ELSEVIER

Environmental Modelling & Software 16 (2001) 309–330

Environmental
Modelling & Software

www.elsevier.com/locate/envsoft

Using spatial interpolation to construct a comprehensive archive of Australian climate data

Stephen J. Jeffrey *, John O. Carter, Keith B. Moodie, Alan R. Beswick

Queensland Centre for Climate Applications, Department of Natural Resources, Indooroopilly, Qld. 4068, Australia

Received 14 September 2000; received in revised form 6 December 2000; accepted 22 January 2001

Abstract

A comprehensive archive of Australian rainfall and climate data has been constructed from ground-based observational data. Continuous, daily time step records have been constructed using spatial interpolation algorithms to estimate missing data. Datasets have been constructed for daily rainfall, maximum and minimum temperatures, evaporation, solar radiation and vapour pressure. Datasets are available for approximately 4600 locations across Australia, commencing in 1890 for rainfall and 1957 for climate variables. The datasets can be accessed on the Internet at http://www.dnr.qld.gov.au/silo. Interpolated surfaces have been computed on a regular 0.05° grid extending from latitude 10°S to 44°S and longitude 112°E to 154°E. A thin plate smoothing spline was used to interpolate daily climate variables, and ordinary kriging was used to interpolate daily and monthly rainfall. Independent cross validation has been used to analyse the temporal and spatial error of the interpolated data. An Internet based facility has been developed which allows database clients to interrogate the gridded surfaces at any desired location.

Keywords: Climate; Precipitation; Database; Ground-based data; Interpolation

1. Introduction

A complete and accurate source of rainfall and climate data is a prerequisite for the efficient modelling of a wide variety of environmental processes. While the nature of the individual model may vary, most have the fundamental requirement of a dataset that is complete on a temporal and/or spatial basis. To date, this problem has restricted the research efforts of many workers due to the fact that observational records are typically incomplete, making it difficult to construct a continuous climate record. In particular, such data may: (1) be recorded for discrete periods, not spanning the entire time period of interest; (2) contain short intermittent periods where data have not been recorded; and (3) contain either systematic or random errors (Peck, 1997). Furthermore, erroneous data must be removed once detected, which is a difficult problem in itself. While these points focus on the incomplete temporal aspects of observational data, another inherent problem is the spatial distribution of recording stations. The density of stations in observational networks is of particular interest to those who use models which require point data. In many applications, the success or at least accuracy of point simulations can be critically dependent upon the availability of observational data within an acceptable distance of the location under investigation. Ideally the nearest recording station would be situated such that its climatology was identical to that of the location of interest. However, due to the sparsity of observational networks, the distance to the nearest station can be of the order of several hundred kilometres. As a result, the only available data may not be representative of the climatology at the desired location.

The fact that observational data are both spatially and temporally incomplete has led to the widespread interest in remotely sensed data. While in principle such data have the potential to overcome the deficiencies of observational data, its implementation has been severely limited by a number of factors. In particular, satellite data are unavailable for dates prior to the 1970s and, more importantly, remotely sensed information is in most cases an indirect measure of climatic elements. In order

* Corresponding author. Fax: +61-7-3896-9843.
E-mail address: stephen.jeffrey@dnr.qld.gov.au (S.J. Jeffrey).

1364-8152/01/$ - see front matter
PII: S1364-8152(01)00008-1

to derive quantitative information from remotely sensed data, one must first construct acceptable calibration models to transform the measured signal. Given this already difficult task, one may also be required to extract the desired component from a number of background effects. While calibration models are rapidly improving, and remotely sensed data are becoming widely available, ground-based observational data remain the preferred source of meteorological data.

The reconstruction of serially incomplete data records has been the subject of a number of review articles (see for example Lam, 1983; Bennett et al., 1984). An associated problem which has received less attention is that of record length. The duration of available data records may impact upon the quality and interpretation of modelled results in two ways. First, unbiased statistical analyses may require all datasets to be of equal length, and furthermore, some analyses may impose the additional constraint that all datasets span identical time periods. Second, certain applications can be critically dependent upon the use of long term climate records. In these situations, short term datasets can only be used if they are correctly interpreted in a historical context. As climate data typically consist of a large number of stations with limited observation periods, the issue of record length must be considered carefully. If reliable algorithms are not used to enable reconstruction of long term records from short term datasets, a large proportion of the observed data may be rendered useless.

The majority of observational data is collected and processed by government agencies. These agencies are typically responsible for the installation and operation of the various data recording methods, and also the quality control of the observed data. Many users of such data accept it as being accurate and are not aware of the inconsistencies and biases which can occur in such data. While the existence of erroneous data may corrupt the output from a numerical model, these problems can be minimised by rejecting input data that is considered to be statistically incorrect. The existence of serially incomplete records is a more serious problem as it can result in the rejection of entire datasets. While numerous techniques for estimating missing data values have been implemented (Creutin and Obled, 1982), it is undesirable that individual researchers should have to expend considerable resources to develop their own databases. These problems could be overcome through the development of a single unified archive of quality climate data, which is publicly accessible. This would relieve individual workers of the problems associated with generating continuous climate records from sets of intermittent data. While extensive archives of observed and gridded datasets are available, these are usually restricted to mean datasets and/or monthly time steps. In many modelling applications one requires continuous daily time step records which are not widely available. Furthermore, spatial modelling usually requires access to high resolution gridded surfaces which are rarely available on a continental scale.

The purpose of this paper is to describe the construction of a climate database which was developed to specifically address the aforementioned issues regarding spatially and temporally incomplete datasets. In particular we describe database construction in Section 2, followed by an analysis of the interpolation error in Section 3. Discussion and concluding remarks are presented in Sections 4 and 5, respectively.

2. Database construction

A climate database has been constructed using observational data collected by the Australian Bureau of Meteorology. The database consists of continuous daily climate records at point locations, and sets of interpolated daily surfaces. The gridded surfaces were constructed to facilitate spatial modelling and the compilation of point datasets. Point records were constructed for stations which had long term observational datasets. At each location, the available data were used as a base to construct a complete and continuous daily climate record. Where observed data were unavailable, the interpolated surfaces were used to provide estimates. For the remainder of the paper we shall refer to the procedure of supplementing observed data with spatially interpolated data as *patching*. The resultant continuous rainfall and climate records constructed in this manner will be referred to as patched datasets.

Patched datasets have been constructed for approximately 4600 locations across Australia, shown in Fig. 1. At these locations, continuous daily time step records are available for rainfall, maximum and minimum tem-

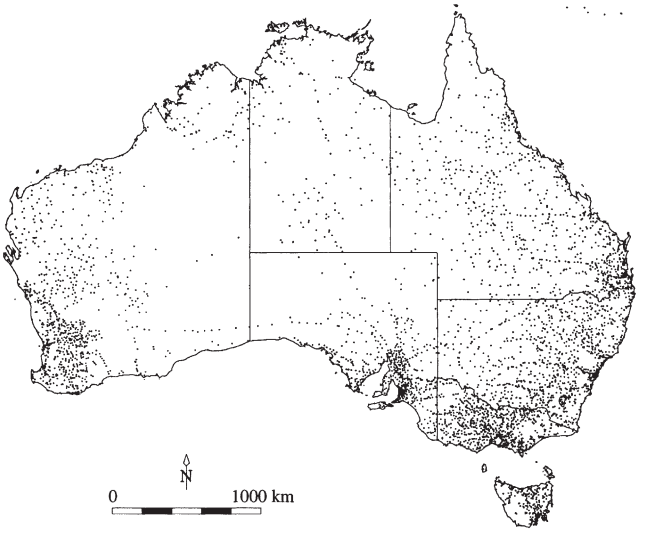

Fig. 1. Location of stations with patched datasets for rainfall and climate variables.

perature, Class A pan evaporation, solar radiation and vapour pressure. Both the gridded and point datasets cover the time period 1 January 1957 to the present, although a small number (about 10) commence as early as 1890. One should note however that all rainfall datasets commence in 1890. The contents of the database are summarised in Table 1.

2.1. Observational data

The Bureau of Meteorology maintains a network of rainfall and climate recording stations which report data using a variety of protocols. The majority of climate stations (i.e. those stations which report data for a comprehensive range of climate variables) report electronically, and consequently there is only a short delay between the time of observation and when the data become available. In contrast, only about 20% of rainfall recording stations report data using electronic means. At these locations, observational data are manually recorded on paper forms and submitted to the Bureau. Thus there can be a time delay from 1–2 days up to several months before data from these stations become available. The number and distribution of stations which report data by the various channels have important consequences for how one constructs a climate database from observational records. For the purposes of the following discussion, we shall refer to data for all dates up to but excluding the 12 months preceding the current date, as being *historical*. Data for dates within the previous 12 month period will be referred to as *near real-time* data.

The historical datasets maintained by the Bureau of Meteorology can undergo minor alterations for two principal reasons. First, additional climate records may be added to the existing data archive as new observational records become available, and as resources permit. Second, the historical data may be subject to ongoing error checking which may result in the rejection of some data. Hence, the historical datasets may slowly change, but do not undergo major revisions. Conversely, the Bureau's near real-time datasets are frequently updated to reflect the availability of new data and the results of error checking. To expand upon this point, it will be instructive to consider the processing procedure for daily data. On any given date, the available daily data for the preceding day can be extracted from the Bureau's database. This dataset will consist almost entirely of records from stations which report via electronic means and may have undergone initial error checking. In the case of climate variables, such as temperature, evaporation, etc., the initial dataset will not change significantly since most climate recording stations report electronically. Alterations may of course arise through subsequent error checking. The situation for rainfall records is considerably different as it may be several months before reports are received from rainfall stations which report using non-electronic means. As a result, the initial rainfall dataset may evolve over several months as new records arrive, in addition to alterations arising from error checking. To demonstrate this point, Fig. 2 shows how the number of rainfall reports for 15 July 1999 has changed with time. The figure shows the time delay that can occur between observation and date of entry in the database.

