Interpolation of 1961–97 Daily Temperature and Precipitation Data onto Alberta Polygons of Ecodistrict and Soil Landscapes of Canada

SAMUEL S. P. SHEN

Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada

PETER DZIKOWSKI

Conservation and Development Branch, Alberta Agriculture, Food and Rural Development, Edmonton, Alberta, Canada

GUILONG LI AND DARREN GRIFFITH

Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada

(Manuscript received 3 November 2000, in final form 7 May 2001)

ABSTRACT

Soil quality models developed for ecodistrict polygons (EDP) and the polygons of the soil landscapes of Canada (SLC) to monitor the concentration of soil organic matter require daily climate data as an important input. The objectives of this paper are (i) to provide a method that interpolates the daily station data onto the 894 SLC polygons and 150 EDP in the province of Alberta, Canada, so that the interpolated data fit not only climate mean but also climate variability, especially for the precipitation field, and hence can be used as realistic climate input to soil quality models and (ii) to understand the variability of the Alberta daily climate, such as precipitation frequency. The procedure interpolates the station data onto a dense network of grid points and then averages the gridpoint values inside polygons. The procedure and results for maximum temperature, minimum temperature, and precipitation are reported in detail. The interpolation uses the observed daily data for the period 1 January 1961–31 December 1997 (13 514 days) within the latitude–longitude box (45°–64°N, 116°–124°W). Because the precipitation field can have a short spatial correlation length scale and large variability, a hybrid of the methods of inverse-distance weight and nearest-station assignment is developed for interpolating the precipitation data. This method can reliably calculate not only the number of precipitation days per month, but also the precipitation amount for a day. The temperature field has a long spatial correlation scale, and its data are interpolated by the inverse-distance-weight method. Cross-validation shows that the interpolated results on polygons are accurate and appropriate for soil quality models. The computing algorithm uses all the daily observed climate data; despite that, some stations have a very short time record or only summer records.

1. Introduction

Agricultural use of climate data has increased considerably during the last two decades because of the rapid development of information technology, and the rate of the increase will accelerate in the future (Changnon and Kunkel 1999). This paper reports the interpolation methods used to provide the daily climatic input data required by soil quality models in Alberta, Canada.

Alberta extends from 49° to 60°N latitude and from 110° to 120°W longitude. Agriculture is found throughout the province, extending over 1000 km from the 49th parallel, where it is the most prominent industry, to the Peace River region in the northwest. The total land area of Alberta is 63.8 million hectares (or 0.638 million square kilometers), of which about one-third, 20.8 million hectares, is occupied farm land.

Alberta Agriculture, Food and Rural Development (AAFRD), a provincial government department, in partnership with the agricultural industry, has been developing a strategy for sustainable agriculture. Government agencies are becoming more aware of the need to be accountable and are trying to include specific measurable results in business plans. AAFRD is committed to environmental sustainability and is working with researchers to develop quantitative measures. Soil quality is one of the initial indicators of environmental sustainability being developed. One aspect of sustainability is to ensure land management practices maintain or improve soil quality.

Soil organic matter is one of the key soil attributes associated with soil quality. Soils with higher levels of

Corresponding author address: Dr. Samuel Shen, Department of Mathematical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada.
E-mail: shen@ualberta.ca

organic matter are generally considered to be of better quality and tend to have (i) better nutrient-retention characteristics for good crop growth; (ii) better water infiltration rates, resulting in slower rates of water erosion of soil; and (iii) better structure, reducing susceptibility to wind erosion.

Several models are being used by AAFRD to assess soil quality in Alberta. Among them are EPIC (Erosion/Productivity Impact Calculator) and WEPP (Water Erosion Prediction Project). The EPIC model was developed to assess the effect of soil erosion on soil productivity (Sharpley and Williams 1990). EPIC operates in a daily time step and requires daily climate data (radiation, maximum and minimum temperature (T_{max} and T_{min}), precipitation, relative humidity, and wind speed) and information on land-management practices. The WEPP model, also operating in daily time step, simulates the soil water content in multiple layers of soil relevant to plant growth and/or decomposition. It also simulates the effects of tillage processes and soil consolidation (Flanagan and Livingston 1995). These soil quality models apply current knowledge of crop growth and soil processes, which are influenced by climate conditions, to assess the effect of changes in land management practices, such as adoption of reduced tillage and annual cropping or perennial cover, on soil organic matter.

Alberta is developing a method to monitor changes in soil quality by using models on a provincewide scale verified by research plot data. In order to operate, most models require a complete daily climate dataset as input, with no missing data. The models quantitatively estimate the effect on soil quality due to changes in land management practices under the climate conditions used in the models. It is very important to have actual observed daily climate data on which to run the models, to compare the model results with carefully measured soil data. In a similar way, it is important to have daily climate data for operational use of the models to quantitatively estimate changes in soil organic matter. The daily data must adequately represent the actual weather conditions that occurred. Soil-erosion research has shown that severe weather events, especially heavy rainfalls or high winds, are primarily responsible for the bulk of soil erosion, which reduces soil quality.

The soil quality models are being developed in Alberta to run on Ecodistrict polygons (EDP) and Soil Landscapes of Canada (SLC) polygons. These polygons represent sufficiently uniform soil and climate conditions, suitable for provincewide land capability assessment and for the soil quality monitoring intended. The province is divided into 150 EDP (Fig. 1; Ecological Stratification Working Group 1995) and 894 SLC polygons (Fig. 2; Shields et al. 1991). There is a mismatch between the climate data available, which have been recorded at points, and the data needed for polygons. It is not clear how a climate parameter for a polygon would be directly measured.

We define the climate value for a polygon as the spatial average of the climate parameter. Therefore the polygon climate value is an estimated value. Interpolation and averaging are essential tools to obtain that estimate with minimal error.

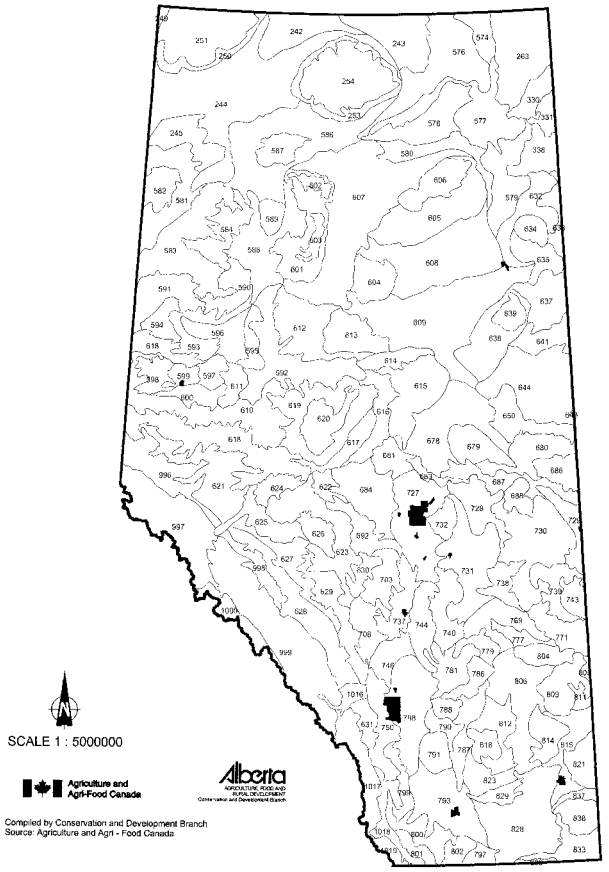

FIG. 1. The 150 EDPs in Alberta.

Various methods may be used to interpolate scattered data onto polygons, such as the Thiessen polygon method. We chose to interpolate all the available station data onto a regular grid with 10-km spacing and average the values for the grid points inside a polygon. In our current study, the 10-km spacing was chosen as well-suited to the size of the polygons and the station density in Alberta. The grid was not so dense as to cause excessive computation. This station-to-grid-to-polygon approach was successfully used earlier by Mackey et al. (1996) to recharacterize climate subregions in the province of Ontario, Canada, and was reviewed by Shen (1998).

Many methods are available to interpolate data onto grid points, such as nearest-station assignment, inverse-distance weighting, kriging, and thin-plate splines. Most of the interpolation methods are best for fitting the mean for a period of a month or longer. However, soil erosion is mostly influenced by extreme weather events, such as heavy precipitation or high winds. Thus, the input data to the soil quality models must adequately represent the real sequence of weather events in the recorded data.

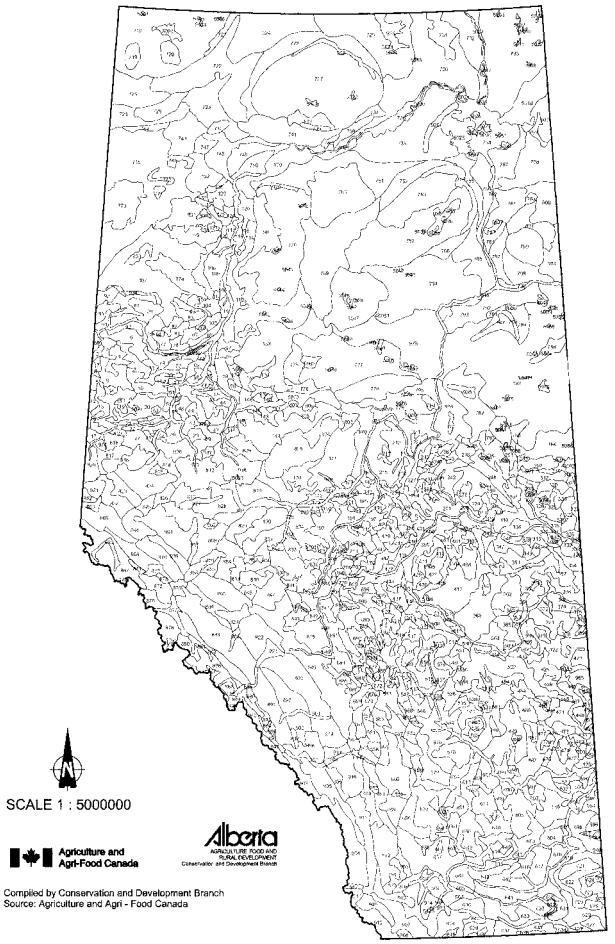

FIG. 2. The 894 SLC polygons in Alberta.

