(4.25, 4.25),\\
\tau_1^2 &\sim IG(4.25, 0.0425),\\
\tau_2^2 &\sim IG(4.25, 0.0425),\\
\sigma_m^2 &\sim IG(1.1, 1.0), \quad m = 1, \ldots, 4,\\
\beta_j &\sim N(0, 10000) \text{ for all } j,\\
\gamma_m &\sim N(0, 10000), \quad m = 1, \ldots, 4.
\end{aligned}
$$

Proper priors on the variances of unobserved quantities (τ_Z^2, τ_0^2, τ_1^2, and τ_2^2 in our model) are required to ensure propriety of the joint posterior. We specify inverse gamma priors, parameterized such that $IG(a, b)$ implies a mean of $b/(a-1)$. Lacking reliable previous information about the likely magnitudes of these variances, we use "common sense" to determine prior means and then give the priors far smaller weight than the data. The inverse gamma prior on τ_Z^2 is roughly equivalent to the information contained in 1106 previously-observed $z_{s,t}$'s with mean squared deviation of 1. Since there are 70,745 observations in the dataset, this prior has only 1/64 as much weight as the data. We expect the variability of the subregion-specific slopes and quadratic coefficients to be much smaller than that of either the subregion-specific intercepts or the $z_{s,t}$'s. Accordingly, the inverse gamma priors on τ_0^2, τ_1^2, and τ_2^2 correspond to 8.5 prior observations with mean squared deviations of 1, .01, and .01 respectively. These priors have 1/8 as much weight as the 68 sets of subregion-specific estimated intercepts, slopes, and quadratic coefficients. See Section 4.2 for results of a sensitivity study to assess the influence of the above priors on estimation of parameters of interest.

Because the biases and measurement-error variances of the four measurement methods, as well as the overall slope on year, are of primary inferential interest in this analysis, we want the data to drive estimation of these parameters with essentially no influence of priors. Accordingly, we specify extremely vague priors on σ_m^2, $m = 1, \ldots, 4$ and the overall coefficients $\boldsymbol{\beta}$. For the remaining model parameters, no families of semiconjugate prior distributions exist. In addition, we have no reliable prior information about these parameters. Thus we specify the following vague priors:

$$
\phi_Z \sim U(0.005, 1.0),
$$

$$\rho_Z \sim TN_{[0.01,0.99]}(0.5, 100),$$
$$\phi_i \sim U(-0.1744, 0.1744),\ i = 1, 2, 3.$$

Distances between sites were measured in units of 10 miles, so the uniform prior on ϕ_Z represents the belief that the range (the distance at which spatial correlation decays to 0) is between 10 and 2000 miles. The prior on ρ_Z is a vague normal centered at 0.5, truncated to positive values, and bounded slightly away from 1.0 to avoid numerical problems. Using a normal rather than a uniform prior slightly simplifies the MCMC sampling algorithm for this parameter. The priors on ϕ_0, ϕ_1, and ϕ_2 are chosen to ensure that $\mathbf{\Sigma}(\phi_0)$, $\mathbf{\Sigma}(\phi_1)$, and $\mathbf{\Sigma}(\phi_2)$ are positive definite. Note that $\Sigma_{\phi_0}^{-1} = (I_R - \phi_0 A)$, where A is the adjacency matrix of the subregions, is assumed *a priori* to be positive definite. If $\lambda_j,\ j = 1, \ldots, R$ denote the eigenvalues of A, then the eigenvalues of $\Sigma_{\phi_0}^{-1}$ are $1 - \phi_0\lambda_j,\ j = 1, \ldots, R$. An upper endpoint for the uniform prior on ϕ_0 is obtained by the requirement that $1 - \phi_0\lambda_j > 0,\ j = 1, \ldots, R$. We choose the lower endpoint to make the interval symmetric around 0. Precisely the same argument applies to ϕ_1 and ϕ_2.

4 MODEL FITTING, COMPARING, AND CHECKING

4.1 Candidate Models

We considered three specific cases of the general model in (4). To facilitate convergence of the MCMC samplers used for model fitting, we centered all continuous covariates. Thus the first stage of model "QQ" (quadratic in time and quadratic in latitude and longitude) was:

$$\begin{aligned}
y_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{Z}, \sigma^2_{m_i} \sim\ & N[\alpha_{0,r_i} + (\beta_1 + \alpha_{1,r_i}) \cdot ((t_i - 1910) - 64.144) \\
& +(\beta_2 + \alpha_{2,r_i}) \cdot ((t_i - 1910)^2 - 4375.14) \\
& +\beta_3 \cdot (lat_{s_i} - 42.85) + \beta_4 \cdot (long_{s_i} + 113.89) \\
& +\beta_5 \cdot (lat^2_{s_i} - 1848.54) + \beta_6 \cdot (long^2_{s_i} - 12996.53) \\
& +\beta_7 \cdot (lat_{s_i} \cdot long_{s_i} + 4887.39) \\
& +\beta_8 \cdot (elev_{s_i} - 7035.68) + z_{s_i,t_i} + \gamma_{m_i},\ \sigma^2_{m_i}].
\end{aligned}$$

Model LQ (linear in time but with a quadratic spatial trend surface) omitted β_2, and model LL (linear/linear) omitted β_5, β_6, and β_7 as well as β_2.

In all of these models, AGR data values were treated as if they were measured at point sites. Our experience with a different modeling approach that explicitly accounts for the areal nature of AGR measurements is described in section 4.6.

4.2 Sensitivity Analysis

We used MCMC methods to fit our models. Using MCMC requires a delicate balance: priors on variance components must be strong enough to enable the MCMC sampler to converge while still enabling the data to drive posterior inference (Gelfand and Sahu, 1999). To investigate the influence of priors on the inferences of greatest interest in this analysis, we fit model LL using four different priors on the variances that require proper priors, as shown in Table 1. All other priors were fixed at the values given in Section 3.4.

[Insert Table 1 about here]

In the remainder of this section, the models are compared with respect to goodness of fit, MCMC convergence, and inferential results. Because of the preferable performance of Prior 1 for model LL (described below), we used only it in fitting models LQ and QQ.