As noted above, the historical data undergo minor adjustments, so the entire dataset is updated approximately every 3 years. This allows users to take advantage of improvements in the observational records which may have arisen through error checking or the incorporation of additional data. The near real-time data (or components thereof) are extracted from the Bureau on a daily basis. Each day the rainfall and climate records for the preceding 24 h observational interval are downloaded and entered into the database. However, data for preceding days may have been modified, so existing data files

Table 1
Summary of data currently available in the database

Variable	Starting year	Patched	Gridded
Daily rainfall	1890	yes	yes
Monthly rainfall	1890	yes	yes
Maximum temperature	1957	yes	yes
Minimum temperature	1957	yes	yes
Class A pan evaporation	1970	yes	yes
Mean sea level pressure	1957	no	yes
Relative humidity	1957	no	yes
Solar radiation	1957	yes	yes
Vapour pressure	1957	yes	yes
Vapour pressure deficit	1957	no	yes

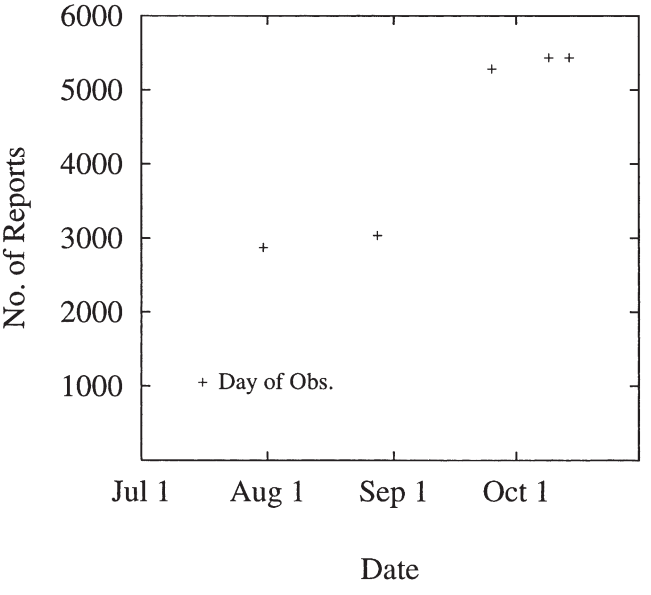

Fig. 2. Number of stations reporting daily rainfall for 15 July 1999. The figure shows the delay incurred between observation and when station reports are transmitted to the Bureau of Meteorology.

which have been updated are also downloaded. Owing to computational constraints it is only possible to monitor data files for the preceding 12 month period, in order to detect when modifications have been made. Modifications to data files for dates outside this 'moving time frame' will be incorporated when the historical dataset is next updated.

2.2. Construction of datasets

To construct continuous climate records from observational data, one must realise that hydrometeorological datasets are invariably plagued by the problem of missing data values. Data can be lost for a variety of reasons (Peck, 1997) such as failure of recording instruments, non-reporting of measured data, reporting incorrect data, and data rejection by error checking routines. Procedures for estimating missing data values are well documented (see for example Creutin and Obled, 1982) and usually consist of some form of spatial interpolation, whether it be in an explicit form or embedded within complex algorithms such as weather generators or artificial neural networks (Kuligowski and Barros, 1998). In the present work, missing data values have been estimated by explicit spatial interpolation of observational data.

The estimation of missing values raises the issue as to how interpolated approximations should be computed. Given the observed data for a particular day, one could compute interpolated estimates at all the preselected point locations that lack observational data. Alternatively, for each daily dataset one could compute (and optionally store) an interpolated surface spanning the domain of interest. Missing data values could then be conveniently extracted from the nearest grid cell on the surface. Such an approach has the distinct advantage that it provides data estimates at all locations on the computed grid, and is not limited to the predefined point locations. The latter approach was adopted, and a combination of kriging and smoothing splines were used to generate interpolated surfaces covering Australia. All surfaces were computed on a regular 0.05° grid extending from latitude 10°S to 44°S and longitude 112°E to 154°E.

The patched datasets are revised and updated on a daily basis because, as noted earlier, the archive of observational data changes daily through: (1) the incorporation of new daily data for the previous day; and (2) modification of existing data through error checking and the addition of new data values. For both new daily data and modified existing data, the patching procedure is essentially the same: interpolation algorithms are used to compute gridded surfaces, and those locations (without observational data) requiring interpolated values are identified. Interpolated estimates are computed for those locations (or taken directly from the gridded surfaces) and then used in conjunction with the observational data to update or modify the patched datasets. An overview of the data processing streams required to construct the patched datasets is shown in Fig. 3.

2.2.1. Interpolation of climate data

All climate variables were interpolated using a trivariate thin plate smoothing spline (Wahba and Wendelberger, 1980) with latitude, longitude and elevation as independent variables. Elevation was expressed in kilometres to minimise the validated root mean square interpolation error (Hutchinson, 1995). Latitude and longitude were in units of degrees. All surfaces were fitted by minimising the Generalised Cross Validation (GCV) error (Wahba, 1990) with the constraint of first order smoothness imposed.

To aid in the identification and removal of erroneous data, the interpolated climate surfaces were constructed using a two-pass interpolation scheme. Observed data were interpolated in a first pass and residuals computed for all data points. The residual is the difference between the observed and interpolated values. Data points with high residuals may be indicative of erroneous data and

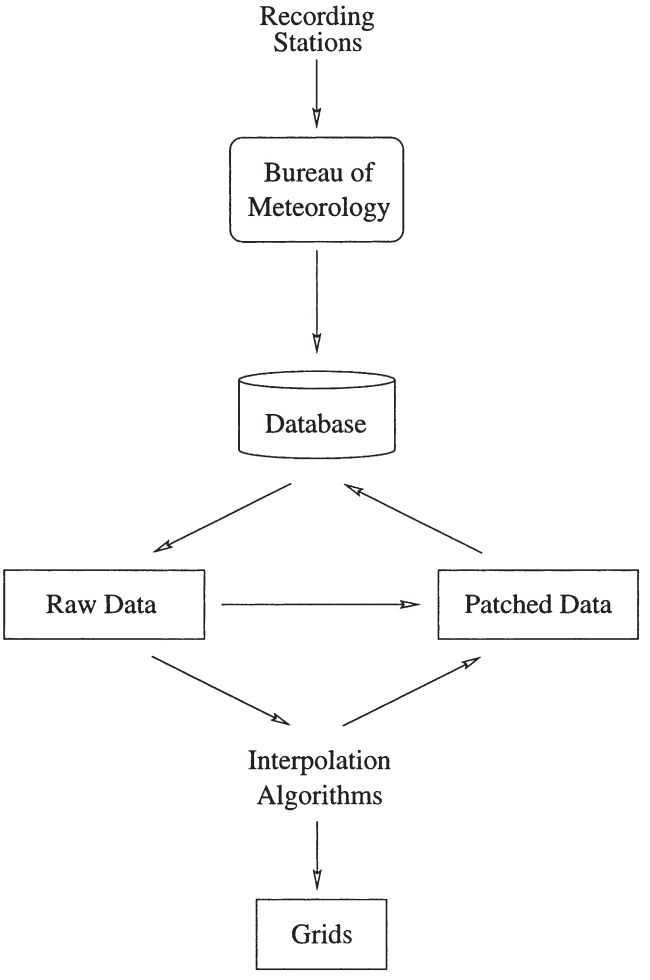

Fig. 3. Overview of the data processing streams required to construct the patched datasets.

were excluded from a subsequent interpolation which generated the final surface from which the patched datasets were constructed. A datum was rejected if the associated residual exceeded a fixed threshold. The thresholds were: 1.4°C and 1.6°C for maximum and minimum temperatures, respectively; 2.7 mm for pan evaporation; 3.5 hPa for pressure; 3.0 hPa for vapour pressure; 1.5 hPa for VPD; 10.0% for relative humidity; and 16.0% for percent extra-terrestrial radiation. The preceding thresholds were selected as they typically remove about 1–2% of the observed data. If the procedure rejected more than 5% of data from a given dataset, the threshold was adjusted so that not more than 5% of data would be rejected. Experience has shown the two-pass interpolation scheme can accurately detect and remove gross outliers in observed data. In some cases genuinely correct data may be removed, but this cost must be weighed against the effect that erroneous data can have on an interpolated surface.

Although a thin plate spline was used to interpolate all climate variables, some have characteristics which require further comment. In the following sections, we provide brief notes regarding the processing of various climate elements.

2.2.1.1. Pan evaporation The patched datasets have been constructed using data recorded from Class A pans which measure potential evaporation. Observational data prior to 1970 have not been included in the patched datasets because various measuring devices were in use before 1970, resulting in inconsistent and unreliable data. These problems have been described elsewhere (Hounam, 1961). For dates preceding 1970 the patched datasets have been supplemented with daily long term means for pan evaporation.

2.2.1.2. Pressure The spatial variation in observed pressure can be reduced if the topographic component is removed. This is done by converting the observed (station level) pressure into mean sea level pressure (Colquhoun, 1965):

$$P_o \approx P e^{\frac{z_g M_d g}{R(T+z_g\frac{\alpha}{2}+V_w H_{cf})}} \tag{1}$$

where P_o is the mean sea level pressure (hPa), P is the observed station level pressure (hPa), z_g is the geopotential height (m), M_d is the molecular weight of dry air (0.028965 kg mol^{-1}), g is the gravitational constant (9.80665 m s^{-2}), R is the universal gas constant (8.31451 J mol^{-1} K^{-1}), α is an average temperature lapse rate (0.0065 K m^{-1}), V_w is the saturation vapour pressure of aqueous water (hPa), H_{cf} is the humidity correction factor (K/hPa) and T is the temperature (K), usually set equal to the dry-bulb temperature. The geopotential height, humidity correction factor and saturation vapour pressure of aqueous water can be computed given the wet- and dry-bulb temperatures, station level pressure, and the station's elevation (z) and latitude (e.g. Manual of Barometry, 1963). If wet- and dry-bulb temperatures are unavailable, an approximate mean sea level pressure can be computed using a simple exponential decay model (Bohren and Albrecht, 1998):

$$P_o \approx P e^{\frac{z_g M_d g}{R(T_o-\alpha z_g)}} \tag{2}$$

where T_o (=298.0 K) is an average air temperature at sea level.