According to Karl and Knight (1998), the increase of precipitation in the United States and Canada over the last 90 years is mainly due to extreme precipitation events. Thus, it is important for the interpolated data to retain the heavy precipitation events.

Preliminary investigation showed that although some interpolation methods provide good estimates of the monthly mean precipitation, they also result in too many days with precipitation and therefore too little precipitation each day. This attribute of interpolation, if left uncorrected, would lead to underestimation of precipitation intensity and hence soil erosion.

Among the commonly used interpolation methods, the nearest-station-assignment method yields a good estimate of variance. In order to have a good assessment of climate variability, when Agriculture Canada calculated the ecodistrict climate normals for 1961–90, it used the Thiessen polygon approach for interpolation, which is equivalent to the nearest-station assignment method. The details can be found from the following Web site: http://res.agr.ca/CANSIS/NSDB/ECOSTRAT/climate_normals_1961-91.html.

The main objective of this paper is to describe a method that interpolates the daily station data onto polygons so that the interpolated data fit not only the monthly mean but also retain the appropriate number of days with precipitation and hence provide more realistic and complete (no missing data) daily climate input for use by soil quality models in Alberta.

The interpolation methods must satisfy the soil quality monitoring project needs, which are (i) to provide the best fit for both the monthly mean and the days with precipitation, (ii) to dynamically adapt to the number of stations to use all the daily data available, and (iii) to provide realistic and complete (no gaps in the dataset) daily input data for polygons to be used in soil quality models.

Our method for precipitation is a hybrid of inverse-distance weighting and nearest-station assignment. To achieve the best fit, we use all the daily observed climate data available for the period 1 January 1961 to 31 December 1997. Even stations are used that have a very short record, some with as little as two months of data. To overcome the technical difficulty of interpolating data with varying and incomplete data sources, a dynamic searching algorithm is developed in this research.

The accuracy of interpolation is assessed by cross-validation for both observed stations and polygons. The cross-validation results show that our method is reliable and appropriate for preparing realistic daily weather data for use in soil quality models.

This paper is arranged as follows. Section 2 describes the data source used for interpolation. Section 3 describes our interpolation method. Section 4 describes the assessment of the interpolation error by cross-validation. Conclusions and discussion are in section 5.

2. Data

The basic daily climate data needed for soil quality models are the following seven parameters: maximum and minimum temperature, precipitation, wind speed, wind direction, relative humidity, and incoming solar radiation. This paper reports the details only on the interpolation procedures and results for temperature and precipitation. The wind, humidity, and radiation data are only discussed briefly in various places of the paper, because there are only a few stations with these quantities.

Daily weather conditions are observed at stations distributed in a variety of networks throughout Alberta and its vicinity. Each of the networks serves a different purpose. Environment Canada operates a network of about 30 stations that support the real-time, synoptic-scale weather forecasting activities by providing hourly and daily observations via telecommunications. Environment Canada also operates a network of about 200 stations that record twice-daily readings of temperature and precipitation for characterizing the climate. Alberta Environment operates, during the summer months only, a network of about 100 stations that record the daily cli-

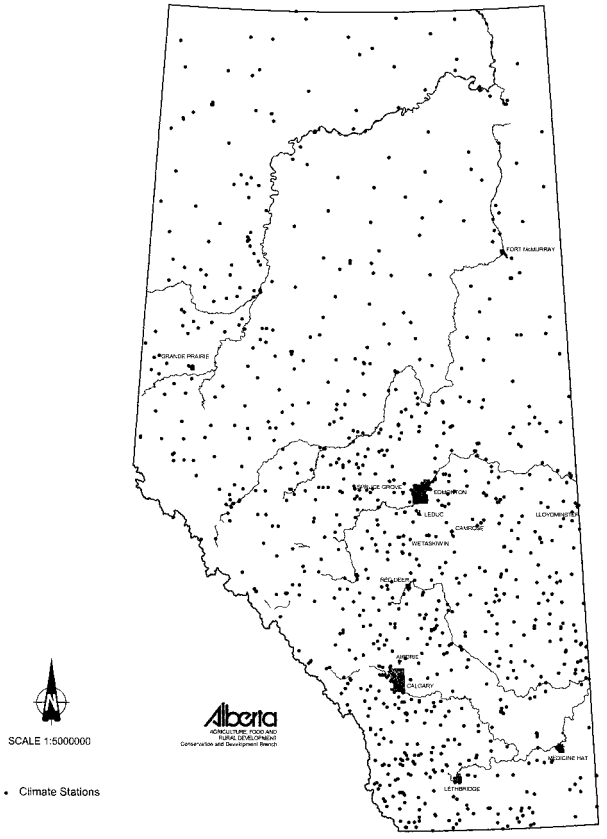

FIG. 3. Locations of the 927 temperature and precipitation stations (1961–90) in Alberta.

mate data to support forest-fire protection activities. When the physical condition of a station, most often the location, was changed, the station identification (ID) was changed accordingly. We counted the number of stations according to the station ID. Thus, it can happen that one station name, such as Edmonton, corresponded to several station IDs and hence several stations. Figure 3 shows the locations of these 927 climate stations in Alberta from 1 January 1961 to 31 December 1990.

Large differences exist in station density across Alberta as well as in data availability from the stations. The southern, more populated, and also more agriculturally intensive area has greater station density than the northern portion, which is mostly forested but also has agricultural activity. Most stations measure only daily precipitation and the daily maximum and minimum temperature. The synoptic stations usually record relative humidity, wind speed, and wind direction. Only a few stations in Alberta record solar radiation. A summary of the archived data inventory for temperature and precipitation is given in Table 1. The data were accessed in two phases. The first phase was for 1961–90 in order to match the period of climate normals. The second phase was to update the daily dataset to include the most recent data available.

Alberta is contained in the latitude–longitude box 49°–60°N, 110°–120°W with the southwest corner cut off by the Rocky Mountains. The climate stations within Alberta, and those outside the borders of the province, but within a box 4° longitude to the east and west, and 4° latitude to the north and south, were used for interpolation. Thus, our data were from the stations inside the latitude–longitude box 45°–64°N, 116°–124°W.

TABLE 1. The T_{max}, T_{min}, and Pcpn observational station summary.

	1961–90	1991–97	
	927	507	Alberta stations
	268	452	U.S. stations
	135	47	Saskatchewan stations
	37	25	British Columbia stations
	3	15	Northwest Territories stations
Total	1370	1046	Stations

Daily climate data were purchased from the Atmospheric Environment Service (AES) in Canada, and the National Oceanic and Atmospheric Administration (NOAA) in the United States. All the available daily data, except apparently incorrect data, were used for interpolation. The data had gone through quality control at both AES and NOAA, but some unusual data still existed. Additional quality control tests were used to eliminate some apparently incorrect data, such as records with maximum temperature less than minimum temperature, temperature greater than 45°C and less than −70°C, daily precipitation greater than 250 mm, daily radiation greater than 50 MJ m^{-2}, wind speed greater than 150 km h^{-1} and wind direction greater than 360°.

Most of the stations had incomplete data records, varying from some having a few missing days, to some having only a few months of data. A complete dataset would be the number of stations multiplied by the number of days, which we call the station-days. For the climatological period 1961–90, of the total possible station-days, only 32% had data for temperature, and 39% had data for precipitation.

3. Interpolation method

a. Review of conventional interpolation methods

Interpolation belongs to a very useful branch of mathematics, called approximation theory. Statistical interpolation considers fields having random fluctuations. Various interpolation methods have been invented to solve interpolation problems in practical applications. Many traditional interpolation methods are summarized in a comprehensive book by Cressie (1993).

Commonly used interpolation methods for meteorological applications include nearest-station assignment, inverse-distance weighting, inverse-distance-square weighting, Thiessen-polygon method, orthogonal-polynomial approximation, Lagrange method, interpolation by splines, kriging, and interpolation by empirical orthogonal functions. Each method has its merits and is

applicable according to temporal length scale, spatial length scale, stationarity, and variability of the field under consideration. In the following, a few relevant methods are briefly reviewed and assessed for their suitability for processing a long time series (37 yr) of daily climate data for use by soil quality models.