4.3 Model Fitting and Convergence Assessment

We coded our samplers in C. For each model, we initialized three parallel chains at overdispersed starting values. We assessed MCMC convergence of all model parameters (except the over 70,000 spatiotemporal errors $\mathbf{z}$) by means of trace plots and the Gelman and Rubin (1992) convergence diagnostic, as modified by Brooks and Gelman (1998). A commonly-applied rule of thumb is that MCMC chains have converged satisfactorily if the .975 quantile of the PSRF is < 1.20 for all model parameters. Brooks and Gelman's multivariate potential scale reduction factor (MPSRF), a very conservative upper bound on the values of all the univariate PSRFs for a given model, was also computed. By default, both the Gelman and Rubin and the Brooks and Gelman diagnostics are applied to the second half of sampler output; in the case of our samplers, this was iterations 1001-2000 for LL models and iterations 5001 - 10000 for LQ and QQ.

Table 2 summarizes convergence analysis using the Brooks and Gelman diagnostic. With 2000 sampler iterations, each of the LL models using informative priors on precisions of random effects showed little evidence of convergence problems, with the .975 quantile of the PSRF exceeding 1.20 for only a few parameters and, in those cases, not by much. In contrast, with 2000 iterations for the LL model with vague priors on all variances, the .975 quantile of the univariate corrected PSRF exceeded 1.20 for nearly 2/3 of model parameters (102 out of 155), and in several cases, the values were huge (e.g. 75.0, 111.2, 139.9). The model with vague priors is omitted from all subsequent discussion because of sampler convergence failure.

[Insert Table 2 about here]

Use of squared covariate values in linear models induces correlations between the predictor variables, which in turn causes correlation between their coefficients. High cross-correlations among parameters slow MCMC sampler convergence. Thus it is not surprising that 10,000 iterations were required for reasonable sampler convergence for models LQ and QQ.

As an example of computer run-time, each $10,000$-iteration chain of model LQ took just under 11 hours on a 733-MHz Pentium III PC. The Bayesian Output Analysis software package (Smith, 2000) was used for all convergence assessment and output analysis.

4.4 Model Comparison and Selection

4.4.1 Comparisons Among the Three Priors for Model LL

We employed two methods of model comparison and selection. First, we used the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) as a global criterion for model comparison. For a generic parameter vector $\boldsymbol{\theta}$ (possibly including random effects and/or latent variables), the DIC is defined as $\bar{D} + p_D$. Letting $l(\boldsymbol{\theta})$ denote the likelihood and $\hat{\boldsymbol{\theta}}$ denote the posterior mean, then $D = -2 \ log \ l(\theta)$, $\bar{D} = E_{\boldsymbol{\theta}|\mathbf{y}}(D)$, and $p_d = \bar{D} - D(\hat{\boldsymbol{\theta}})$ where $D(\hat{\boldsymbol{\theta}}) = -2 \log l(\hat{\theta})$. $\bar{D}$ measures the quality of the model fit to the data, while p_D, the effective number of parameters, is a penalty for model complexity. For our models, p_D in part measures the contribution of the spatially-varying random coefficients $\boldsymbol{\alpha}$ and the spatiotemporally-correlated errors $\mathbf{Z}$ to the total number of parameters. Due to correlations among these variables, p_D is smaller than their actual total number. For each model, we computed a separate DIC value based on the last half of the iterations from each of the three independent sampler chains used to fit that model (the same iterations that were combined to estimate the posterior marginal distributions of model parameters.) Our overall estimate of the DIC and its standard error for that model are the mean and standard deviation, respectively, of the three independent estimates. Comparisons are shown in Table 3. For model LL, priors 1 and 2 are not distinguishable with respect to the DIC, whereas prior 3 is markedly poorer.

[Insert Table 3 about here]

Table 4 shows parameter estimates and credible sets for model LL using the two most different priors, 1 and 3. Note that inferences on the model's mean structure parameters, as well as the quantities of primary inferential interest – the differences between relative biases and the ratios of measurement-error variances among the four measurement methods – are practically unaffected by the choice of priors. Not surprisingly, since only the priors on variance components differ, inference regarding magnitudes of these variance components is not so robust. In particular, the posterior mean of σ_Z^2 is only half as large under Prior 3 as under Prior 1, with a compensating increase in the posterior means of the measurement error variances $\sigma_m^2,\ m = 1, \ldots, 4$ under Prior 3.

[Insert Table 4 about here]

Based on these results for model LL, we concluded that inference regarding the quantities of greatest interest was robust to different prior specifications and that we could safely fit our more complex models using only Prior 1.

4.4.2 Comparisons Among Models LL, LQ, and QQ

Estimated DICs for models LQ and QQ are too imprecise to be very useful in comparing these models with each other and with model LL. This is explained at least in part by the fact that it was more difficult to estimate ϕ_Z in the models with the more complicated (quadratic) spatial trend surface. Since there are over 70,000 z's compared to only 155–229 other parameters, estimation of the "effective number of parameters," and consequently of the DIC, is very sensitive to the estimated correlation between the z's.

Since the DIC did not help in choosing among models LL, LQ, and QQ, and since these models are nested, we next checked whether the parameters included only in the larger models appeared to be important. As shown in Table 4, in the LQ model, the 95% credible sets for the coefficients of lat^2, $long^2$, and $lat{\cdot}long$ all exclude zero. We concluded that the quadratic trend surface was preferable to the linear trend surface. (For this reason we chose not even to fit the model with a linear trend surface and quadratic time effect.) However, for the QQ model, the posterior mean of the coefficient for year^2 was -0.00176, and the 95% credible set was (-0.0106, 0.0074). We concluded that the quadratic term in time was not a useful predictor, and chose to base our inference on model LQ.

4.5 Model Checking

Having determined that model LQ was the best of the three models, to assess its adequacy we computed the posterior predictive p-value (Gelman, Meng and Stern, 1996) for each observation. These indicated that the model predicted values poorly at certain sites in the Cascade and Sierra Nevada mountain ranges, suggesting that the combined effect of longitude and elevation on *SWE* is more complicated than that specified by model LQ. Although we may explore this issue further in future studies, the vast majority of observations were predicted reasonably and we concluded that model LQ had adequate fit for our purposes.