Mean sea level pressure is interpolated using a two-dimensional smoothing spline, with latitude and longitude as independent variables.

The frequency and times at which pressure, wet-bulb and dry-bulb temperatures are recorded varies between reporting stations. Since the number of stations which report 9 a.m. data exceeds the number reporting 3 p.m. data by a factor of four, the interpolated surfaces are constructed using 9 a.m. data. One should note, however, that '9 a.m.' data may include reports ranging from 8 a.m. to 10 a.m. owing to different time zones and the effects of daylight saving.

2.2.1.3. Vapour pressure Vapour pressure can be calculated given the wet- and dry-bulb temperatures and station level pressure (Letestu, 1973):

$$VP = \begin{cases} VP_s - 7.866\times10^{-4} P(T_d - T_w)\left(1.0+\dfrac{T_w - T_i}{610.0}\right): & T_w \geq T_i \\ VP_s - 7.110\times10^{-4} P(T_d - T_w)\left(1.0+\dfrac{T_w - T_i}{671.2}\right): & T_w < T_i \end{cases} \tag{3}$$

where T_w and T_d are the wet- and dry-bulb temperatures (K), respectively, and the saturated vapour pressure is given by:

$$VP_s = \begin{cases} \log^{-1}\{10.79574\left(1.0-\dfrac{T_t}{T_w}\right) - 5.02800\log\left(\dfrac{T_w}{T_t}\right) \\ +1.50475\times10^{-4}\left(1.0-\log^{-1}\left[8.29690\left(1.0-\dfrac{T_w}{T_t}\right)\right]\right) & : T_w \geq 0 \\ +0.42873\times10^{-3}\left(\log^{-1}\left[4.76955\left(1.0-\dfrac{T_t}{T_w}\right)\right]-1.0\right)+0.78614\} \\ \log^{-1}\{9.09685\left(1.0-\dfrac{T_t}{T_w}\right) - 3.56654\log\left(\dfrac{T_t}{T_w}\right) \\ +0.87682\left(1.0-\dfrac{T_w}{T_t}\right)+0.78614\} & : T_w < 0 \end{cases} \tag{4}$$

where T_t is the triple point of water (273.16 K) and T_i is the ice point of water (273.15 K). If pressure is unavailable, an approximate pressure can be computed using the station elevation:

$$P \approx 1013.2127 - 0.1195z + 5.1681 \times 10^{-6} z^2, \quad -250m < z < 3250m. \tag{5}$$

The polynomial coefficients in Eq. (5) were computed by least squares fitting a second-order polynomial to tabulated pressure–elevation data (List, 1966).

Vapour pressure computed using 9 a.m. data is interpolated using a three-dimensional smoothing spline.

2.2.1.4. Vapour pressure deficit Vapour pressure deficit (VPD) may be defined as the difference between the actual vapour pressure and the vapour pressure under saturated conditions. This quantity is intrinsically poorly defined since saturation vapour pressure (SVP) is strongly dependent upon temperature. VPD has been included in the database as it is a direct measure of atmospheric demand for moisture, and consequently of interest to the plant modelling community.

The computation of VPD has two components. First, one must compute the saturated vapour pressure. Since atmospheric pressure is reasonably constant throughout the day, the SVP may be calculated at any desired temperature (e.g. max, min, etc.) using pressure data recorded at the nearest appropriate time. Second, one must compute the actual vapour pressure. This quantity would, ideally, be computed using wet- and dry-bulb temperature and pressure data recorded at the interval nearest the time at which the VPD is being reported. However, owing to the data constraints outlined earlier, vapour pressure is best calculated using 9 a.m. data. This does not present a serious difficulty as vapour pressure can be assumed to be relatively constant throughout the day. This assumption is implicit in the work of Meinke (1996) and can be justified by examination of observed data. The relative difference between 9 a.m., 12 p.m. and 3 p.m. vapour pressure data is shown in Fig. 4. The figure was constructed by computing the absolute relative difference between 9 a.m./12 p.m. data, and 9 a.m./3 p.m. data, for all Australian stations on a given date. The results were then spatially averaged to obtain a mean absolute relative difference for that date. This quantity was computed for a series of dates chosen by randomly selecting a single day out of each month from January 1957 to December 1997. To aid comprehension of the results, a least squares polynomial of order 3 was fitted to both sets of data. The figure shows that the relative variation in vapour pressure throughout the day is of the order of 10–16%. Given measurement error in calculating the vapour pressure due to air flow effects on wet-bulb temperature, it is reasonable to assume constant vapour pressure throughout the day.

Vapour pressure deficit is computed at an *average* daytime temperature:

$$\mathrm{VPD} = 0.75\ \mathrm{SVP}(T_{\max}) + 0.25\ \mathrm{SVP}(T_{\min}) - VP_{9\mathrm{am}} \tag{6}$$

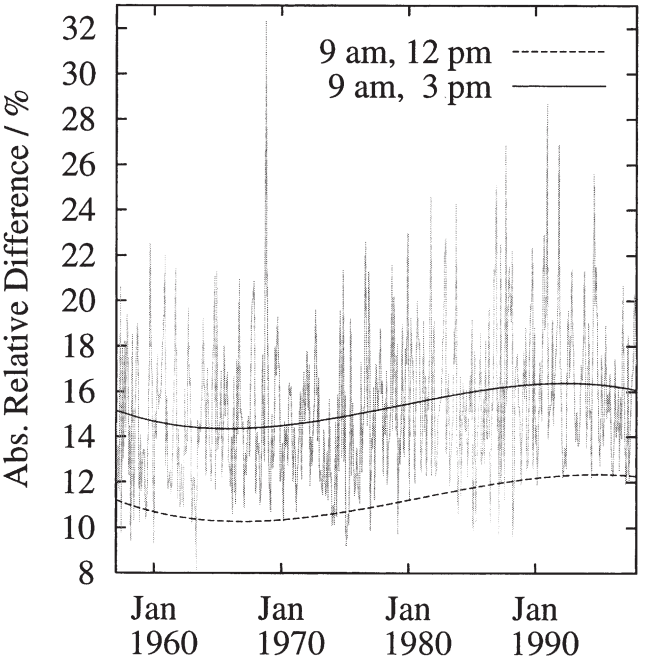

Fig. 4. Absolute relative difference between 9 a.m./12 p.m. vapour pressure data, and 9 a.m./3 p.m. vapour pressure data. Statistics computed were $\frac{|VP(9\mathrm{am}) - VP(12\mathrm{pm})|}{VP(9\mathrm{am})} \times 100\%$ and $\frac{|VP(9\mathrm{am}) - VP(3\mathrm{pm})|}{VP(9\mathrm{am})} \times 100\%$. A least squares polynomial of order 3 was fitted to both sets of data. The actual results for the 9 a.m./3 p.m. datasets are shown in lighter shading.

as this has previously been reported as being of greatest utility for modelling applications (Tanner and Sinclair, 1983; Meinke, 1996).

2.2.1.5. Relative humidity Relative humidity can be defined as the ratio of the actual vapour pressure to the saturation vapour pressure (Bohren and Albrecht, 1998). For data supply reasons, the actual vapour pressure is computed using 9 a.m. data. For modelling applications, the saturated vapour pressure is calculated at maximum temperature. Consequently, the computed relative humidity corresponds to relative humidity at maximum temperature:

$$RH = \frac{VP_{9\mathrm{am}}}{SVP(T_{\max})} \times 100\% \tag{7}$$

2.2.1.6. Solar radiation Solar radiation estimates have been computed using data taken from actual radiation measurements, hours of sunshine duration and estimates of cloud oktas. It was necessary to use data from different sources since only about 25 stations measure radiation directly. Approximately 70 stations measure hours of sunshine duration and about 500 climate recording stations estimate cloud oktas.

In order to estimate actual solar radiation from observational estimates of cloud oktas, empirical tables were constructed using data from stations which simultaneously recorded both oktas and actual radiation. These tables were constructed by optimising the correlation between estimated cloud oktas and the measured radiation. However, since the theoretical extra-terrestrial radiation that is available for atmospheric transmission is dependent upon latitude and time of year, all measured radiation data were converted to percent extra-terrestrial (*%E-T*) radiation. This removes location and time of year effects and allows radiation data for all available dates and stations to be used in the construction of empirical tables. The coefficient of determination (r^2) between percent extra-terrestrial radiation and 9 a.m. estimates of cloud oktas was found to be 0.50. This was improved if an average of 9 a.m. and 3 p.m. cloud oktas was used (r^2=0.60). However, best results were obtained using a two-dimensional table, consisting of both 9 a.m. and 3 p.m. estimates of cloud oktas (r^2=0.71). A further marginal improvement was obtained using 9 a.m., 12 p.m. and 3 p.m. oktas; however, the number of stations which report all three estimates is unacceptably low. The empirical array used to estimate percent extra-terrestrial radiation from cloud oktas is shown in Table 2. Given the 9 a.m. and 3 p.m. cloud okta estimates, the percent extra-terrestrial radiation can be read directly from the table.

To estimate solar radiation from hours of sunshine duration, an empirical equation was constructed using data from stations which simultaneously recorded both sunshine duration and actual radiation. A linear relationship was fitted using regression:

$$\%E\text{-}T = 4.07 \cdot \text{sunshine hours} + 25.50 \quad (8)$$

resulting in a coefficient of determination of r^2=0.85 with the measured (*%E-T*) radiation.

The three sources of radiation data are merged to preferentially include the highest quality data source (Carter et al., 1996). Measured radiation is the most accurate data and is used if available. Sunshine duration measurements are more accurate than cloud okta estimates and are used if measured radiation is unavailable. If neither measured radiation nor sunshine duration is available, cloud okta data are utilised.