1) *Trivariate thin-plate-spline smoothing (Hutchinson 1995, 1998a,b):* the spline interpolation of scattered data is to construct a thin plate that fits a field with minimum mean-square error and satisfies the constraint of continuous curvature. At the three-dimensional data points (x_i, y_i, z_i), the functional values of the thin plate are usually not equal to the observed data $f(x_i, y_i, z_i)$. The thin-plate-spline smoothing emphasizes the global shape of a field. Mathematically, the thin-plate-spline algorithm minimizes the mean-square error under the constraint of smooth curvature. The constraint is integrated into the cost functional by a Lagrange multiplier. Hutchinson (1995, 1998a,b, and references therein) has extensively explored meteorological applications of the method and used it to interpolate both daily and monthly rainfall fields. After a square root transformation of the precipitation data, the skewly distributed rainfall data became more normally distributed. He then interpolated the square root of the rainfall data and obtained a rainfall interpolation result with a small root-mean-square error (rmse).

   Bivariate (x, y) thin-plate-spline smoothing always results in a very smooth field, whose spatial variance is often too small to be realistic. Hutchison (1998b, and in his earlier work) noticed that it was important to include elevation z as an additional independent variable. He applied the trivariate thin-plate-spline method to various climate data (Hutchinson 1998b; Price et al. 2000). The gradients of the climate fields due to topographic variation were successfully recovered.

   Because the resulting field, even with the trivariate thin-plate-spline interpolation, is still very smooth over a flat region, the method is well suited for large-scale rainfalls, such as the daily rainfall in Southeast Asia during the monsoon season, or monthly and annual precipitation in most places of the world. Over the plains area, the elevation variable plays a very small role, and the smoothness implies smaller variance of the fitted field in comparison with the true one. Consequently, the small-scale storms are smeared, and thus it is to be investigated whether the method can be improved to accurately interpolate the highly localized, summer-convective storms (in daily time scale) in plains areas like the Canadian prairies. In conclusion, this method is most suited to interpolate a climate field of large spatial scales or over a mountainous region. Alberta's relatively flat, agricultural area has highly localized, convective precipitation during the summer, which is the ''wet'' season, when about 60% of the annual total precipitation falls in 4 months.
2) *Gradient plus inverse-distance-squared (Price et al. 2000; Nadler and Wein 1998)*: To interpolate climate data $[V(x_i, y_i, z_i), i = 1, 2, \ldots, N]$ onto points (x_p, y_p), one first takes the first-order Taylor expansion,

$$\hat{V}_p = V(x_i, y_i, z_i) + \nabla V(x_i, y_i, z_i) \cdot (x_p - x_i, y_p - y_i, z_p - z_i). \quad (1)$$

   Here (x_i, y_i, z_i) are the local orthogonal coordinates and z_i is elevation. The gradient $\nabla V(x_i, y_i, z_i)$ is approximated by linear regression coefficients. The accuracy of this linear approximation is assumed to be weighted by inverse-distance square (d_{ip}^{-2}). According to Price et al. (2000), although the result of this method is comparable to that of the thin-plate spline when applied to monthly data over a region where both topographic and climatic gradients are small, the trivariate thin-plate-spline smoothing systematically outperforms it. Hence, the method works well in Ontario, Quebec, Manitoba, Saskatchewan, and the agricultural areas of Alberta, but the thin-plate-spline smoothing clearly produced better results over British Columbia and the mountain areas of Alberta. The requirement of the stationarity of the regression coefficients, which approximate the climatic gradient, makes it difficult for the method to be applied to daily precipitation data, which often have a large gradient in the summer.
3) *Precipitation–elevation regressions on independent slopes model (PRISM; Daly et al. 1994)*: PRISM considers (i) the relationship between precipitation and elevation via a digital elevation model, (ii) the spatial scale of orographic effects, and (iii) the topographic characteristics of orographic regimes. The method is regarded as a useful approach for interpolating the monthly precipitation field in a mountain region. However, the factors (i) and (ii) are difficult to assess quantitatively, when a precipitation field is not stationary. Hence, the method's application to daily data is limited.
4) *Inverse-distance weighting (Jones et al. 1986)*: In their milestone paper, Jones et al. (1986) systematically interpolated the global station data onto $5° \times 5°$ grid points for the first time. Their method was the inverse-distance weighting. Despite further development of various types of interpolation methods in the last 14 yr, this seemingly crude method yielded a reliable result in terms of mean temperature. However, the resultant field is still too smooth and the variance at many grid points is too small.
5) *Kriging (Hudson and Wachkernagel 1994; Cressie 1993)*: Kriging is a commonly used method in geology. It minimizes the mean-square error between the estimated field and the true field, when the covariance field is known. The covariance field is de-

scribed by a variogram, which has three parameters to be fitted: variance, spatial length scale, and correlation between two points of a large distance. Recently, Hudson and Wachkernagel (1994) modified the method by explicitly including the elevation factor and applied the modified method to the January temperature in Scotland. Cross-validation shows that the correlation between the kriged data and the true data is around 0.9. However, kriging requires a field to be relatively stationary in time and homogeneous in space. These requirements make it a poor choice to apply to daily climatic data, particularly precipitation. This was also pointed out by Daly et al. (1994).

6) *Empirical orthogonal function method (Smith et al. 1998)*: Empirical orthogonal functions (EOF) are the eigenfunctions of the covariance function of a field. They can reflect inhomogeneous properties such as teleconnections. It has been shown that EOF has become the most effective tool in dealing with spatially inhomogeneous climate fields. Smith et al. (1998) used EOFs to interpolate the monthly sea surface temperature data over the tropical Pacific. They trained the EOFs by using the observed data from 1982 to 1995. The trained EOFs were then used for interpolation by minimizing the mean-square error. Thus, the EOF approach is a kriging for an inhomogeneous climate field. This method, however, may yield unreasonable results when the field is highly nonstationary, since there is no data to train the EOFs, which, by definition, are stationary. The daily climate data field is often highly nonstationary, and hence the EOF approach is not suitable for processing daily data.

In summary, most optimization methods are best for fitting mean conditions; the resulting fields are too smooth and do not adequately preserve the number of days with precipitation in a month or a year. Normally, an optimization method requires estimating parameters that are often assumed to be stationary. The nonstationary daily data often do not produce a robust estimate of parameters. For example, the covariance structure, required for kriging algorithms, for daily data is usually nonstationary and is hard to assess accurately. Simple interpolation methods, such as the nearest-station assignment, although not optimized, can often retain the variability of daily data. Retaining variance of climate fields is essential when putting climate data into soil quality models, particularly the erosion models. On the other hand, if the monthly or annual data are processed, the optimization methods are more accurate since the monthly and annual climate fields are close to being stationary. In this paper, a hybrid interpolation method is described that incorporates the nearest-station-assignment and inverse-distance-weighting methods.

b. The method of nearest-station assignment

To obtain the spatial average values of a climatic quantity over each polygon, a regular 10 km $\times$ 10 km grid was used to cover Alberta. Each grid point was assigned the observed value of the nearest station that had data for the day. The arithmetic average of the climate parameter values of all the grid points inside the polygon was the daily value of the climate parameter for the polygon. Thus, the method is called the ''nearest-station assignment.''

In general, a polygon had at least one grid point. Some small polygons had no regular grid points; thus the polygon centroid was selected as an additional interpolation (grid) point. Because the polygon was small when compared with the length scales of temperature and precipitation, this centroid represented the climate conditions over the entire small polygon, much as a station located there would do so. Two EDP and 116 SLC polygons have no regular grid points. Their centroids were used as the interpolation (grid) points. With the regular grid points and the centroids of the small polygons, the number of interpolation points was 6633 for EDPs and 6746 for SLCs over the entire area of Alberta. For any given day, every polygon could acquire its interpolated values of climate parameters by the nearest-station-assignment method.

This method assigns to a grid point the observed climate data directly from the nearest station. It should not yield a large bias when the observational stations are sufficiently dense. However, this method is by no means optimal since no computational optimization is implemented. When the observational stations are very sparse and the climate conditions are complex, this method will result in substantial spatial errors for a climate parameter that varies over short length scales.

Since the desired result is the spatial average of a parameter for a polygon, rather than the gridpoint values, this interpolation should be related to averaging. An averaging method used in geography is the Thiessen polygon approach, which considers only a station's representative area determined by the bisectors between each pair of stations. The implicit assumption is that the length scale of the observing network and the climate parameter are similar and interchangeable. This spatial averaging method is, in fact, equivalent to our interpolation from station data to polygon grid points, which uses the method of nearest-station assignment.

Canadian ecodistrict climate normals for 1961–90 prepared by Agriculture Canada were developed by using the Thiessen polygon method. More information is available on the Web site given in section 1. The nearest-station-assignment method is the same as the Thiessen polygon method, but the computing for the nearest-station-assignment method is much simpler than that of the Thiessen polygon method for daily data.

The procedure for the Thiessen polygon method is as follows. For a given day, one identifies all the stations

that have data. With these stations, Thiessen polygons are drawn to cover Alberta, a procedure requiring a great amount of computing work.

The equivalence between the nearest-station-assignment method and the Thiessen polygon method can be shown as follows. The EDP B intersects with M_B Thiessen polygons P_{Bi} ($i = 1, 2, \ldots, M_B$). The temperature over the EDP B is then determined by these M_B stations and their Thiessen polygons by the formula

$$T_B = \frac{1}{|B|} \sum_{i=1}^{M_B} w_{Bi} T_{Bi}, \tag{2}$$

where $|B|$ is the area of the EDP B, and w_{Bi} is the area of the intersection of B and the Thiessen polygon P_{Bi}. When using the grid points, the above formula can be approximated by

$$T_B = \frac{1}{N_B} \sum_{i=1}^{M_B} N_{Bi} T_{Bi}, \tag{3}$$

where N_B is the total number of grid points inside the EDP polygon B, and N_{Bi} is the number of grid points inside the intersection of B and the Thiessen polygon P_{Bi}. The accuracy of this approximation is proportional to the density of the grid points and also depends on the spatial length scales of the interpolated parameters. For the daily temperature and precipitation under our consideration and the 10 km $\times$ 10 km grid, the above two formulas approximate each other extremely well. The approximation may be incorrect for only some very small scale storms.