4.6 Accounting for Areal Measurements

An AGR measurement, unlike the other three SWE measurements, is an average over an area (i.e. a flight path). Properly accounting for this fact requires modification of the likelihood for the AGR measurements. If y_i is an AGR measurement and B_i is the corresponding flight path of area $|B_i|$, the likelihood for y_i under model LL is given by

$$\begin{aligned} y_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{Z}, \sigma_4^2 \quad \sim \quad & N\Bigg(\alpha_{0,r_i} + (\beta_1 + \alpha_{1,r_i})t_i + \beta_2 \frac{1}{|B_i|}\int_{B_i} latitude_s \; ds + \beta_3 \frac{1}{|B_i|}\int_{B_i} longitude_s \; ds \\ & + \beta_4 \frac{1}{|B_i|}\int_{B_i} elevation_s \; ds + \gamma_4 + \frac{1}{|B_i|}\int_{B_i} z_{s,t_i} \; ds, \sigma_4^2\Bigg). \end{aligned}$$

In addition, the covariance of two AGR measurements y_i and y_j is

$$\frac{1}{|B_i||B_j|}\int_{B_i}\int_{B_j}\mathrm{Cov}(z_{s,t_i}, z_{u,t_j})\ ds\, du\ ,$$

and the covariance of an AGR measurement and an observation y_j at point site s_j from one of the other networks is

$$\frac{1}{|B_i|}\int_{B_i}\mathrm{Cov}(z_{s,t_i}, z_{s_j,t_j})\ ds\ .$$

To identify the areas B_i, we obtained from the NOHRSC the vertices of all western U.S. AGR flight paths. These vertices define each flight path as a series of connected line segments. However, segment-level AGR measurements are not available; indeed, NOHRSC records the data only as averages over entire flight paths.

To determine whether explicitly accounting for the areal averaging would change our inference, we carried out two parallel analyses. The first analysis used exactly the same model and fitting method described previously for model LL and treated the AGR measurements as point-site measurements. In the second analysis, we used two numerical integration algorithms within our MCMC sampler to approximate the integrals in the above likelihood and spatial correlation structure. Using the midpoint rule, the estimate of the average value of the spatiotemporal errors z across the width of a flight path at a particular vertex is the single value of z measured at the midpoint, i.e. at the vertex itself. Then, using the trapezoidal rule, the average value of z's over an entire rectangular segment is estimated by the average of the cross-sectional averages from the vertices forming the endpoints of the segment. Then the average over the entire flight path may be estimated as the weighted average of the segment-specific averages, with weights proportional to the lengths of the segments. The same weights are applied in approximating the integrals for covariates and correlations. Thus, this method of numerical integration may be carried out within the MCMC sampler by replacing each single AGR observation with a set of appropriately-weighted observations located at pseudo-sites defined by the vertices of the flight path. The weights for all observations corresponding to a single flight path sum to one.

To save computing time, we carried out a pilot study of parallel analyses on a subset of our dataset, specifically those 122 sites in the Upper Colorado River basin located north of the 40th parallel of latitude, which comprise four subregions. Of the 3240 measurements in this dataset, 178 were obtained by the AGR method. Including all flight vertices would have added 736 pseudo-sites to the analysis. Because computing time for these models is approximately proportional to the cube of the number of distinct sites, this was prohibitive. Therefore we deleted some flight vertices such that the shortest segments defined by the remaining consecutive sets of vertices were 4 miles long. After this procedure, there were only 157 pseudo-sites.

Since segment-level measurements of AGR were unavailable, all pseudo-sites corresponding to flight vertices for a single observation were assigned the areal average as the measured value y_i of SWE. Note, however, that we were able to use the correct covariate values for latitude, longitude, and elevation at pseudo-sites.

Table 5 shows that none of the quantities of primary inferential interest (in particular differences between relative biases and ratios of measurement-error variances) are substantially different in the two parallel analyses. However, estimates of nuisance parameters are different. The attempt to account for areal averaging introduced sites that were closer together than those in the original dataset, and this affected the estimation of ϕ_Z, the parameter capturing spatial correlation between the spatiotemporal errors. Values of this parameter in turn affect estimation of ρ and σ_Z^2.

[Insert Table 5 about here]

Because accounting for areal measurements is computationally intensive and did not affect important inferences in our pilot study, we chose not to do it in modeling our full dataset. However, this computing approach would be very useful for data involving areal averages measured over larger regions.

5 RESULTS

Iterations 5001-10000 (model LQ) or iterations 1001-2000 (model LL) from each of the 3 chains for each model were combined for all parameter estimation regarding that model. Results are summarized in Table 4 and Figures 2 and 3. Note that SWE tends to increase with increasing elevation at a rate of about 8 inches per 1000 feet. Also note that longitudes in the western

U.S. are negative, with larger magnitudes being farther west. Under the LL model, the estimated coefficients of latitude and longitude clearly indicate that, as expected, SWE tends to increase as one moves to the north and to the west (the latter being towards the Pacific Ocean, which is the main source of moisture for the western U.S. snowpack). Coefficients relating to spatial location are more difficult to interpret under model LQ, but predictions are similar under the two models. Table 6 illustrates this similarity by giving predictions (based on posterior means of parameters) for two observations in the dataset, both taken in 1936 by the snow course method.

[Insert Table 6 about here]

As for temporal trend, note that regardless of model or prior, the 95% credible set for the coefficient of year lies to the left of zero. This indicates that the overall (i.e. spatially averaged) trend in SWE in the western U.S. has been negative in the past 90 years. The situation is more complex when examined by subregions, however. Figure 2 is a map of the means of the posterior distributions (using model LQ) of the subregion-specific offsets to the intercept. Large positive values are estimated in the southernmost subregions (in Arizona and New Mexico), where most of the observed values of SWE are 0; without the positive offsets, predicted values would be negative due to the influence of latitude and elevation. The high degree of spatial correlation among these offsets is apparent in the clustering of regions with the same shade. Figure 3 is a map of the posterior means of the subregion-specific slopes on year (again using model LQ). (These are the sums of the overall slope and the subregion-specific offset.) Subregions in which the 95% central posterior intervals for the slope were entirely negative are cross-hatched. There were no subregions in which the 95% posterior interval was entirely to the right of zero. Positive spatial correlation is apparent in this map, although it does not appear as strong as that among the intercept offsets. The intervals provide substantial evidence of a negative slope in western Washington and Oregon, northeastern California, and parts of Idaho and Montana, while there is little evidence of non-zero slope in the southern half of the western U.S.