Percent extra-terrestrial radiation is interpolated using a three-dimensional smoothing spline with latitude, longitude and elevation as independent variables. Elevation is incorporated as atmospheric transmittance of radiation is affected by cloud and optical air mass, both of which are influenced by elevation. The interpolated surface of *%E-T* radiation is then converted, pixel by pixel, to a gridded surface of incident radiation (MJ m^{-2}). Actual radiation is not interpolated as this quantity is parametrically dependent upon latitude, and thus incorporates the component of spatial variation due to solar angle. This component is explicitly removed in the transformation to *%E-T* radiation and is expected to result in reduced interpolation error.

An algorithm for estimating radiation using geostationary infrared satellite data (Dedieu et al., 1987) has recently been implemented. This procedure is currently undergoing validation testing, but as yet has not been employed in the construction of the climate database.

2.2.2. Interpolation of rainfall data

Interpolating daily rainfall is intrinsically difficult due primarily to its inherent spatial and temporal variability. Moreover, the short range spatial correlation displayed by many rainfall events makes it virtually impossible to resolve complex rainfall patterns using data from existing rain gauge networks. These facts have led to widespread interest in the development of various approaches which seek to supplement gauge data with information from various other sources. Techniques used to date include stochastic rainfall models, weather generators, radar and satellite remotely sensed data, downscaling of General Circulation Model outputs, genetic algorithms and artificial neural networks. This is by no means a comprehensive list and despite recent advances in these

Table 2
Average percent extra-terrestrial radiation as a function of 9 a.m. and 3 p.m. cloud cover

		9 a.m. oktas								
		0	1	2	3	4	5	6	7	8
3 p.m. oktas	0	74.0	73.6	72.9	72.4	71.2	71.0	70.1	67.0	57.5
	1	72.9	71.9	71.0	70.0	69.0	67.9	66.2	62.3	56.4
	2	71.1	69.8	69.1	67.8	66.5	65.5	63.9	58.6	53.4
	3	69.3	68.2	67.1	65.7	65.2	63.5	61.6	56.2	49.7
	4	68.0	66.9	65.0	64.2	63.3	61.5	59.3	54.3	47.2
	5	66.5	65.4	63.6	61.7	61.5	60.3	57.6	52.0	45.3
	6	63.9	63.0	60.7	59.8	58.5	57.6	55.1	49.8	42.2
	7	59.6	58.1	55.8	54.6	53.4	52.2	49.8	43.6	34.8
	8	50.2	49.1	46.9	46.3	43.4	44.5	40.1	33.2	24.5

fields (e.g. cf Hsu et al., 1997, 1999), there is still a fundamental requirement to interpolate observational rainfall data when supplementary information is unavailable.

In the absence of topographic influences, the short range spatial variability of rainfall is expected to decrease as the accumulation period increases, i.e. monthly values will be less variable than daily values. Unfortunately, topographic influences cannot be neglected, as evidenced by the complex patterns displayed by long term statistics. The effect of topography on precipitation does, however, suggest that accumulated rainfall could be interpolated more reliably if one could remove the component of rainfall variability due to topographic influences. This approach has been exploited successfully by a number of workers (Stidd, 1973; Richardson, 1977; Hutchinson et al., 1993) who have shown that normalised rainfall can be reliably interpolated for time steps ranging from monthly to hourly. The normalisation approaches utilise the fact that an appropriately chosen fractional power of observed rainfall data has an approximately normal distribution. Hutchinson et al. (1993) have proposed fitting a normal distribution via maximum likelihood which is truncated at a small observational threshold. The normalisation parameters are fitted locally using observational data and hence they can vary spatially. When tested on rainfall data across the United States, Hutchinson et al. (1993) reported an overall mean fractional power of 0.45 when applied to monthly data, and 0.55 for daily data.

In the present work, normalisation parameters have been computed directly from observational data with the fractional power fixed at 0.50, i.e. the fractional power does not vary spatially. The procedure used for interpolating accumulated monthly rainfall data can be summarised as follows.

1. Normalisation parameters are computed using observational data raised to the appropriate fractional power (0.50 in this case).
2. The normalisation parameters are spatially interpolated using a trivariate thin plate smoothing spline with latitude, longitude and elevation as independent variables. Since most of the topographic information is captured by the long term mean and variance of monthly rainfall, the fitted surface is highly complex.
3. The fractional power of observed monthly rainfall data is normalised to remove the long term component of monthly rainfall which is strongly topographically related. The residual is the short term (or monthly) departure from the long term rainfall pattern and is due primarily to the broad scale synoptic conditions existing throughout the accumulation period.
4. The normalised residual is spatially interpolated using ordinary kriging (see for example Isaaks and Srivastava, 1989) with zero nugget and variable range. The nugget is set to zero to enforce exact interpolation and the range is adjusted locally to the average distance between the target and data points. All data points located outside a 1.0° perimeter of the target are excluded from the interpolation; however, this boundary may be adjusted to ensure that a minimum of 25 data points are used in the interpolation. The covariance structure is assumed to be of an exponential form.
5. At each grid cell, the interpolated residual is converted to monthly rainfall by reversing the normalisation. The parameters used to effect the back transformation are the interpolated mean and variance for that particular grid cell.

Interpolated daily rainfall surfaces can be computed by interpolating daily rainfall directly, or they can be generated from interpolated monthly rainfall. Using the first procedure, one directly interpolates rainfall data collected for the 24 h period of interest. Alternatively, daily rainfall surfaces can be derived from interpolated monthly surfaces by partitioning the monthly total onto individual days. In order to do this, the daily distribution of rainfall at each grid cell must be computed. In other words, determine the proportion of monthly total rainfall that was recorded on individual days throughout the month. The daily distribution can be estimated by interpolating the daily rainfall for each day in the month, thus generating an estimate of the daily distribution of rainfall for each grid cell. The interpolated total monthly rainfall is then partitioned onto individual days in accordance with the computed distribution. An alternative to computing the daily distribution is to assume that each grid cell has the same distribution as the nearest recording station. A comparison of the accuracy of surfaces derived by the partitioning of monthly data according to nearest neighbour and computed distributions is presented in Section 3. The nearest neighbour approach has been used in the construction of all patched datasets.

In Section 3 we compare the cross validated error statistics for the three methods used to interpolate daily rainfall. These statistics indicate that interpolating daily rainfall directly is more accurate than deriving daily rainfall surfaces from interpolated monthly rainfall. However, the differences are minor and all issues need to be considered when evaluating the relative merits of the two approaches. In particular, it is important to consider the quality of the monthly rainfall data compared with that of the daily data.

The interpolation of daily data is problematic for several reasons. For example, many stations do not report daily rainfall on weekends and public holidays. Rain events occurring during such periods are often reported as an accumulated total. Since the collection period was not the standard 24 h observational interval, accumulated rainfalls cannot be used in daily interpolations. This

problem can cause serious data loss when interpolating historical data. The interpolation of near-real time data is made even more difficult by the fact that data are only available from the subset of stations that report electronically. SYNOPS stations report both zero and non-zero rainfall, but telegraphic stations report only non-zero rain events (Mills et al., 1997). This point can seriously compromise the integrity of the near-real time data set as it is biased due to the non-reporting of zero rainfall by the telegraphic stations. To alleviate this problem, the near-real time datasets can be supplemented by 'assumed zeroes' for those telegraphic stations that did not report. However, this procedure requires an accurate list of stations that report telegraphically. If, for example, such a station ceases telegraphic correspondence, false zeroes will continue to be added to the dataset until the problem is identified.

Deriving daily rainfall surfaces from monthly data can avoid some of the aforementioned problems associated with daily datasets, but at the expense of introducing other problems. For example, the interpolation of monthly data can use accumulated rainfall amounts, whereas daily interpolations cannot. This is because accumulated amounts can be incorporated into the monthly total, providing the accumulation period does not bridge monthly boundaries. However, the benefits obtained through the use of accumulated data must be considered with the problems introduced by the fact that monthly rainfall is usually computed by summing the daily amounts. If valid data are available for the entire month, the summed monthly total is correct. However, if daily data (and/or accumulated data) are unavailable for some portion of the month, the summed monthly total may be incorrect and consequently, the daily surfaces derived from the incorrect monthly total will be affected. In contrast, surfaces derived by direct interpolation of the daily data would be correct: those days for which data were available would be unaffected, while grids computed for those days without data may have surface quality diminished through reduced data content, they would nevertheless not be affected.

The interpolated surfaces used in the construction of the patched rainfall datasets have been constructed as follows. Prior to the end of the month, daily rainfall surfaces are generated by ordinary kriging of the available (daily) data. These surfaces are continually reinterpolated throughout the month as the near real-time datasets are updated with additional and error-checked data. At the end of the month, or typically a few days thereafter, the total monthly rainfall becomes available. The monthly total is spatially interpolated and used to derive daily rainfall surfaces which then supersede those surfaces computed using daily data. As a result, all daily rainfall surfaces, except those for the current month are derived from interpolated grids of normalised total monthly rainfall.

2.3. Composition of datasets

To enable the user to identify the data source and quality control status, flags have been logged to indicate the quality and origin of the data. The quality control flags are predominantly as supplied by the Australian Bureau of Meteorology. A data source flag indicates if the data have been drawn from actual observation, spatial interpolation, or in the case of rainfall, deaccumulation. If a given rainfall amount is the accumulated total for a period exceeding the standard 24 h interval, the total rainfall is partitioned onto the individual days comprising the accumulation period. The accumulated value is partitioned in accordance with the daily distribution of rainfall obtained from the interpolated surfaces.

3. Estimation of interpolation error

An accurate estimate of interpolation error may be critical for three reasons. First, the output from numerical models must be correctly interpreted in the context of the model's sensitivity to input data. Obviously the outputs of any numerical model must be critically appraised if the model is sensitive to perturbations in the input data that is below the expected magnitude of interpolation error. Second, if management decisions are going to be based upon interpolated data, or the results of numerical models utilising such data, a desirable prerequisite is an accurate estimate of the error bounds on the interpolated data. Finally, accurate error statistics facilitate the intercomparison of both datasets and interpolation algorithms. Benchmark results allow one to compare the performance of a given algorithm on different datasets, and also the results of different algorithms on an individual dataset.