Three aspects of interpolation require discussion: the error, the variance of the results, and the computing algorithm. The first aspect is that the nearest-station-assignment method may contain some sizable errors, yet it does not generate a large estimation bias. If the station density is high, this method can obviously yield very accurate results.

The second aspect is variance. Since the nearest-station method uses only one station's data for a grid for a given day, the interpolated grid should adequately preserve the variance of a single point, although the nearest station may change from day to day. The interpolation to a grid point from every other method is a linear combination of the data from several stations. It is well known that a linear combination of data reduces variance, particularly the variance that measures noise.

The third aspect is the algorithm of finding the nearest stations every day. When considering regional climatology, researchers generally believe that the climate parameter over a polygon is represented by nearby stations. All the stations within and adjacent to each polygon should be identified. The nearest stations can be found in many ways, such as by the use of moving circular or rectangular windows (Isaacs and Srivastava 1989). After many computing experiments, we selected the following searching strategy. Each EDP has maximum and minimum latitude and longitude values that define a rectangle that encloses the polygon. An expansion of the rectangle on each side by 1.8° (about 200 km) in the north–south direction and 3.4° (also about 200 km) in the east–west direction forms a larger rectangle. This 200 km is the spatial length scale for the temperature field (Hansen and Lebedeff 1987). Thus, the values 1.8 and 3.4 are chosen to make the extensions in the north–south and east–west directions have about the same spatial length. The stations contained inside this larger rectangle are regarded as the subset of nearby stations for the polygon. The distances from each grid point inside the polygon to the stations inside the expanded rectangle are computed, and the results form a distance table. The distance sorting finally identifies the stations used for the interpolation to the particular grid point for each climate parameter.

c. Interpolation by inverse-distance weighting

This method is based upon the assumption that the influence of the nearby observed data on an interpolated point solely depends on the inverse of the distance between the interpolated point and the data point. Let $\hat{\mathbf{g}}_j$ be the interpolated point, T_i be the observed data at the station $\hat{\mathbf{r}}_i$, and $\hat{T}_j$ be the estimated value of the quantity T at the point $\hat{\mathbf{g}}_j$. Then the inverse-distance-weighting scheme is

$$\hat{T}_j = \left(\sum_{i=1}^{N} \frac{1}{d_{ij}} \right)^{-1} \sum_{i=1}^{M_j} \frac{T_i}{d_{ij}}, \tag{4}$$

where $d_{ij} = |\hat{\mathbf{r}}_i - \hat{\mathbf{g}}_j|$ is the distance between $\hat{\mathbf{r}}_i$ and $\hat{\mathbf{g}}_j$, and M_j is the total number of the stations "nearby" $\hat{\mathbf{g}}_j$. If the station $\hat{\mathbf{r}}_i$ is on the grid point $\hat{\mathbf{g}}_j$, then

$$\hat{T}_j = T_i. \tag{5}$$

The stations T_i (where $i = 1, 2, \ldots, M_j$) are chosen according to the distance table for the grid $\hat{\mathbf{g}}_j$. Station $\hat{\mathbf{r}}_1$ is the station with data that is nearest to the grid $\hat{\mathbf{g}}_j$, and station $\hat{\mathbf{r}}_2$ is the second nearest station. The eight nearest stations with $d_{ij} \leq 200$ km for temperature data and with $d_{ij} \leq 60$ km for precipitation data are chosen for interpolation. Here 60 and 200 km are approximate spatial correlation length scales of precipitation and temperature, respectively (Huff and Shipp 1969; Hansen and Lebedeff 1987; more discussion in the next two paragraphs). If less than eight stations are within the specified distance, then only the stations present are used for interpolation. For example, if six stations are within 200 km, then the interpolation for temperature uses only these six stations. In the northern part of Alberta, there may be no stations within the specified distance. Then, the nearest-station-assignment method is used for interpolation since the nearest-station-assignment method does not specify a distance; hence only one station is used. Figure 4 is the flow chart of the interpolation algorithm.

Also because the station distribution in northern Al-

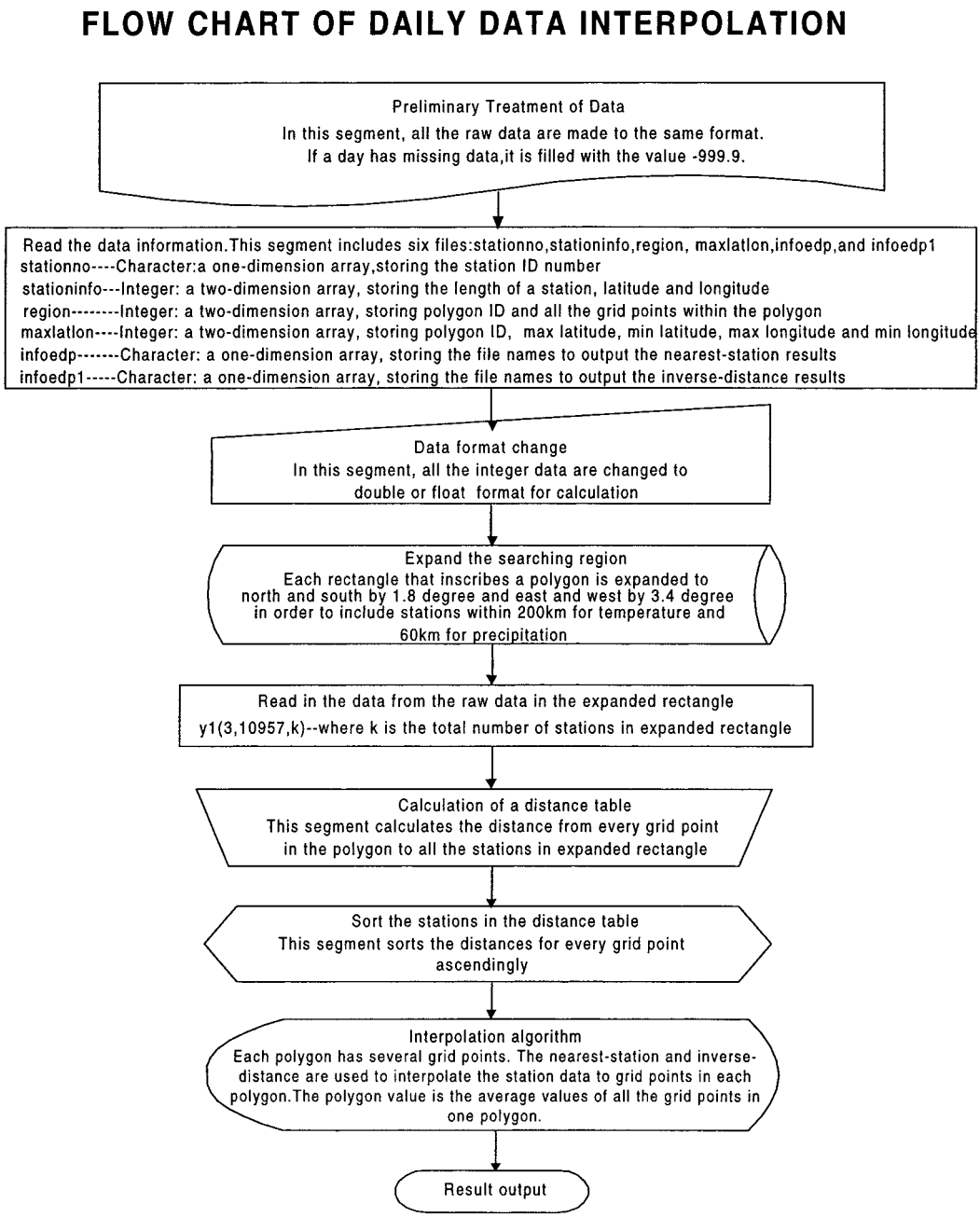

FIG. 4. Flow chart of the interpolation procedures.

berta is very sparse, the expanded searching rectangle for a polygon might not include any station. In this case, the rectangle associated with the polygon should be extended so that it includes at least four stations with data. No universally applicable rule states how many stations should be chosen. Here, we choose four stations, following the method used by McGinn et al. (1992). For temperature, because only the stations within 200-km distance from the grid point are used for interpolation, it can happen that for a given grid point inside the polygon, there is only one or no station within the specified distance; hence, only the nearest station is used for interpolation, and consequently, the inverse-distance-weighting interpolation is equivalent to the nearest-station-assignment method. Thus, in the data-sparse areas, the inverse-distance method reverts to the nearest-station method. However, our inverse-distance method is so flexible that inside the same polygon, some grid points use several stations and some use only one station, which is dynamically determined by the unique circumstances of polygon shape and station locations.