[Insert Figures 2 and 3 about here]

Finally, consider systematic differences attributable to measurement system. (Note that the smaller estimates of all the measurement-method-specific intercepts under model LQ than under model LL are due to the effect of centering the quadratic and interaction terms that appear only in the larger model. As shown in the table above, predicted values of SWE are not systematically lower under model LQ than under model LL.) Under the criterion that the data provide evidence of a difference in relative biases between two measurement systems if the 95% posterior credible set for the difference does not include zero, we conclude that the SNOTEL system yields systematically somewhat higher SWE measurements than the snow course and aerial marker systems, while the AGR system yields substantially higher SWE measurements than the other three systems. Similarly, using the criterion that the data provide evidence of a difference in measurement-error variances if the 95% posterior credible set for the ratio of the two measurement-error variances does not include 1.0, we conclude that there are meaningful differences between all pairs of measurement systems, with the ordering from smallest to largest measurement-error variance being SNOTEL, snow course, AGR, and aerial marker.

6 DISCUSSION

One of the main objectives of this work was to characterize the temporal trend (from 1910-1998) in April 1 SWE, both over the entire western U.S. and regionally. Our results indicate that there is a significant negative trend in SWE overall, but there are regional differences. In the northern Rockies and the Cascades of the Pacific Northwest, the trend tends to be negative, with SWE decreasing at a rate of about 0.1 to 0.2 inches per year. In contrast, SWE in the intermountain region and southern Rockies has not changed significantly. These results reinforce more tenuous conclusions made by previous authors. For example, Changon et al. (1993) and McCabe and Legates (1995) studied snow course data taken from 1951-1985 and 1948-1987, respectively, at 275 and 311 sites, respectively. Both previous analyses indicated a decreasing trend in SWE at most sites in the Pacific Northwest and an increasing trend at some sites in the southern Rockies. Actual point or interval estimates of trend were not given, however. Our analysis, in addition to being based on more than six times as many sites and a time period twice as long, provides, through the posterior distributions of model unknowns, for much more specific inferences and conclusions.

Our other main objective was to compare the systems used to measure SWE. In this regard the most surprising finding was the large bias observed in AGR measurements relative to the other measurements. This bias has not previously been documented and certainly is cause for concern. Of course, lacking a “gold standard” in this situation, it is unknown whether the relative bias results from overestimation by the AGR system or underestimation by the other three systems, or both. Clearly, possible explanations for the differences in mean SWE should be sought, and any analysis that combines AGR data with other data without adjusting for these differences should be regarded with suspicion.

Using a dataset much less extensive than ours, Serreze et al. (1999) found no evidence of a relative bias between SNOTEL and snow course measurements. We found differently: after adjusting for effects of latitude, longitude, elevation, and year the SNOTEL measurements were systematically higher on average by slightly more than one inch.

Regarding measurement-error variances, we found that the SNOTEL system is more reliable than the snow course system. This is consistent with what Huang and Cressie (1996) conjectured regarding these two systems. We also found that both the SNOTEL and snow course systems are more reliable than the aerial marker system, which is reasonable in light of the latter’s cruder measurement technique. The inferior reliability of the AGR system relative to the two ground-based systems, however, does not comport with claims made by Carroll et al. (1995), who extol the virtues of the AGR system. One factor contributing to the larger variability of AGR measurement errors could be the discrepancy in the date of actual measurement from the nominal April 1 date. The variance of the deviations in actual date of measurement from April 1 is equal to zero for SNOTEL and is approximately 33% larger for AGR measurements than for snow course measurements. Thus, on average there is more time for snowmelt or additional snowfall to occur between the times an AGR measurement is taken and nearby measurements from other systems are taken. On the other hand, the AGR measurements are areal averages rather than point values, so they are not subject to the very small-scale spatial variation to which the other measurements are subject. Clearly, to resolve these issues will require further investigation.

ACKNOWLEDGEMENTS

We thank an associate editor and a referee for their comments which improved the quality of this article. The work of Cowles, Zimmerman, and Christ was supported by the United States Environmental Protection Agency under grant R826887-01-0. McGinnis was supported by the National Science Foundation under grant EAR-9634329. This research has not been subject to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

Aguado, E. (1990), "Elevational and latitudinal patterns of snow accumulation and departures from normal in the Sierra Nevada," *Theoretical and Applied Climatology*, 42, 177-185.

Brooks, S.P. and Gelman, A. (1998), "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, 7, 434-455.

Carroll, S.S. (1995), "Modeling measurement errors when estimating snow water equivalent," *Journal of Hydrology*, 172, 247-260.

Carroll, S.S. and Carroll, T.R. (1989), "Effect of uneven snow cover on airborne snow water equivalent estimates obtained by measuring terrestrial gamma radiation," *Water Resources Research*, 25, 1505-1510.

Carroll, S.S. and Carroll, T.R. (1990), "Estimating the variance of airborne snow water equivalent estimates using computer simulation techniques," *Nordic Hydrology*, 21, 35-46.

Carroll, S.S., Carroll, T.R., and Poston, R.W. (1999), "Spatial modeling and prediction of snow-water equivalent using ground-based, airborne, and satellite snow data," *Journal of Geophysical Research*, 104, 19623-19629.

Carroll, S.S. and Cressie, N. (1996), "A comparison of geostatistical methodologies used to estimate snow water equivalent," *Water Resources Bulletin*, 32, 267-278.

Carroll, S.S. and Cressie, N. (1997), "Spatial modeling of snow water equivalent using covariance estimated from spatial and geomorphic attributes," *Journal of Hydrology*, 190, 42-57.

Carroll, S.S., Day, G.N., Cressie, N., and Carroll, T.R. (1995), "Spatial modelling of snow water equivalent using airborne and ground-based snow data," *Environmetrics*, 6, 127-139.