Independent cross validation has been used to investigate the interpolation error for a number of climate variables. This technique requires withholding a particular datum from the interpolation, and then using the interpolation algorithm to estimate the withheld value using the remaining data points. Each data point may be sequentially cross validated, and consequently n interpolations are required to independently cross validate n data points. The mean interpolation error can then be estimated by computing the average of the differences between the n observed values and their corresponding interpolated estimates. This technique is widely used for assessing interpolation schemes, but one should be aware of its shortcomings. Specifically, cross validation will usually overestimate the interpolation error because the interpolated estimate is being computed at a location where data are genuinely available. In addition, the computed surface and hence the cross validated estimate may be altered by the removal of the point being cross vali-

dated. In practice these issues are unavoidable but have less impact as the number of data points increases.

When multiple datasets are available, the mean interpolation error can be computed for each dataset. In the current context, the mean error can be computed for a series of daily datasets for any desired climate variable. The time series of error estimates will then illustrate how the mean interpolation error varies with time, due to effects such as seasonal fluctuation.

For climate applications it can generally be assumed that the mean interpolation error is dependent upon the location and density of data points. Therefore the interpolation error will (in most cases) vary spatially since neither the topography nor the spatial distribution of data recording stations is uniform. Neglecting other effects, interpolation error will usually be lower in areas of high station density compared to areas where stations are sparsely distributed.

From the preceding discussion it can be seen that a complete error analysis of an interpolation system requires two components. First, a temporal analysis reveals how the error varies in time. The mean interpolation error for a given date is computed by averaging the interpolation error incurred at the discrete station locations. This procedure is repeated for a series of dates, yielding an estimate of the mean interpolation error as a function of time. Since the error has been averaged over all station locations, no spatial information is provided. Secondly, a spatial analysis shows how the error varies with location. The mean interpolation error for a given recording station is computed by averaging the interpolation error incurred on all dates analysed. The mean error is computed for all recording stations, thus providing an estimate of the mean interpolation error as a function of location. Since the error has been averaged over all dates, no temporal information is provided. Consequently the spatial and temporal analyses provide complementary information.

To evaluate the accuracy of the interpolated data used to construct the patched datasets, it was necessary to examine the interpolation error for the entire time frame spanned by the datasets. This immediately raises the point that stations typically record data for discrete and often intermittent periods of time. As a result, it is not possible to select a series of data files covering the period of interest and attempt to compute valid error analyses. The intermittent nature of many recording stations precludes this approach because the number of stations that have data recorded for the entire period, and hence the number upon which the error averages would be computed, is small. To overcome this problem, one must define a threshold for the minimum number of days of cross validated data that is deemed acceptable to compute mean errors for temporal analyses. Obviously the threshold must be defined in accordance with the total number of days analysed, such that there is an acceptable number of stations to perform the spatial analysis.

Spatial and temporal error analyses have been performed by independently cross validating data for daily and monthly rainfall, maximum and minimum temperature, Class A pan evaporation, vapour pressure and mean sea level pressure. The daily datasets used for the analyses were chosen by randomly selecting a single day from each month in the analysis period. The period examined was 1957–1997 for climate variables, and 1900–1997 for daily and monthly rainfall. Therefore a total of 492 days of data was used for the climate analyses, and 1176 days of data used for the rainfall analyses. Error statistics were computed for rainfall stations which had at least 300 days of cross validated data available. Climate reporting stations were required to have at least 150 days of cross validated data, for inclusion in the error analysis. The following statistics were computed: root mean square error (RMSE), mean absolute error (MAE), mean error (ME) or bias and the coefficient of determination (r^2). All error statistics were computed using the observed and independently cross validated data. It should be noted that the coefficient of determination would be unrealistically overestimated if the seasonal data signal was not removed. This is readily understood if one compares observed and interpolated data for climate variables such as temperature or evaporation. These variables display a seasonal fluctuation that is approximately sinusoidal, and the magnitude of this variation is large, relative to the interpolation error. Consequently, the computed coefficient of determination is dominated by the seasonal fluctuation and does not accurately reflect the correlation (or lack thereof) between the relevant datasets. For this reason the mean monthly signal was removed from both the observed data and cross validated estimates, prior to the computation of r^2.

3.1. Climate variables

The results of the spatial error analyses are presented in Figs. 5–9. In each case, the coefficient of determination and mean absolute error is shown in parts A and B, respectively. These statistics were selected as r^2 is a relative index which indicates the proportion of data variance explained by the interpolation algorithm, and the mean absolute error is the error that one could expect in an interpolated estimate where an observed datum is unavailable. RMS errors are not presented as these statistics can be unrealistically influenced by a small number of large errors. The bias or mean error is not presented as it was found to be negligible (see Table 3).

The spatial analyses were compiled by computing mean error statistics at discrete points corresponding to the location of recording stations. To aid in the interpretation of the results, the error statistics were spatially interpolated using a two dimensional smoothing spline.

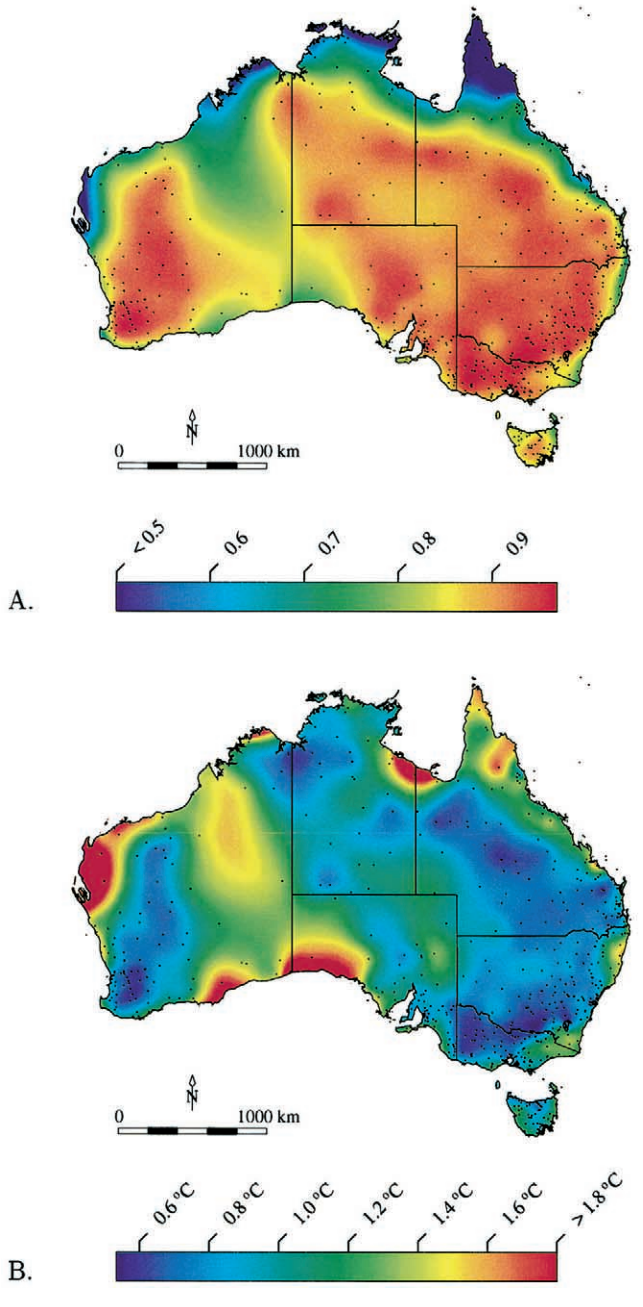

Fig. 5. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed daily maximum temperature data and cross validated estimates. All estimates were computed using a 3-D thin plate smoothing spline. Dots indicate the location of recording stations used in the analysis.

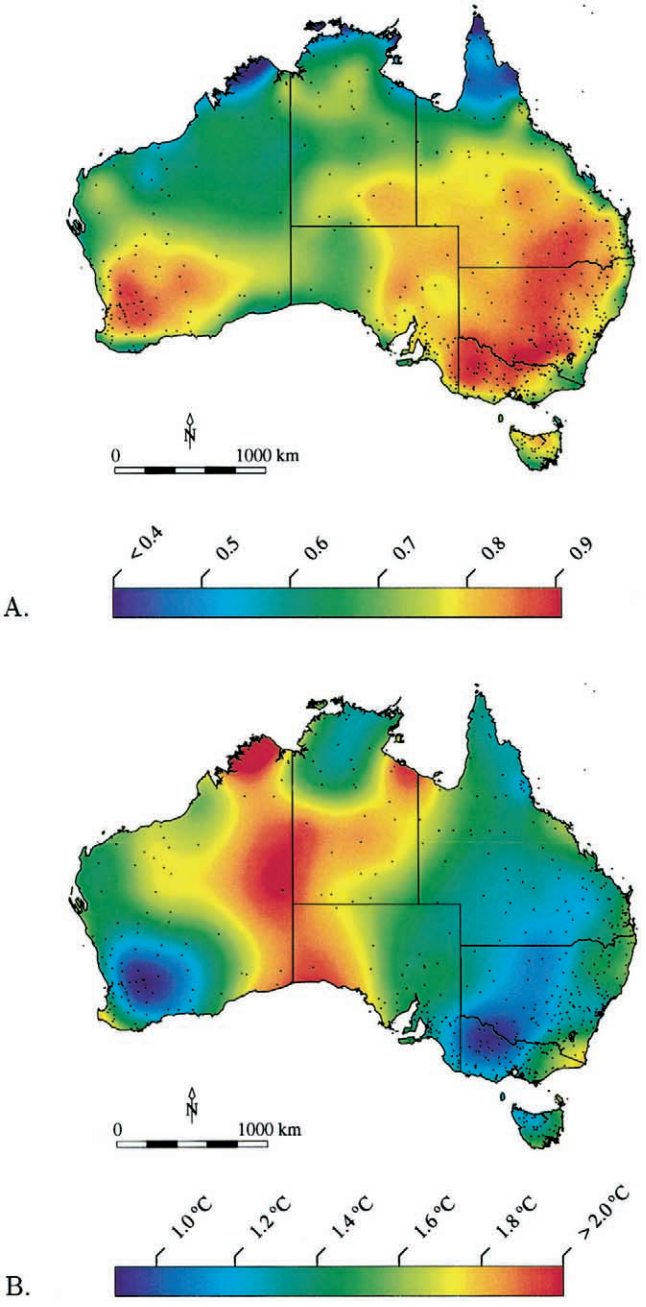

Fig. 6. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed daily minimum temperature data and cross validated estimates. All estimates were computed using a 3-D thin plate smoothing spline. Dots indicate the location of recording stations used in the analysis.