Our inverse-distance method is somewhat different

from the conventional ones described in most geostatistics books or those commonly used in the literature (Haining 1990; Isaaks and Srivastava 1989). The main differences involve (i) the station-searching method and (ii) the use of the nearest-station method if specified length scales are exceeded in regions with low station density. The spatial length scale is a measure of the coherence of a climate field. If two points are nearby, the correlation of their climate parameters is close to 1, and, in general, the correlation decreases as the distance between the two points increases. Usually, for a homogeneous field, when the correlation is equal to 0.4, the distance is defined as the length scale of the field. Although our highly inhomogeneous daily climate fields do not obey this decaying rule, the length scales are still a good reference for certain coherence. In this paper, 200 and 60 km are used as the length scales of temperature and precipitation, respectively. These length scales are approximate values and have large ranges of uncertainties. From the study of Huff and Shipp (1969) on the storm data in Illinois, 60 km is a reasonable value for precipitation length scale. The choice of this value should also consider the station density. If the station density is very large, one may choose 40 km to reduce the noise from distant stations. The length scale for temperature is estimated from the study of Hansen and Lebedeff (1987). They considered annual data. (Monthly data have similar results.) We obtained our length value by dividing theirs, 1200 km, by $6 \approx \sqrt{30}$, assuming that the daily temperature anomalies are independent from each other. The choice of the length scales has also been validated by examining numerous Alberta daily weather maps.

The use of faraway stations most likely introduces more noise when the inhomogeneous teleconnection patterns are not known. Our station searching method inherently adjusts to the station density. Our searching method and computational algorithm automatically exclude the more distant stations when there are stations closer to a grid point. Inclusion of a distant station's data, which do not represent the grid point, can only further distort the interpolation result from the true field and lead to oversmoothing.

The inverse-distance method (4) is a point estimation, but we need the spatial value for each polygon. We used the regular 10 km $\times$ 10 km grid, as in the previous section, to cover Alberta with at least one grid point in each polygon. The inverse-distance method is used for each grid point in a polygon. The climate parameter over the polygon B is the arithmetic average of the $\hat{T}_j$ for all the grid points $\hat{\mathbf{r}}_j$ inside B (see Mackey et al. 1996). Hence,

$$T_B = \frac{1}{N_B} \sum_{j=1}^{N_B} \hat{T}_j. \tag{6}$$

If a station is inside the polygon, then some grid points must be near the station, and, hence, this station's weight is very large. Thus, if the north–south and east–west dimensions of a polygon are about the same, then the climate values over the polygon are determined mainly by the station(s) inside the polygon. However, if a polygon is long and narrow, a station outside of but near the polygon may also contribute to the polygon data.

In the data-dense region, the field resulting from inverse-distance weighting is smoother than that from the nearest-station assignment, but the inverse-distance weighting might have oversmoothed the field and reduced extremes. The fields of the monthly precipitation and daily mean temperature may be smooth enough, and the inverse-distance weighting may yield reasonable results. However, for daily precipitation, the interpolation method often results in too many days with precipitation in a month, which raises the precipitation frequency of a polygon. For example, if even one of the polygon's nearest stations recorded nonzero precipitation for a given day, inverse-distance weighting will yield a nonzero precipitation for the polygon, even though all other nearest stations for this polygon may actually have recorded zero precipitation for the day. This can significantly increase the number of days with precipitation. For example, the 1961–90 June climatological mean of EDP727 has 24.5 days with precipitation using the data from the inverse-distance method, but the days with precipitation for EDP727 according to our scheme are only 13.7. The latter is more reasonable and very similar to the 1961–90 normal value from actual recorded data at nearby stations. See the cross-validation section below for more details.

The inverse-distance-square weighting, also called the power-2 inverse-distance weighting, follows the same computational procedures. Because of the higher power of the inverse distance, the field is more localized. Thus, if a polygon has stations inside, the climate parameter over the polygon is subject to little influence from the station data outside of the polygon. As compared with inverse-distance weighting, the inverse-distance-square weighting yields a less smooth climate field. In fact, as the power-n of the inverse-distance weighting approaches infinity, the power-n inverse-distance weighting becomes the nearest-station assignment.

d. Precipitation frequency and hybrid method

A polygon's precipitation frequency involves the polygon's number of days with precipitation in a month, the exact days of precipitation, and the amounts of precipitation (Osborn and Hulme 1997). Unfortunately, the inverse-distance method cannot correctly determine the precipitation frequency and yields too many days with precipitation and too little precipitation per day for a grid point, and hence for a polygon, as compared with the observed data. The results from the inverse-distance method have smaller variance both in space and time. Since the rmse, mean absolute error (mae), and mean biased error (mbe), defined by formulas (9)–(11) in sec-

tion 4, are mean properties of differences between the observed data and the interpolated data, it is not surprising that the inverse-distance method yields more accurate results than the nearest-station method when these measures of error are used.

Because the results from the inverse-distance method fit the monthly mean very well, the monthly total computed from the method is accurate. Our cross-validation for monthly totals for five stations supports this conclusion. Thus, the monthly totals for a grid point, and hence a polygon, are computed from the inverse-distance method.

One way to estimate precipitation frequency is to have stations everywhere in a polygon. Of course, doing so is impossible and even then would not resolve the situation for larger polygons where inevitably some stations would record precipitation and others would not on a given day. Therefore, we use the precipitation of a polygon's centroid to define whether the polygon had precipitation on a given day. If the centroid has precipitation, then we define that the polygon also has precipitation that day. The centroid of a polygon is rarely the location of a station. We use the centroid's nearest station as the best indicator for the centroid's precipitation, and hence the polygon's precipitation and also as the indicator of which days have precipitation. The monthly total precipitation of a polygon is the sum of the daily polygon precipitation determined by the inverse-distance method. The precipitation frequency is computed from the nearest-station method. For this reason, our method is called a ''hybrid method.'' For a given day t, the precipitation over polygon is computed from the following hybrid formula:

$$\mathrm{Pcpn}_{\mathrm{edp}}(t) = \mathrm{Pcpn}_{\mathrm{iedp}}(m) \times \frac{\mathrm{Pcpn}_{\mathrm{center}}(t)}{\mathrm{Pcpn}_{\mathrm{center}}(m)}, \quad (7)$$

where $\mathrm{Pcpn}_{\mathrm{iedp}}(m)$ is the monthly total precipitation of the EDP and is computed by the inverse-distance method, $\mathrm{Pcpn}_{\mathrm{center}}(m)$ is the monthly total precipitation of the station(s) nearest to the centroid, and $\mathrm{Pcpn}_{\mathrm{center}}(t)$ is the precipitation of the station nearest to the centroid for the given day t.

Please note that

$$\sum_{t=1}^{M} \mathrm{Pcpn}_{\mathrm{edp}}(t) = \mathrm{Pcpn}_{\mathrm{iedp}}(m). \quad (8)$$

Namely, the monthly total precipitation for a polygon is not changed after using this precipitation frequency formula (7). The hybrid formula (7) combines the inverse-distance and nearest-station methods. It is empirical and data driven and requires cross-validation, which is given in the next section.

4. Cross-validation and accuracy of interpolation

The most effective method now commonly used to assess the error of climate data estimation is cross-validation (Cressie 1993). The procedure compares estimated data for a point with observed station data at that point. Of course, the station data is withheld from the estimation. The data from other stations are interpolated to the station location. The statistics for the difference, or errors, between the true data and the interpolated data are used to evaluate the interpolation scheme's accuracy.

To evaluate the interpolation accuracy, three types of errors were computed.

1) Root-mean-square error:

$$\mathrm{rmse} = \left\{\frac{1}{K}\sum_{t=1}^{K}[X_{\mathrm{true}}(t) - X_{\mathrm{estimate}}(t)]^2\right\}^{1/2}, \quad (9)$$

where K is the number of days used for cross-validation studies, t is time with units of 1 day, and X denotes a climate parameter at a cross-validation location.

2) Mean absolute error:

$$\mathrm{mae} = \frac{1}{K}\sum_{t=1}^{K}|X_{\mathrm{true}}(t) - X_{\mathrm{estimate}}(t)|. \quad (10)$$

3) Mean biased error:

$$\mathrm{mbe} = \frac{1}{K}\sum_{t=1}^{K}[X_{\mathrm{true}}(t) - X_{\mathrm{estimate}}(t)]. \quad (11)$$

Since the polygon values are obtained from gridpoint values, the cross-validation is performed for both grid points and polygons. For gridpoint cross-validation, five long-term stations distributed from south to north are considered. They are, sorted from south to north, Lethbridge CDA (3033890; 49°42′N, 112°47′W), Lacombe CDA (3023720; 52°28′N, 113°45′W), Edmonton INTL A (3012205; 53°18′N, 113°35′W), Beaverlodge CDA (3070560; 55°12′N, 119°24′W), and High Level A (3073146; 58°37′N, 117°10′W). The total number of days for cross-validation is 13 514. Hence, the total number of data entries for cross-validation is 67 570 minus the days without data at the cross-validation stations. Since day-to-day temperature and precipitation anomalies are normally independent of each other, the data for each day may be considered as an independent sample.

The rmse, mae, and mbe results for Edmonton, Lacombe, Lethbridge, and Beaverlodge are comparable. The magnitude of the errors for the five stations is shown in Table 2.

For the inverse-distance method and for all stations except High Level A, the rmse for T_{max} ranges from 1.37° to 3.19°C, T_{min} from 1.79° to 3.22°C, and Pcpn from 1.75 to 2.84 mm. For the nearest-station-assignment method again excluding High Level A, the rmse for T_{max} ranges from 1.97° to 2.91°C, T_{min} from 2.48° to 3.72°C, and Pcpn from 2.39 to 3.30 mm.