Cayan, D.R. (1996), "Interannual climate variability and snowpack in the western United States," *Journal of Climate*, 9, 928-948.

Changnon, D., McKee, T.B., and Doesken, N.J. (1993), "Annual snowpack patterns across the Rockies: Long-term trends and associated 500-mb synoptic patterns," *Monthly Weather Review*, 121, 633-647.

Gelfand, A.E. and Sahu, S.K. (1999), "Identifiability, improper priors, and Gibbs sampling for generalized linear models," *Journal of the American Statistical Association*, 94, 247-253.

Gelman, A., Meng, X.-L., and Stern, H. (1996), "Posterior predictive assessment of model fitness via realized discrepancies" (with discussion), *Statistica Sinica*, 6, 733-760.

Gelman, A. and Rubin, D.B. (1992), "Inference from iterative simulation using multiple sequences" (with discussion), *Statistical Science*, 7, 457-472.

Gleick, P.H. (1987), "Regional hydrologic consequences of increases in atmospheric CO_2 and other trace gases," *Climate Change*, 10, 137-161.

Hartman, R.K., Rost, A.A., and Anderson, D.M. (1995), "Operational processing of multi-source snow data," *Proceedings of the Western Snow Conference*, pp. 147-151.

Huang, H.C. and Cressie, N. (1996), "Spatio-temporal prediction of snow water equivalent using the Kalman filter," *Computational Statistics and Data Analysis*, 22, 159-175.

Isaacson, J.D. and Zimmerman, D.L. (2000), "Combining temporally correlated environmental data from two measurement systems," *Journal of Agricultural, Biological, and Environmental Statistics*, 5, 64-83.

Lettenmaier, D.P. and Sheer, D.P. (1991), "Climatic sensitivity of California water resources," *Journal of Water Resources Planning and Management*, 117, 108-125.

McCabe, A.J. and Legates, S.R. (1995), "Relationships between 700 hPa height anomalies and 1 April snowpack accumulations in the western USA," *International Journal of Climatology*, 14, 517-530.

McGinnis, D.L. (1997), "Estimating climate-change impacts on Colorado Plateau snowpack using downscaling methods," *Professional Geographer*, 49, 117-125.

Palmer, P.L. (1988), “The SCS snow survey water supply forecasting program: Current operations and future directions,” *Proceedings of the Western Snow Conference*, Kalispell, MT, pp. 43-51.

Royle, J.A. and Berliner, L.M. (1999), “A hierarchical approach to multivariate spatial modeling and prediction,” *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 29-56.

Serreze, M.C., Clark, M.P., McGinnis, D.L., Pulwarty, R.S., and Armstrong, R.A. (1999), “Climatological characteristics of the western U.S. snowpack from SNOTEL data,” *Water Resources Research*, 35, 2145-2160.

Smith, B.J. (2000). Bayesian Output Analysis Program (BOA), Version 0.5.0 for S-PLUS and R. `http://www.public-health.uiowa.edu/BOA`.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), “Bayesian measures of model complexity and fit,” *Journal of the Royal Statistical Society, Series B*, in press.

Wikle, C.K., Berliner, L.M., and Cressie, N. (1998), “Hierarchical Bayesian space-time models,” *Environmental and Ecological Statistics*, 5, 117-154.

Table 1: Prior Distributions on Variance Parameters of Model LL.

Parameter	Prior 1	Prior 2	Prior 3	Vague
τ_Z^2	IG(553.0, 553.0)	IG(553.0, 553.0)	IG(2211, 2211)	IG(.0001,.0001)
τ_0^2	IG(4.25, 4.25)	IG(17.0, 17.0)	IG(34.0, 34.0)	IG(.0001,.0001)
τ_1^2	IG(4.25 , 0.0425)	IG(17.0, 0.17)	IG(34.0, 0.34)	IG(.0001,.0001)

Table 2: Brooks and Gelman Convergence Diagnostics.

Model	Prior	Iterations	MPSRF	# parameters evaluated	# times .975 quantile of corrected PSRF > 1.20	Largest .975 quantile of corrected PSRF
LL	Vague	2000	123479.7	155	102	139.9
LL	1	2000	52.5	155	4	1.259
LL	2	2000	47.1	155	2	1.663
LL	3	2000	25.0	155	4	2.132
LQ	1	10000	88.4	158	0	—
QQ	1	10000	92.8	229	6	1.890

Table 3: Model Comparisons Using Deviance Information Criterion. Values in parentheses are the corresponding standard errors.

Model	Prior	DIC	Deviance	Effective number of parameters
LL	1	489,101 (187)	480,752 (243)	8,349 (56)
LL	2	488,896 (124)	480,485 (158)	8,410 (34)
LL	3	498,407 (302)	492,825 (367)	5,582 (64)
LQ	1	489,179 (1330)	473,289 (1992)	7945 (343)
QQ	1	490,523 (1027)	475,405 (1400)	7559 (226)

Table 4: Parameter Estimates for the Linear/Quadratic and Linear/Linear Models. Measurement methods are coded as follows: snow course = 1, SNOTEL = 2, aerial marker = 3, AGR = 4. Parameters labeled "diff" are the relative biases between two measurement methods, with diff j, k = $\beta_j - \beta_k$. Similarly, parameters labeled "ratio" are the ratios of measurement-error variances, with ratio $j, k = \frac{\sigma_j^2}{\sigma_k^2}$.