The location of the recording stations used to construct the spatial maps are indicated by dots. Only those stations with at least 150 days of cross validated data were included in the analysis.

The results of the temporal error analyses are presented in Figs. 10–14. The temporal plots show how the interpolation error varies in time due to influences such as seasonal fluctuation, and the number of stations reporting data at any point in time. The seasonal fluctuation is best demonstrated by evaporation, and is easily seen if the time scale is expanded, as shown in Fig. 15a. The figure shows that the interpolation error decreases mid-year, which is the Southern Hemisphere winter. However, the reduced error does not imply that the interpolation is more accurate in winter — it is simply a consequence of the fact that evaporation decreases during winter. Fig. 15b confirms this as the relative error

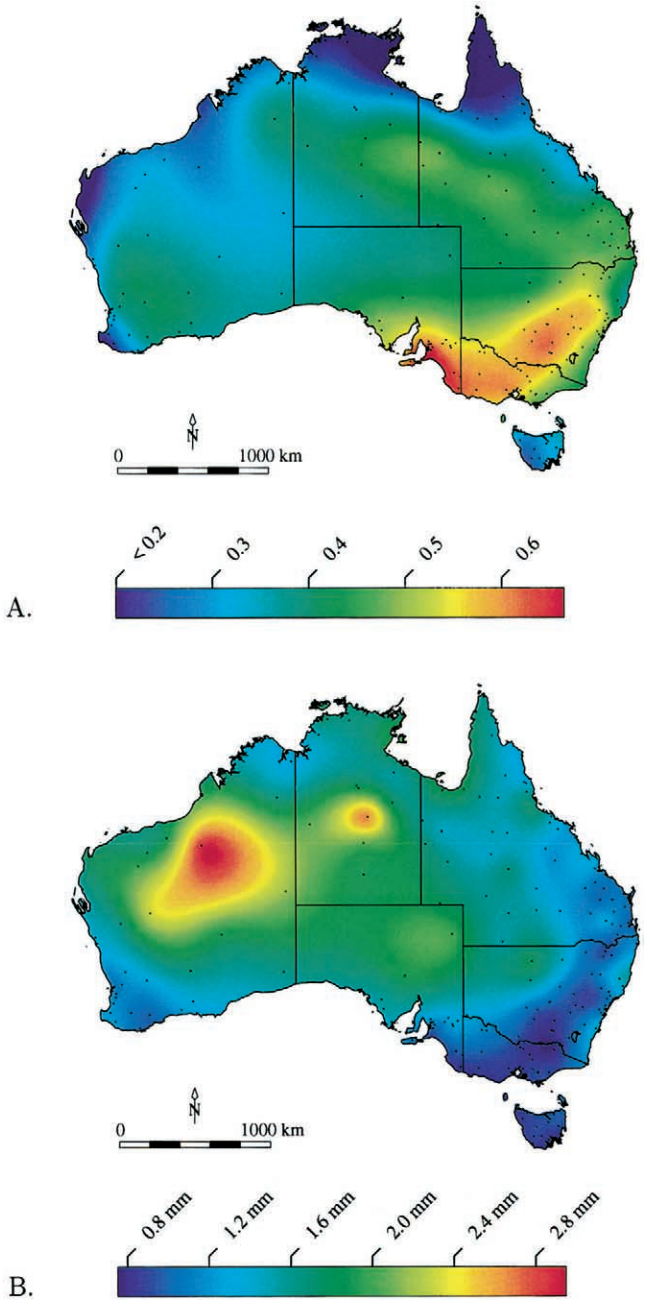

Fig. 7. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed daily Class A pan evaporation data and cross validated estimates. All estimates were computed using a 3-D thin plate smoothing spline. Dots indicate the location of recording stations used in the analysis.

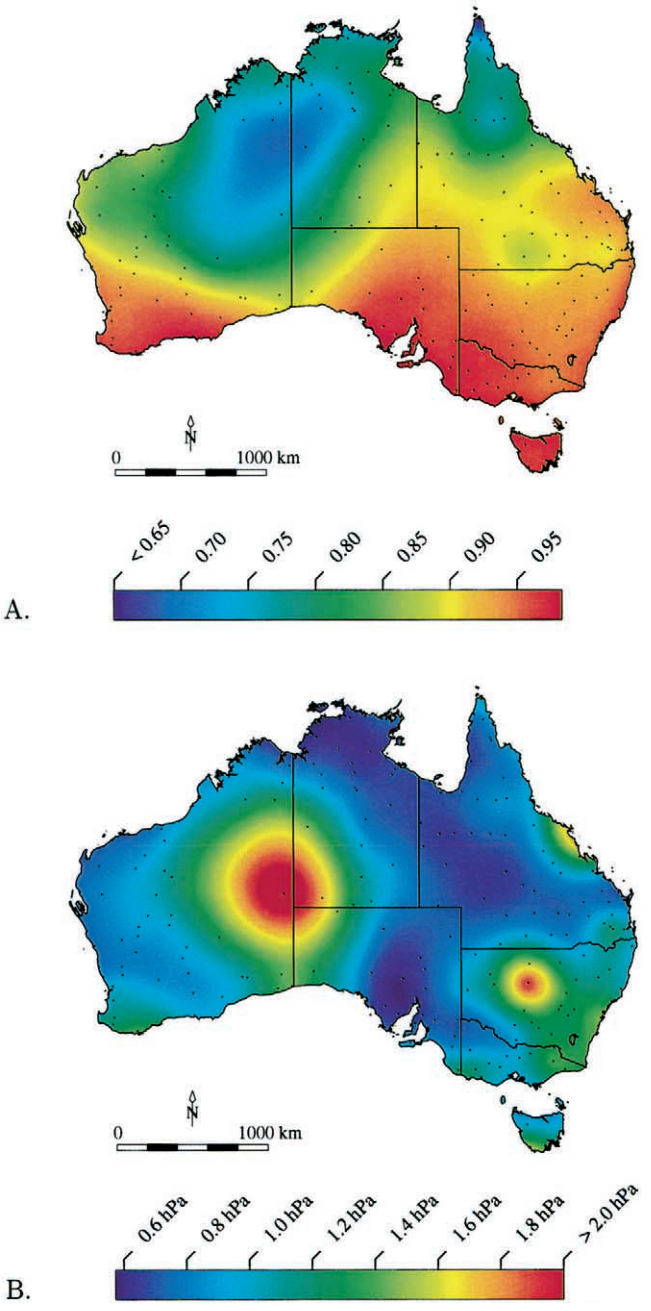

Fig. 8. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed daily mean sea level pressure data and cross validated estimates. All estimates were computed using a 2-D thin plate smoothing spline. Dots indicate the location of recording stations used in the analysis.

does not exhibit the same seasonal fluctuation as the absolute error.

Figs. 10–14 also show the number of recording stations used in the error analyses. While the number of stations shown closely represents the number of reported values, there were some instances (e.g. 1991) when the database did not contain all available data. The additional data will be entered into the database when the historic datasets are next updated.

To obtain an overall view of the accuracy of the interpolation systems used to construct the patched datasets, the error statistics can be averaged over both space and time to obtain global statistics. These statistics provide a simple estimate of the skill, and hence error, in estimating climate data for locations where observed data are unavailable. However caution must be exercised when using global statistics as both the temporal and

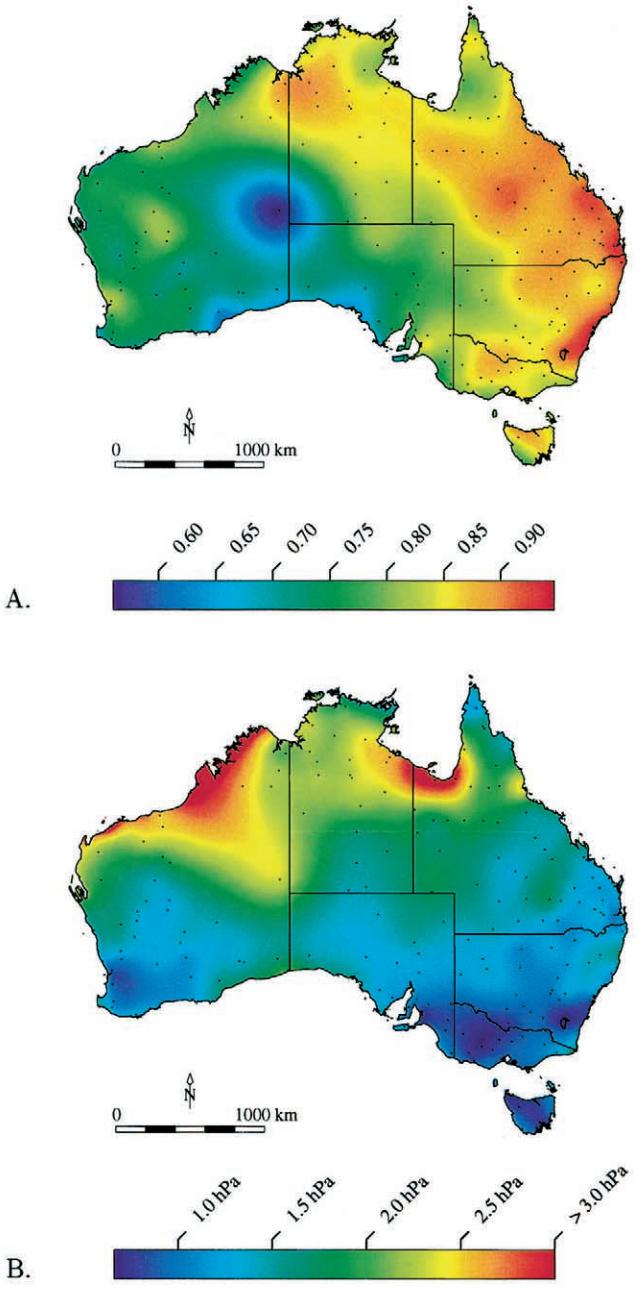

Fig. 9. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed daily vapour pressure data and cross validated estimates. All estimates were computed using a 3-D thin plate smoothing spline. Dots indicate the location of recording stations used in the analysis.

spatial variation are being ignored. Summary statistics are presented in Table 3.