The errors are small for the Lacombe, Edmonton, and Beaverlodge stations, since these areas are flat and have higher station density. The station density in the Leth-

TABLE 2. Errors assessed by cross-validation for five long-term stations from south to north (Units: T_{max} and T_{min} in degrees Celsius, Pcpn in millimeters). (First number following section name is station ID, second numbers are lat and lon in degrees and minutes with symbols and spacing omitted.)

		T_{max}	T_{min}	Pcpn
Lethbridge [3033890, (4942, 11247)]				
Inverse-distance method	rmse	2.63	2.99	2.00
	mae	2.25	2.46	0.60
	mbe	1.86	1.89	−0.05
Nearest-station method	rmse	1.97	2.69	2.97
	mae	1.30	1.89	0.87
	mbe	−0.15	−0.05	−0.04
Lacombe [3023720, (5228, 11345)]				
Inverse-distance method	rmse	1.37	1.79	1.75
	mae	0.96	1.38	0.64
	mbe	−0.11	−0.53	−0.16
Nearest-station method	rmse	2.10	2.48	2.39
	mae	1.40	1.83	0.78
	mbe	−0.25	−0.66	−0.10
Edmonton [3012205, (5318, 11335)]				
Inverse-distance method	rmse	1.75	2.16	2.36
	mae	1.10	1.57	0.81
	mbe	−0.07	−0.67	−0.03
Nearest-station method	rmse	2.10	2.83	3.30
	mae	1.40	2.10	1.08
	mbe	−0.15	−0.79	0.02
Beaverlodge [3070560, (5512, 11924)]				
Inverse-distance method	rmse	2.09	2.63	1.95
	mae	1.50	1.93	0.76
	mbe	0.63	1.01	−0.15
Nearest-station method	rmse	2.91	3.72	2.48
	mae	2.13	2.71	0.88
	mbe	0.84	1.04	−0.11
High Level A [3073146, (5837, 11710)]				
Inverse-distance method	rmse	3.19	3.22	2.84
	mae	2.21	2.41	0.74
	mbe	0.23	−1.12	−0.20
Nearest-station method	rmse	5.30	4.75	3.31
	mae	4.05	3.32	1.15
	mbe	1.88	−0.24	0.21

bridge area is also higher, but the topographic influence makes the errors slightly higher than Lacombe and Edmonton. The errors are larger for the northernmost station, High Level A. The station density in this area is much lower. The mean station distance, defined by the sum of the mutual distances between any two points divided by the total number of distances, is about 105 km. Also the data stream of this station is short, less than 10 yr as compared with others of 37 yr. Thus, the cross-validation errors for this station are not considered representative, and the large errors of the nearest-station-assignment method for this station are not included in the error summary of the above paragraph.

The rmse, mae, and mbe above are considered measures of the goodness of fit to mean conditions. Our computational results show that the error for T_{max} is usually smaller than that for T_{min}. The nearest-station-assignment method usually produces rmse and mae around 20%–30% larger than the inverse-distance method. This result is expected, since the inverse-distance method yields a smooth field and the true daily weather distributes randomly on the positive or negative side of the smooth field. Thus, the field generated by the inverse-distance method has a smaller variance than the true field.

We also checked other algorithms such as the inverse-distance-square method, quadrant search, and different length scales for precipitation, and found the following.

1) The inverse-distance-square method consistently yielded a larger error than the inverse-distance method.
2) The errors of the quadrant inverse-distance method were slightly but consistently larger than those from the inverse-distance method. The quadrant searched required at least a station in each quadrant. If a quadrant did not have a station within the distance range (200 km for temperature and 60 km for precipitation), then the nearest-station-assignment method was applied to the nearest station in the quadrant.
3) For precipitation, the inverse-distance method yielded smaller errors than the nearest-station method, and 100- and 60-km length scales did not make much difference, with 100 km giving a slightly smaller error for the High Level A station. Here, the two distances, 60 and 100 km, were considered to test the sensitivity of the results to the length scale. These two distances were combined with the four methods: nearest-station assignment, inverse distance, inverse-distance square, and quadrant search. Hence, seven, not eight, cases were evaluated since the nearest-station method is irrelevant to the length scale in computing algorithm.

The seven experiments were (temp is temperature)

a) nearest-station method (temp: ≥50 km, Pcpn: ≥5 km),
b) inverse-distance method (temp: 50–200 km, Pcpn: 5–60 km),
c) inverse-distance method (temp: 50–200 km, Pcpn: 5–100 km),
d) inverse-distance-square method (temp: 50–200 km, Pcpn: 5–60 km),
e) inverse-distance-square method (temp: 50–200 km, Pcpn: 5–100 km),
f) inverse-distance-quadrant method (temp: 50–200 km, Pcpn: 5–60 km), and
g) inverse-distance-quadrant method (temp: 50–200 km, Pcpn: 5–100 km).

The advanced, optimized interpolation methods, such as interpolation by EOF, cannot be used here either, because of nonstationarity (Shen et al. 1994; Smith et al. 1998).

In order for the cross-validation errors to be representative, the cross-validation should not include stations too close to the cross-validation site, since not all the grid points have stations nearby. Thus, in the above,

TABLE 3. Sample variances of the data from cross-validation stations (Units: °C² for T_{max} and T_{min}, mm² for Pcpn).

Station name	Observed data			Inverse-distance			Nearest-station		
	T_{max}	T_{min}	Pcpn	T_{max}	T_{min}	Pcpn	T_{max}	T_{min}	Pcpn
Edmonton Intl A	7.14	6.37	3.96	7.06	6.17	3.43	7.08	6.48	3.94
Lacombe CDA	7.24	6.10	3.76	7.20	5.91	3.53	7.42	6.22	4.05
Lethbridge CDA	7.65	6.71	3.79	7.06	6.17	3.29	7.70	6.49	3.79
Beaverlodge CDA	7.45	6.68	3.97	7.35	6.85	3.54	7.84	7.61	4.05
High Level A	7.08	6.77	3.31	7.16	6.74	3.82	7.70	7.35	3.57

the 50 km for temperature means that the stations less than 50 km away from the cross-validation site are not used for interpolation. This value for precipitation is 5 km, since the length scale for precipitation is much smaller.

Three types of errors, rmse, mae, and mbe, were calculated. The results of the above experiments (a) and (b) are shown in Table 2. They indicated that the inverse-distance method appears to generate more accurate results. The inverse-distance method oversmoothed the interpolated fields, particularly the precipitation field. We computed the sample variances for the data of the five cross-validation stations, and the variance results are shown in Table 3. Here, the variances are computed from the daily anomaly data. For each day in a year, the 1961–90 mean is computed. The anomalies for a station are with respect to this mean. The variance of the station is computed according to these anomaly data for 1961–90. Table 3 indicates that the inverse-distance method reduces the sample variance. For temperature, the reduction is small. The average of the five stations is less than 5%. For precipitation, the reduction is over 10%. (The results from the High Level A station were not representative and hence excluded in the above discussion because of the station's short record, namely 1072 days during the 30-yr cross-validation period, 1961–90.)

The variance for the precipitation resulting from the nearest-station method is almost the same as that of the observed data. Thus, considering the need to preserve the second moment, that is, variance, the nearest-station method was selected as the preferred interpolation method. This is the reason why the precipitation frequency was computed by the hybrid formula (7) in section 3d.

Let us consider the number of precipitation days per month for the five cross-validation stations. For each cross-validation, three datasets exist: the observed data at the station, the interpolated data from the inverse-distance method, and the interpolated data from the nearest-station method. The cross-validation results for the Lacombe (3023720) station are shown in Table 4. The number of days with precipitation from the observed data and that from the nearest-station interpolation are about the same, while that from the inverse-distance interpolation is too large by about 50%–100%. Such a big percentage is unexpected, although it is not surprising that the inverse-distance method yields too many precipitation days. Cross-validation results from other stations support the same conclusion.

We also validated the hybrid method on the five EDP and five SLC polygons in which the five cross-validation stations are located. The inverse-distance method yielded a result of daily precipitation and formula (7) revised the result. The revised result had a larger variance than the one generated by the inverse-distance method. Table 5 shows the variance of the precipitation for the five polygons, before and after the revision. The variance of the revised precipitation is about 10%–20% higher than that of the inverse-distance results, which is the size of increase we intended to achieve. The revision does not change the monthly total precipitation, but it changes the temporal distribution and hence the amount of daily precipitation. The revised precipitation overcomes the problem of the oversmoothed inverse-distance results, which have too many precipitation days and too little precipitation each day. Table 6 shows the precipitation

TABLE 4. Number of days with precipitation per month at Lacombe Station [3023720 (52°28′N, 113°45′W)].

Month	Observed	Inverse-distance	Nearest-station
1	9.30	16.13	8.92
2	7.13	12.33	7.07
3	6.83	13.90	6.69
4	6.87	12.80	5.68
5	10.33	16.87	10.19
6	13.53	20.43	13.92
7	14.20	21.23	14.22
8	12.47	18.97	12.61
9	11.00	16.73	11.31
10	5.80	12.00	5.31
11	7.03	12.97	6.33
12	7.60	14.50	7.62

TABLE 5. Precipitation variances of the inverse-distance results and revised results over the cross-validation polygons (Units: mm²).