Parameter	LQ, Prior 1 Posterior mean	LQ, Prior 1 95% credible set	LL, Prior 1 Posterior mean	LL, Prior 1 95% credible set	LL, Prior 3 Posterior mean	LL, Prior 3 95% credible set
year	-0.035	(-0.071, -0.0031)	-0.035	(-0.066, -0.0049)	-0.040	(-0.079, -0.0074)
latitude	0.783	(-2.71, 3.99)	3.55	(3.35, 3.74)	3.51	(3.36, 3.67)
longitude	22.71	(18.40, 27.53)	-2.77	(-2.62, -2.92)	-2.51	(-2.38, -2.64)
lat^2	0.096	(0.065, 0.121)				
long^2	0.120	(0.100, 0.142)				
lat $\cdot$ long	0.0456	(0.015, 0.075)				
elevation	7.92	(7.73, 8.07)	7.99	(7.9, 8.1)	7.8	(7.7, 7.8)
γ_1	17.35	(13.64, 21.10)	20.46	(16.39, 24.74)	20.48	(17.12, 23.72)
γ_2	18.57	(14.87, 22.36)	21.72	(17.66, 25.98)	21.89	(18.54, 25.16))
γ_3	17.16	(13.41, 20.98)	20.34	(16.28, 24.71)	20.36	(17.07, 23.68)
γ_4	25.29	(21.47, 29.29)	28.46	(24.28, 32.77)	28.19	(24.83, 31.61)
diff 1,2	-1.23	(-1.47, -1.00)	-1.26	(-1.44,-1.08)	-1.41	(-1.61, -1.21)
diff 1,3	0.19	(-0.29, 0.66)	0.12	(-0.35, 0.57)	0.12	(-0.36, 0.61)
diff 1,4	-7.94	(-8.94, -6.97)	-8.01	(-8.95, -7.09)	-7.71	(-8.68, -6.72)
diff 2,3	1.42	(0.92, 1.91)	1.38	(0.90,1.87)	1.53	(1.02, 2.04)
diff 2,4	-6.71	(-7.71, -5.72)	-6.75	(-7.68, -5.82)	-6.30	(-7.24, -5.32)
diff 3,4	-8.13	(-9.19, -7.03)	-8.13	(-9.20, -7.15)	-7.83	(-8.93, -6.74)
σ_1^2	52.35	(47.61, 58.55)	50.83	(49.75, 52.08)	60.09	(58.48, 61.35)
σ_2^2	46.36	(41.07, 53.39)	44.76	(43.10, 46.51)	54.72	(52.63, 56.82)
σ_3^2	135.05	(125.00, 146.41)	136.37	(128.21, 142.86)	146.21	(138.89, 153.85)
σ_4^2	100.80	(87.41, 116.82)	94.48	(82.64, 107.53)	108.46	(95.24, 123.46)
ratio 1,2	1.13	(1.08, 1.18)	1.14	(1.10, 1.18)	1.10	(1.06, 1.14)
ratio 1,3	0.39	(0.36, 0.42)	0.38	(0.35, 0.40)	0.41	(0.39, 0.44)
ratio 1,4	0.52	(0.45, 0.60)	0.54	(0.47, 0.62)	0.56	(0.49, 0.63)
ratio 2,3	0.34	(0.31, 0.38)	0.33	(0.31, 0.35)	0.37	(0.35, 0.40)
ratio 2,4	0.46	(0.39, 0.54)	0.48	(0.41, 0.55)	0.51	(0.44, 0.58)
ratio 3,4	1.35	(1.16, 1.54)	1.44	(1.24, 1.66)	1.35	(1.17, 1.55)
ϕ_0	0.164	(0.140, 0.174)	0.17	(0.145,0.174)	0.17	(0.161, 0.173)
$\sigma_{\alpha_0}^2$	45.83	(31.98, 64.72)	47.36	(33.78, 67.11)	24.36	(18.83, 31.52)
ϕ_1	0.082	(-0.066, 0.166)	0.081	(-0.067, 0.164)	0.091	(-0.026, 0.162)
$\sigma_{\alpha_1}^2$	0.0059	(0.0038, 0.0089)	0.0054	(0.0034, 0.0080)	0.0081	(0.0062, 0.0104)
ϕ_Z	0.029	(0.014, 0.044)	0.033	(0.030,0.035)	0.037	(0.035, 0.040)
ρ_Z	0.41	(0.39, 0.43)	0.43	(0.42,0.44)	0.391	(0.38, 0.40)
σ_Z^2	28.13	(25.93, 31.44)	25.64	(24.57, 26.74)	12.72	(12.04, 13.42)

Table 5: Parameter Estimates for the Pilot Study, With and Without an Explicit Accounting for the Areal Nature of the AGR Measurements. Measurement methods are coded as follows: snow course = 1, SNOTEL = 2, aerial marker = 3, AGR = 4. Parameters labeled "diff" are the relative biases between two measurement methods, with diff $j, k = \beta_j - \beta_k$. Similarly, parameters labeled "ratio" are the ratios of measurement-error variances, with ratio $j, k = \frac{\sigma_j^2}{\sigma_k^2}$.

	No integration		Integration	
Parameter	Posterior mean	95% credible set	Posterior mean	95% credible set
year	0.0044	(-0.13, 0.14)	0.0069	(-0.14,0.13)
latitude	5.41	(4.90, 5.89)	5.38	(4.83, 5.92)
longitude	-4.94	(-5.71, -4.11)	-4.63	(-5.51, -3.74)
elevation	0.0056	(0.0053, 0.0059)	0.0058	(0.0055, 0.0061)
γ_1	18.63	(7.15, 27.83)	19.04	(8.89, 27.69)
γ_2	19.74	(8.34, 28.99)	20.03	(9.88, 28.64)
γ_3	21.67	(10.10, 31.07)	21.02	(10.79, 29.77)
γ_4	27.24	(15.58, 36.60)	28.01	(17.82, 36.75)
diff 1,2	-1.11	(-1.56, -0.67)	-0.99	(-1.39, -0.61)
diff 1,3	-3.04	(-4.88, -1.24)	-1.98	(-3.92, -0.23)
diff 1,4	-8.61	(-10.04, -7.10)	-8.98	(-10.54, -7.41)
diff 2,3	-1.93	(-3.77, -0.10)	-0.99	(-2.88, 0.76)
diff 2,4	-7.50	(-8.98, -6.00)	-7.98	(-9.54, -6.43)
diff 3,4	-5.57	(-7.86, -3.21)	-6.99	(-9.35, -4.47)
σ_1^2	13.19	(10.94, 15.43)	9.90	(7.78, 13.33)
σ_2^2	13.85	(10.76, 17.01)	9.98	(7.40, 14.27)
σ_3^2	27.44	(16.45, 44.25)	22.15	(11.96, 37.31)
σ_4^2	90.37	(71.94, 113.64)	99.51	(79.37, 125.00)
ratio 1,2	0.96	(0.82, 1.11)	1.00	(0.85, 1.18)
ratio 1,3	0.51	(0.30, 0.80)	0.48	(0.27, 0.80)
ratio 1,4	0.15	(0.11, 0.19)	0.10	(0.07, 0.14)
ratio 2,3	0.54	(0.31, 0.85)	0.48	(0.26, 0.81)
ratio 2,4	0.15	(0.11, 0.20)	0.10	(0.07, 0.14)
ratio 3,4	0.31	(0.18, 0.51)	0.22	(0.12, 0.39)
ϕ_0	-0.079	(-0.45, 0.41)	-0.08	(-0.44, 0.40)
$\sigma_{\alpha_0}^2$	67.46	(26.46, 161.29)	63.98	(24.10, 156.25)
ϕ_1	0.012	(-0.43, 0.42)	0.026	(-0.42, 0.43)
$\sigma_{\alpha_1}^2$	0.0095	(0.0039, 0.022)	0.0095	(0.0039, 0.022)
ϕ_Z	0.040	(0.029,0.057)	0.069	(0.032, 0.097)
ρ_Z	0.44	(0.41,0.48)	0.37	(0.34, 0.40)
σ_Z^2	14.78	(12.66, 17.39)	16.57	(14.16, 20.16)