3.2. Anomaly interpolation of climate variables

If one neglects the presence of noise in measured data due to short range or microscale variation, it can be assumed that interpolation error will decrease as the spatial variation of the data decreases. Since surface topography is responsible for most short range variation, the spatial variance in observed data can be reduced by removing the long term mean field which captures most of the topographically related variation. This point provides motivation to test the feasibility of interpolating daily departures from long term means, as opposed to interpolating raw data directly.

Interpolating daily anomalies has been tested for maximum and minimum temperature, Class A pan evaporation and mean sea level pressure. Independently cross validated error statistics were computed as follows.

1. Daily long term means and variances were computed for all stations for which error analyses were to be undertaken.
2. Cross validated means were computed using a trivariate thin plate smoothing spline. All data were scaled according to the relative variance to reflect the reliability of the statistic being interpolated.
3. The departures between the observed daily data and the corresponding (cross validated) daily means were computed and independently cross validated using a two-dimensional smoothing spline.
4. The cross validated anomaly was then added to the cross validated mean to yield the cross validated estimate.
5. Error statistics were computed using the observed data and the cross validated estimates calculated in Step 4.

Error statistics for the anomaly interpolation technique have been summarised and are presented in Table 4. Comparison with the analogous statistics for the direct interpolation of observed data (Table 3) indicates the anomaly technique does not reduce the interpolation error. The temporal error profiles for the direct and anomaly interpolation methods are presented in Fig. 16. To assist in the comparison, least squares polynomials of order 3 and 5 were fitted to the mean absolute error and correlation data, respectively. The temporal plots indicate that the anomaly technique is less accurate than interpolating observed data directly, for most of the time period analysed. However, as the number of data points decreases, the anomaly technique becomes more accurate than direct interpolation. As the number of data records available pre-1957 is very limited, an anomaly interpolation technique may be necessary to accurately interpolate these data. The spatial error analyses (not presented) indicate that the areal distribution of errors arising from the direct and anomaly techniques are similar. However, it must be noted that the spatial analyses were constructed by averaging the error statistics over the entire time period examined. A similar comparison of the spatial distribution of errors for periods with very limited datasets may reveal the anomaly technique to be superior.

Table 3
Summary statistics for the interpolation of observed daily climate data. Statistics were computed by spatially and temporally averaging the cross-validated interpolation error incurred at all stations included in the analysis

Statistic	Climate variable[a]				
	Max. T.	Min. T.	Evap.	Pres.	V.P.
1957–1997[b]					
RMSE	1.4°C	1.9°C	1.8 mm	1.7 hPa	2.2 hPa
MAE	1.0°C	1.4°C	1.3 mm	1.0 hPa	1.5 hPa
ME	0.0°C	0.0°C	0.0 mm	0.0 hPa	0.0 hPa
r^2	0.96	0.91	0.76	0.99	0.85
1990–1997					
RMSE	1.5°C	1.9°C	1.7 mm	1.5 hPa	2.2 hPa
MAE	1.0°C	1.4°C	1.2 mm	0.8 hPa	1.5 hPa
ME	0.0°C	0.0°C	0.0 mm	0.0 hPa	0.0 hPA
r^2	0.97	0.93	0.78	1.00	0.85

[a] Max. T., maximum temperature; Min. T., minimum temperature; Evap., Class A pan evaporation; Pres., mean sea level pressure; V.P., vapour pressure.
[b] Analysis period for evaporation: 1967–1997.

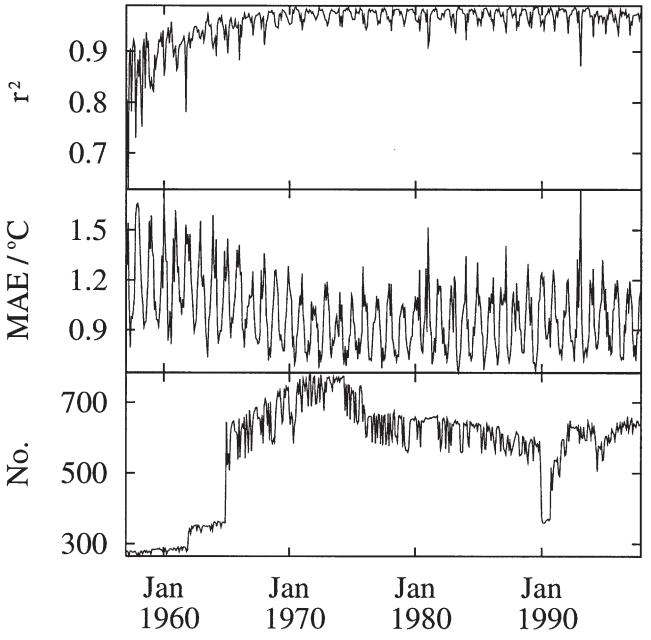

Fig. 10. Cross validated interpolation statistics for daily maximum temperature. All estimates were computed using a 3-D thin plate smoothing spline.

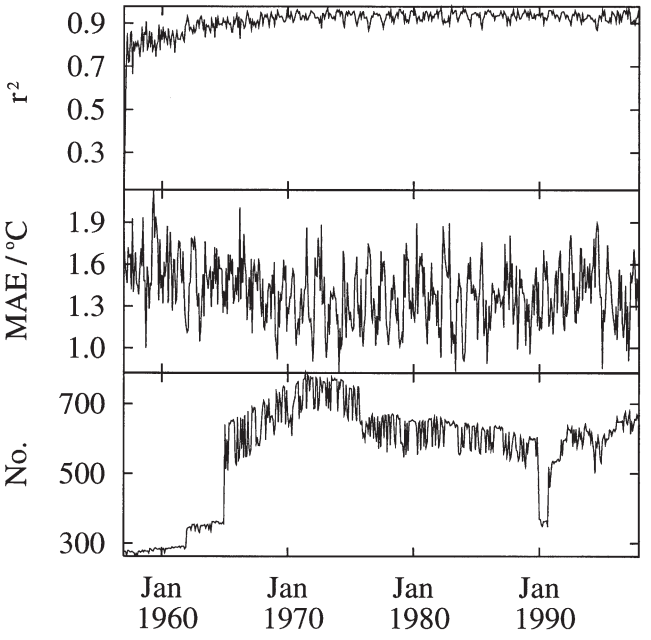

Fig. 11. Cross validated interpolation statistics for daily minimum temperature. All estimates were computed using a 3-D thin plate smoothing spline.

3.3. Rainfall

Spatial and temporal error analyses have been computed for monthly and daily rainfall. Monthly rainfall has been cross validated using a technique which simulates the normalisation procedure outlined in Section 2.2.2. The technique used can be summarised as follows.

1. Means and variances of the chosen fractional power (0.5) of monthly rainfall were calculated for all stations.
2. Cross validated means and variances (from Step 1) were computed using a trivariate smoothing spline, with all data scaled according to its relative variance.
3. The fractional power of monthly rainfall was normalised using the means and variances computed in Step 1.
4. Normalised monthly rainfall was cross validated using ordinary kriging. An exponential variogram was used with zero nugget, sill=2.5, and the range set equal to $1.5\bar{d}$, where $\bar{d}$ is the average distance between

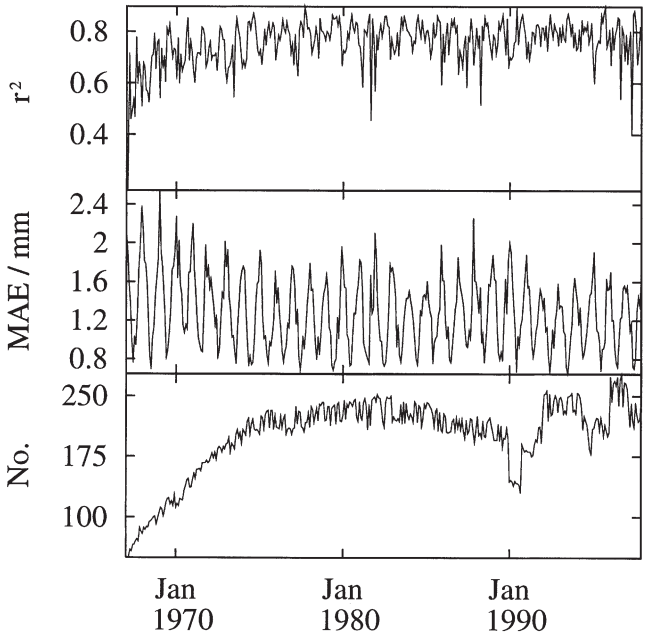

Fig. 12. Cross validated interpolation statistics for daily Class A pan evaporation. All estimates were computed using a 3-D thin plate smoothing spline.