Polygon	Revised	Inverse
EDP727	3.82	3.25
EDP737	4.10	3.41
EDP793	3.65	3.18
EDP598	4.09	3.49
EDP586	3.63	2.82
SLC433	4.04	3.41
SLC518	3.98	3.50
SLC644	3.67	3.22
SLC15	4.06	3.48
SLC723	3.63	3.24

TABLE 6. Number of precipitation days on two polygons computed from inverse-distance and hybrid methods.

	SLC518		EDP793	
Month	Inverse-distance	Hybrid	Inverse-distance	Hybrid
1	18.03	8.60	19.03	7.73
2	14.03	6.83	15.43	5.73
3	15.37	6.19	18.53	7.40
4	14.23	5.93	18.77	6.93
5	18.57	9.33	20.43	9.10
6	21.93	12.67	21.27	9.43
7	23.10	13.40	19.90	7.37
8	21.00	11.70	19.13	7.43
9	18.10	9.83	16.63	7.07
10	13.20	4.87	14.03	4.63
11	14.57	6.57	14.97	5.47
12	16.63	7.20	18.70	7.83

days per month in two of the 10 cross-validation polygons: SLC518 and EDP793. The precipitation frequency results produced by the hybrid method for polygons are comparable to those for cross-validation stations (Table 4).

The hybrid method can also preserve the spatial localization of precipitation, while the inverse-distance method and other smoothing methods spread precipitation domains. The localization is particularly important in summer and significant for the climate input of soil quality models, since water erosion of soil is mainly due to extreme storm events. Figures 5 and 6 show precipitation fields in millimeters for a major storm and for scattered small storms on two given days interpolated by the hybrid method.

The remaining cross-validation question is the goodness of fit for polygons with respect to mean conditions. Since the true value of a polygon average can never be measured, cross-validation experiments cannot be used to directly assess the errors of the polygon data. A rough estimate of the data error of a polygon is given by the above gridpoint mean square error divided by $\sqrt{n}$, where n is the spatial degrees of freedom of the climate field over a polygon. This is 1.0 for a small polygon, 2.0 for a large polygon of two independent grid points, and 3.0 for an even larger polygon of three independent grid points. However, this is only a rough estimate and knowing exactly how many independent grid points are within a polygon is not a trivial task. It can be safely claimed that the upper limit of the polygon data error is 1.8°–3.2°C for temperature and 1.8–2.4 mm for precipitation, and the lower limit is one-half of these amounts.

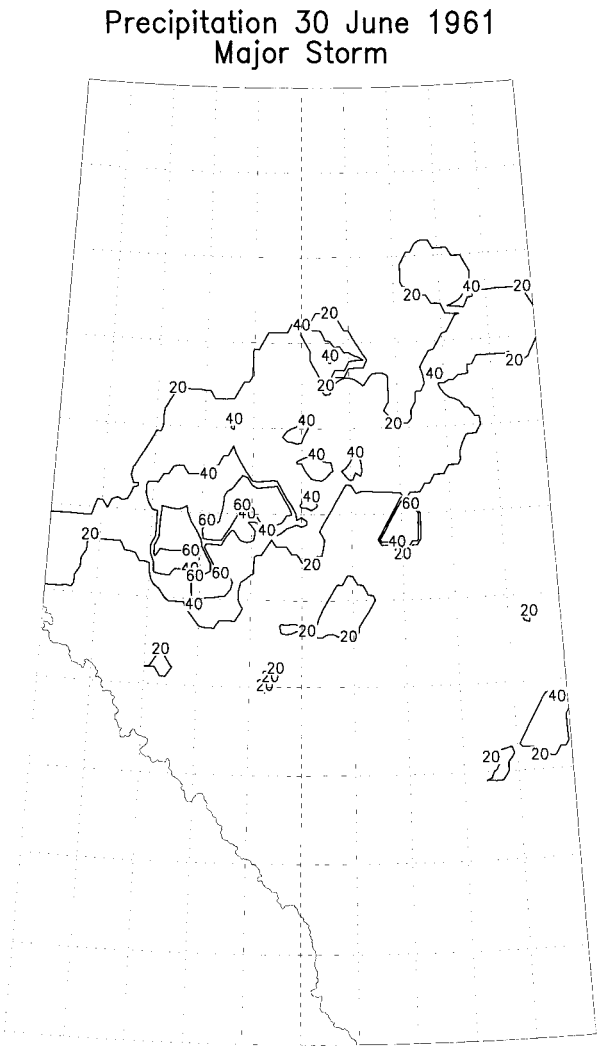

FIG. 5. Precipitation as interpolated by the hybrid method for a major storm on 30 Jun 1961.

5. Conclusions and discussion

We have described a method that interpolates daily station data onto the 894 SLC polygons and 150 EDP in the province of Alberta, Canada. The interpolated daily data fit not only the climate mean but also climate variability, in particular, the number of days with precipitation per month. Hence, the result is a complete 37-yr set of continuous daily data, which provides realistic climate input for use in soil quality models. The interpolated results, the first ever provided for Alberta at this fine scale for the entire province, preserve the variability of the Alberta daily climate data, specifically the number of days with precipitation. Our procedure of interpolating station data onto polygons involves interpolating the station data onto a dense network of grid points and then averaging the gridpoint values inside polygons. Special attention was paid to precipitation because of its short spatial correlation length scale and large variability. A new hybrid method combining inverse-distance weight and nearest-station assignment was developed for interpolating the precipitation data. The polygon's total monthly precipitation was obtained by the inverse-distance-weight method, while the number of days with precipitation and what days had pre-

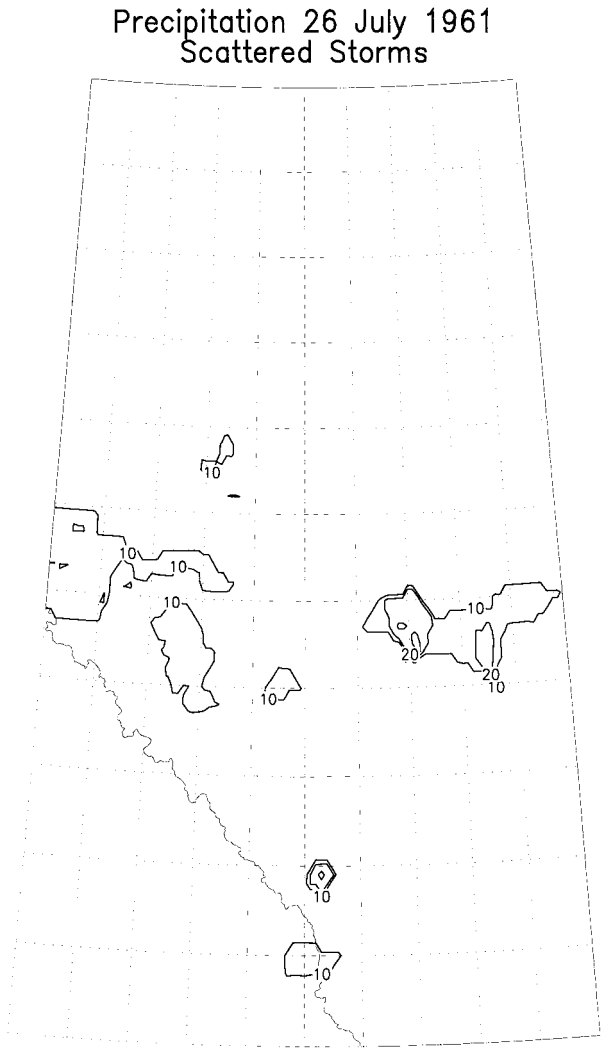

FIG. 6. Precipitation as interpolated by the hybrid method for small scattered storms on 26 Jul 1961.

cipitation in the month were determined by the nearest-station-assignment method. This hybrid method resolves a long-outstanding problem in precipitation interpolation: too many days with precipitation and too little precipitation per day.

Maximum and minimum temperature have long spatial correlation length scale and are recorded at almost all stations. Although interpolation provides good results, it is important to avoid a method that oversmooths the field. It is equally important that one not use a method, such as nearest-station assignment or inverse-distance square, which strongly localizes the temperature field so that the interpolated fields are discontinuous or bumpy. We have chosen to use the inverse-distance-weight method, which well maintains the continuity and variability of the field.

Soil quality models also require the interpolated data for relative humidity, incoming radiation, wind speed, and wind direction. However, there are very few stations observing these parameters. Within Alberta and during the period of 1 January 1961–31 December 1990, there were only 26 stations for relative humidity, 45 stations for wind, and 5 stations for incoming radiation.

Relative humidity and incoming radiation have strong dependence on clouds and hence are highly localized or discontinuous in space. After testing the inverse-distance-weight method, we have found that the results of the nearest-station-assignment method better reflect the true climate conditions over polygons.

The wind field can be tricky and strongly depends on orographic properties. In a large area, the mean wind direction of a day may be nearly homogeneous, but two points nearby can have completely different directions because of geographic characteristics. In the AES dataset, wind speed, and wind direction have few observation stations. We also used the nearest-station-assignment method for the wind field in order to maintain both its temporal and spatial variability. The interpolation scheme is exactly the same as that of relative humidity and radiation and its details have been omitted in this paper. It is understood that dynamic properties should be incorporated to construct a more realistic wind field, which is beyond the scope of this paper.