Table 6: Predictions for Two Observations Under Models LL and LQ. Predictions in the columns headed "Overall" use the overall coefficients only, whereas the other predictions incorporate subregional offsets to the intercept and slope on year for the subregions containing the sites of the observations.

				LL		LQ	
Observation	Latitude (degrees)	Longitude (degrees)	Elevation (feet)	Overall	With subregional offsets	Overall	With subregional offsets
1	40.37	-106.1	9,160	8.37	10.44	8.29	11.43
2	45.17	-117.3	8,200	48.77	38.01	43.47	37.75

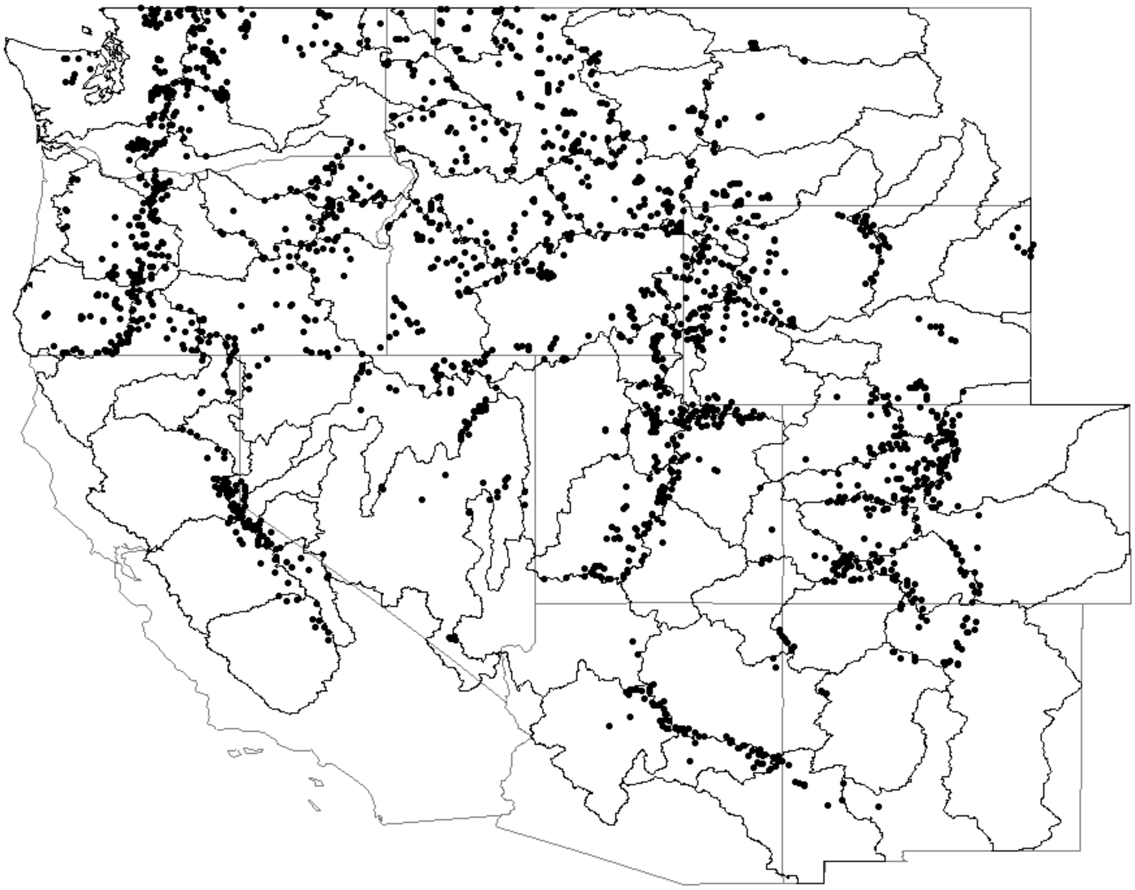

Figure 1: *Geographical locations of all SWE measurement sites.*

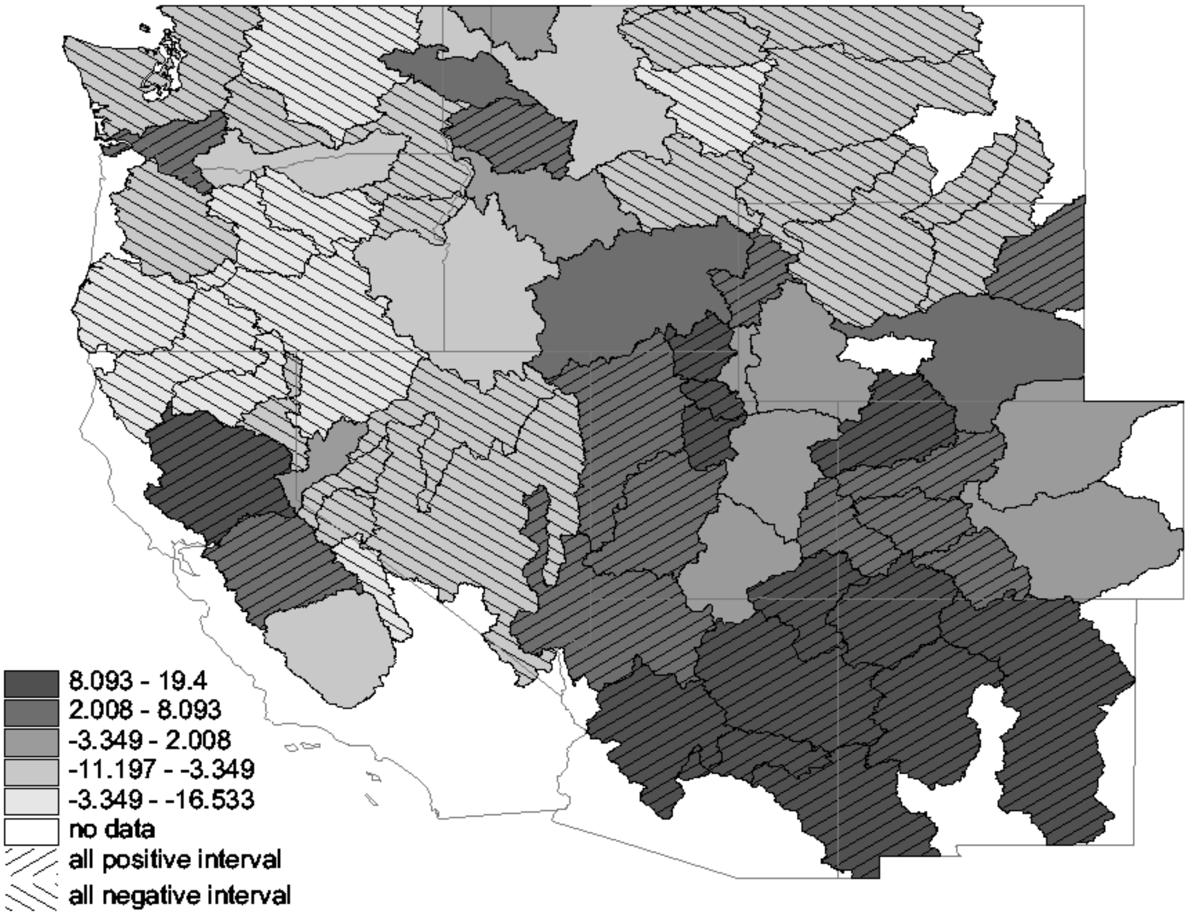

Figure 2: *Subregion-specific offsets to the SWE model overall intercept.*

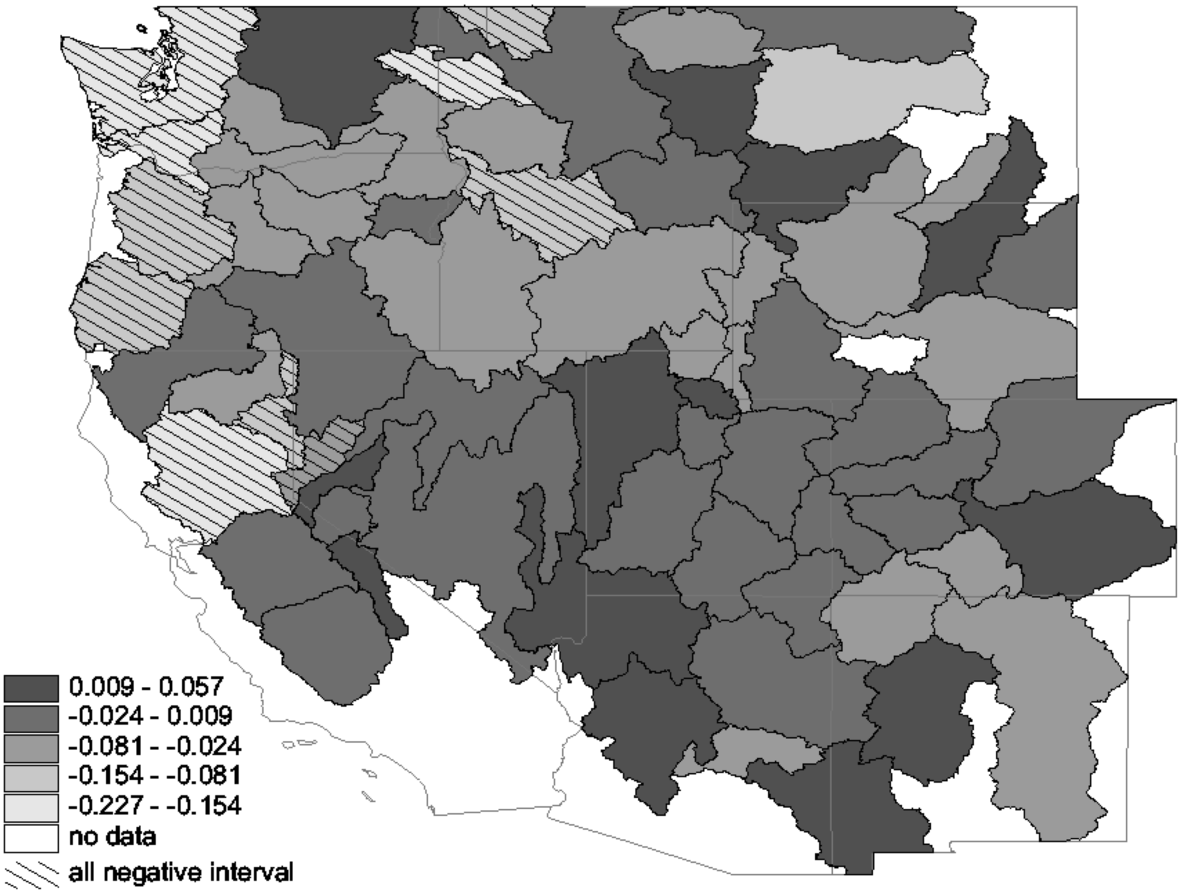

Figure 3: *Subregion-specific slopes of SWE on year adjusted for other covariates.*