It is worth noting that some existing interpolation methods do not make full use of all the observed data. Many interpolation methods require that the data stream from a station be complete. If there are missing data, a temporal interpolation is applied to estimate the missing data. Because of large daily variability of climate parameters, the temporal interpolation can introduce errors into the climate data. Thus, if a station has a very short record, then this station must be excluded from interpolation. Our computing algorithm does not require this temporal interpolation and uses all the daily observed climate data, even using a very short record such as summer-only records, as long as the data have passed the quality control process.

Still other spatial interpolation methods are also available, such as ordinary kriging, block kriging, the empirical orthogonal function approach, and improved kriging, but they require that the climate process be stationary or close to stationary. The monthly or annual means for climate parameters can be approximately stationary, but not the daily values. Some methods such as ordinary kriging (Hudson and Wackernagel 1994; Isaaks and Srivastava 1989) cannot even reliably account for spatial inhomogeneity. The daily weather has large variances and is not a stationary process. It does not have a stable variogram. An assumed variogram, such as a Gaussian type, is certainly far away from truth. Consequently, the kriging results can be very unrealistic. This conclusion agrees with that by Daly et al. (1994). Of course, a more careful selection and fitting of the variogram should improve the result. However, we are still not optimistic about ordinary kriging's application to daily weather data interpolation. For ex-

ample, Higgins et al. (1996), after a test of kriging and other methods, used the inverse-distance method to interpolate U.S. hourly precipitation data.

Any method that takes account of elevation may have great potential in application (Daly et al. 1994; DeGaetano et al. 1993; Dodson and Marks 1997; Hutchinson 1998b). Since the major aim of our project was to develop a dataset for soil quality models in agricultural applications in Alberta, the mountain areas were not considered, and hence elevation-related methods were not explored in this study. Nevertheless, these methods have potential to improve upon the work reported here. One possibility of improvement is the combination of thin-plate-spline smoothing with nearest-station assignment, similar to that done in section 3d. Suppose that thin-plate-spline smoothing can reliably interpolate the mean condition of a climate field. The nearest-station assignment can raise the variance of the result from the thin-plate-spline smoothing and reduce the number of precipitation days in a month. This is deferred to future investigation.

Our procedure interpolated the scattered station data onto a regular grid which could easily be latitude–longitude grid points and suit many applications. Similar to the 10 km $\times$ 10 km resolution, our method can provide a dataset with $0.1^\circ \times 0.1^\circ$ latitude–longitude resolution. One application is a high-resolution version of *Agroclimatic Atlas of Alberta* (Dzikowski and Heywood 1990). With the gridpoint values, the atlas can be easily and accurately generated by software such as the Geographic Information Systems that are used widely in geography and geology, or the Grid Analysis and Display Systems commonly used by meteorologists. Both software require data on regular latitude–longitude grid points. Another application is to use the data for a long-term statistical forecasting. The canonical correlation analysis or singular value decomposition method can use this grid data for a seasonal, 6-month, or 12-month forecasting. These projects will again be deferred to future studies.

The format of the climatic data input for soil quality models is referred to in Shen et al. (2000). Other basic daily climate parameters needed for the models, such as degree-days, are derived from the seven climate parameters: maximum temperature, minimum temperature, precipitation, relative humidity, incoming solar radiation, wind speed, and wind direction.

Acknowledgments. This work was financially supported by Alberta Agriculture, Food and Rural Development, Atmospheric Environment Service, and Natural Sciences and Engineering Research Council. Shane Chetner and Anthony Arendt provided the climate data in a suitable format with database and format documentation. Figures 1–3 were prepared by David Spiess. The authors are grateful to Karen Cannon, Tom Goddard, and David Spiess for helpful discussions and thank the anonymous reviewers for constructive suggestions, which led to a substantial improvement of this paper.

REFERENCES

Changnon, S. A., and K. E. Kunkel, 1999: Rapidly expanding uses of climate data and information in agriculture and water resources: Causes and characteristics of new applications. *Bull. Amer. Meteor. Soc.,* **80,** 821–830.

Cressie, N., 1993: *Statistics for Spatial Data*. John Wiley and Sons, 900 pp.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological prediction over mountainous terrain. *J. Appl. Meteor.,* **33,** 140–158.

DeGaetano, A. T., K. L. Eggleston, and W. W. Knapp, 1993: A method to produce serially complete daily maximum and minimum temperature data for the Northeast. Northeast Regional Climate Center Rep. RR 93-2, 34 pp.

Dodson, R., and D. Marks, 1997: Daily air temperature interpolated at high spatial resolution over a large mountainous region. *Climate Res.,* **8,** 1–20.

Dzikowski, P., and R. Heywood, 1990: *Agroclimatic Atlas of Alberta*. Alberta Agriculture, Food and Rural Development, 31 pp.

Ecological Stratification Working Group, 1995: A national ecological framework for Canada. Agriculture and Agri-Food Canada, Research Branch, Centre for Land and Biological Resources Research and Environment Canada, State of the Environment Directorate, Ecozone Analysis Branch, Ottawa/Hull Rep. A42-65, 125 pp.

Flanagan, D. C., and S. J. Livingston, 1995: USDA water erosion prediction project: Version 95.7 user summary. NSERL Rep. 11, 141 pp.

Haining, R., 1990: *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge University Press, 431 pp.

Hansen, J., and S. Lebedeff, 1987: Global trends of measured surface air temperature, *J. Geophys. Res.,* **92,** 13 345–13 372.

Higgins, W., J. E. Janowiak, and Y.-P. Yao, 1996: A gridded hourly precipitation database for the United States (1963–1993), *NCEP–Climate Prediction Center Atlas*, No. 1, National Centers for Environmental Prediction, 42 pp.

Hudson, G., and H. Wackernagel, 1994: Mapping temperature using kriging with external drift: Theory and example from Scotland, *Int. J. Climatol.,* **14,** 77–91.

Huff, F. A., and W. L. Shipp, 1969: Spatial correlations of storms, monthly and seasonal precipitation. *J. Appl. Meteor.,* **8,** 542–550.

Hutchinson, M. F., 1995: Interpolating mean rainfall using thin plate smoothing splines. *Int. J. Geogr. Inf. Syst.,* **9,** 385–403.

——, 1998a: Interpolation of rainfall data with thin plate smoothing splines. Part I: Two-dimensional smoothing of data with short range correlation. *J. Geogr. Inf. Decis. Anal.,* **2,** 152–167.

——, 1998b: Interpolation of rainfall data with thin plate smoothing splines. Part II: Analysis of topographic dependence. *J. Geogr. Inf. Decis. Anal.,* **2,** 168–185.

Isaaks, E. H., and R. M. Srivastava, 1989: *An Introduction to Applied Geostatistics*. Oxford University Press, 561 pp.

Jones, P. D., S. C. B. Raper, R. S. Bradley, H. F. Diaz, P. M. Kelly, and T. M. L. Wigley, 1986: Northern Hemisphere surface air temperature variations: 1851–1984. *J. Climate Appl. Meteor.,* **25,** 161–179.

Karl, T. R., and R. W. Knight, 1998: Secular trends of precipitation amount, frequency, and intensity in the United States. *Bull. Amer. Meteor. Soc.,* **79,** 231–241.

Mackey, B. G., D. W. McKinney, Y.-Q. Yang, J. P. McMahon, and M. F. Hutchinson, 1996: Site regions revisited: A climatic analysis on Hill's site regions for the province of Ontario using a parametric method. *Can. J. For. Res.,* **26,** 333–354.

McGinn, S. M., O. O. Akinremi, and A. G. Barr, 1992: Description of the Gridded Prairie Climate Database (GRIPCD) for years 1960 to 1989. Alberta Agriculture, Food and Rural Development Res. Rep. 1992-06, 20 pp.

Nadler, I. A., R. W. Wein, 1998: Spatial interpolation of climatic normals: Test of a new method in the Canadian boreal forest. *Agric. For. Meteor.,* **92,** 211–225.

Osborn, T. J., and M. Hulme, 1997: Development of a relationship between station and grid-box rainday frequency for climate model evaluation. *J. Climate,* **10,** 1885–1908.

Price, D. T., D. W. McKenney, I. A. Nadler, M. F. Hutchinson, and J. L. Kesteven, 2000: A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. *Agric. For. Meteor.,* **101,** 81–94.

Sharpley, A. N., and J. R. Williams, Eds.1990: EPIC: Erosion/Productivity Impact Calculator. Part 1. Model documentation. U.S. Department of Agriculture Technical Bulletin Rep. 1768, 235 pp.

Shen, S. S. P., 1998: Methods of spatial interpolation of climatic data onto ecodistrict and SLC polygons. Alberta Agriculture, Food and Rural Development Res. Rep. 1998-01, 24 pp.

——, G. R. North, and K.-Y. Kim, 1994: Spectral approach to optimal estimation of the global average temperature. *J. Climate,* **7,** 1999–2007.

——, K. Cannon, and G. Li, 2000: Alberta 1961–1997 climate data on EDP and SLC polygons: Data derivatives and data formation for soil quality models. Alberta Agriculture, Food and Rural Development Res. Rep. 2002-02, 31 pp.

Shields, J. A., C. Tarnocai, K. W. G. Valentine, and K. B. MacDonald, 1991: Soil landscapes of Canada: Procedures manual and user's handbook. Land Resource Research Centre, Research Branch, Agriculture Canada Publ. 1868/E, 74 pp.

Smith, T. M., R. E. Livezey, and S. S. P. Shen, 1998: An improved method for interpolating sparse and irregularly distributed data onto a regular grid. *J. Climate,* **11,** 2340–2350.