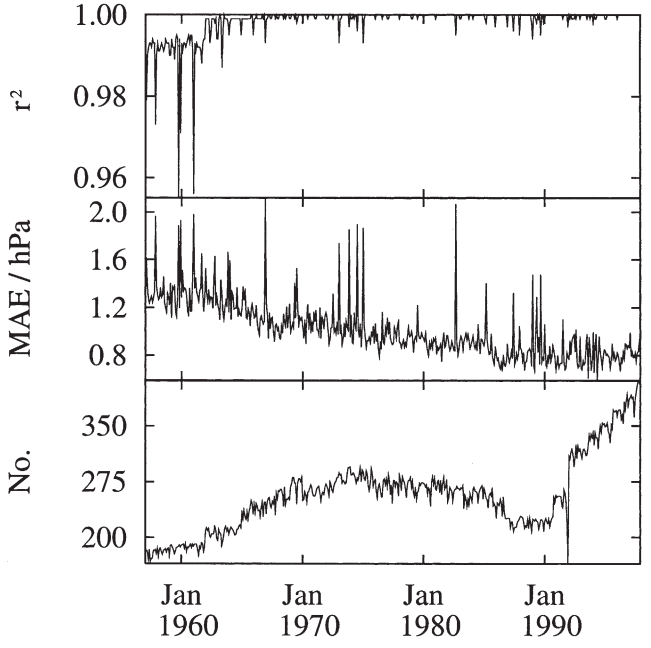

Fig. 13. Cross validated interpolation statistics for daily mean sea level pressure. All estimates were computed using a 2-D thin plate smoothing spline.

the surrounding data points and the location being cross validated.

5. The cross validated monthly rainfall was computed by reversing the normalisation of the cross validated anomaly calculated in Step 4. The cross validated means and variances that were calculated in Step 2 were used to effect the reverse transformation.

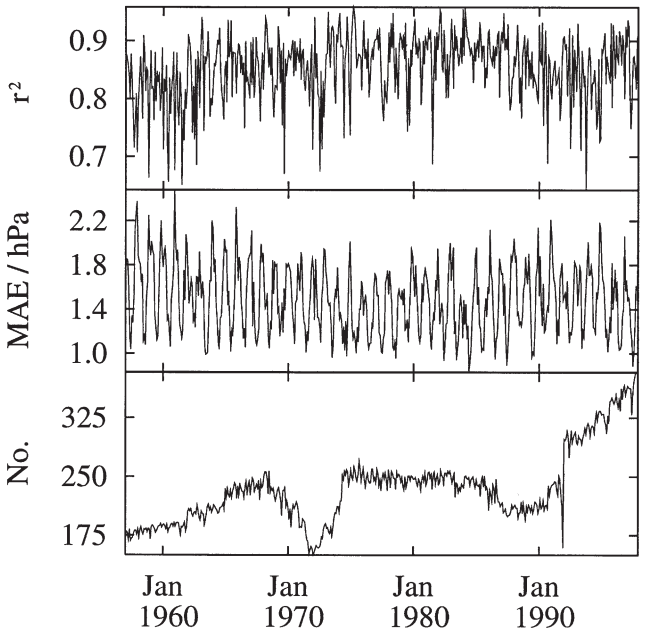

Fig. 14. Cross validated interpolation statistics for daily vapour pressure. All estimates were computed using a 3-D thin plate smoothing spline.

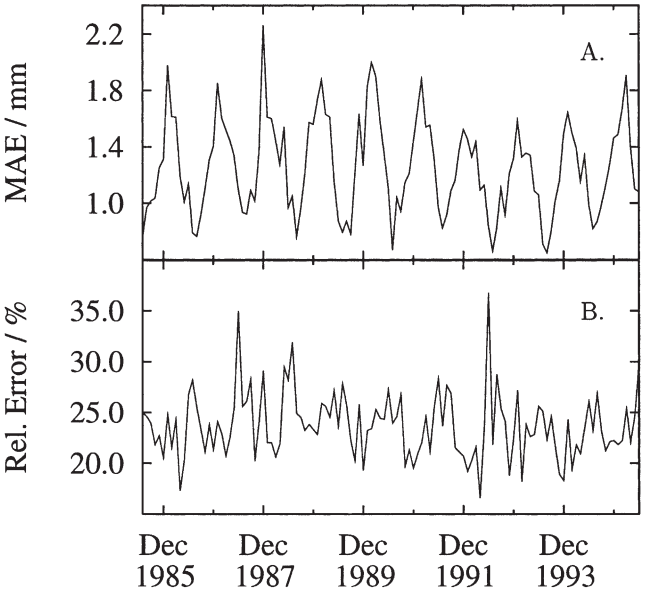

Fig. 15. Mean absolute error (MAE) (A) and relative error (B) between observed Class A pan evaporation data and cross validated estimates.

6. Error statistics were computed using the observed data and the cross validated estimates calculated in Step 5.

The results of the spatial error analyses are shown in Fig. 17 and the temporal results shown in Fig. 18. The temporal results appear to be erratic but this is mainly due to the seasonal variation in the type of rainfall.

Table 4
Summary statistics for the anomaly interpolation of daily climate data. Error statistics were computed by superimposing the interpolated mean and the interpolated residual (obs. data−mean data)

Statistic	Climate variable			
	Max. T.	Min. T.	Evap.	Pres.
1957–1997[a]				
RMSE	1.5°C	1.9°C	1.9 mm	1.6 hPa
MAE	1.1°C	1.5°C	1.4 mm	1.0 hPa
ME	0.0°C	0.0°C	0.0 mm	0.0 hPa
r^2	0.95	0.91	0.73	0.99
1990–1997				
RMSE	1.5°C	2.0°C	1.8 mm	1.4 hPa
MAE	1.0°C	1.5°C	1.3 mm	0.8 hPa
ME	0.0°C	0.0°C	0.0 mm	0.0 hPa
r^2	0.97	0.93	0.74	0.99

[a] Analysis period for evaporation: 1967–1997.

Focusing attention on a reduced time period as in Fig. 19, it is readily seen that the interpolation error increases in summer months due to the onset of convective rain storms. The increased error is exacerbated by the fact that the density of reporting stations is particularly low in the northern part of Australia which experiences monsoonal-type rainfall.

The 12-month mean has been computed to assist interpretation of the error profile shown in Fig. 18. Boxcar filtering effectively removes the seasonal effect and allows one to more effectively interpret the error statistics in the presence of a noisy background.

The spatial results (Fig. 17) show that the areal distribution of errors is highly complex, particularly in areas of detailed topography. As noted earlier, rainfall is strongly influenced by topography and this is evident in the spatial distribution of error. The complex error patterns indicate that the density of reporting stations in the existing gauge network is not sufficiently high to resolve the topographic detail. In some locations it may also be indicative of unresolved errors in the data, such as recording stations being assigned incorrect coordinates.

The spatial and temporal analyses are useful for examining the magnitude of errors in estimated rain amounts, but it is also important to examine the distribution of rainfall amounts. A frequency histogram of observed and cross validated rain amounts is useful in determining if the interpolation algorithm is consistently under- or overestimating rainfall. The histogram shown in Fig. 20

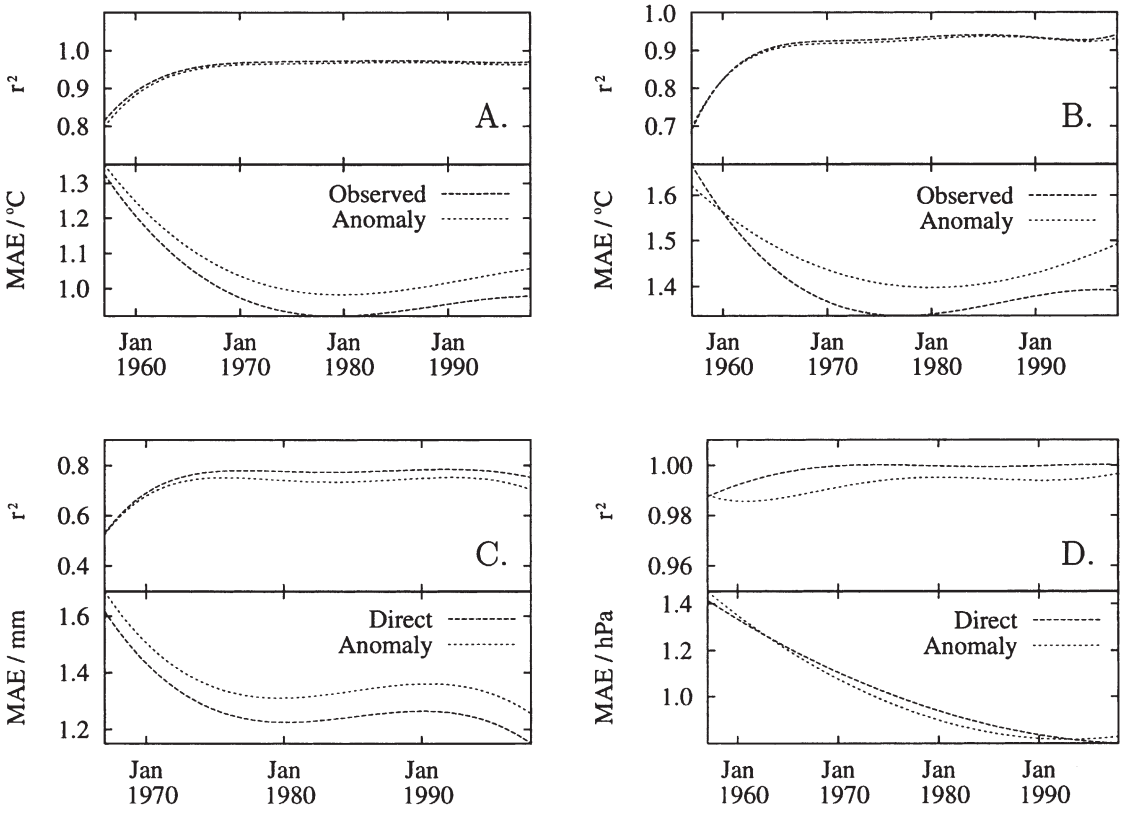

Fig. 16. Comparison of cross validated interpolation statistics for the direct and anomaly interpolation of maximum (A) and minimum (B) temperature, Class A pan evaporation (C) and mean sea level pressure (D). All observed data was interpolated using a 3-D thin plate smoothing spline, with the exception of mean sea level pressure, which utilised a 2-D spline. Anomaly data was interpolated using a 2-D thin plate smoothing spline. Least squares polynomials of order 3 and 5 were fitted to the mean absolute error (MAE) and coefficient of determination (r^2) data, respectively.

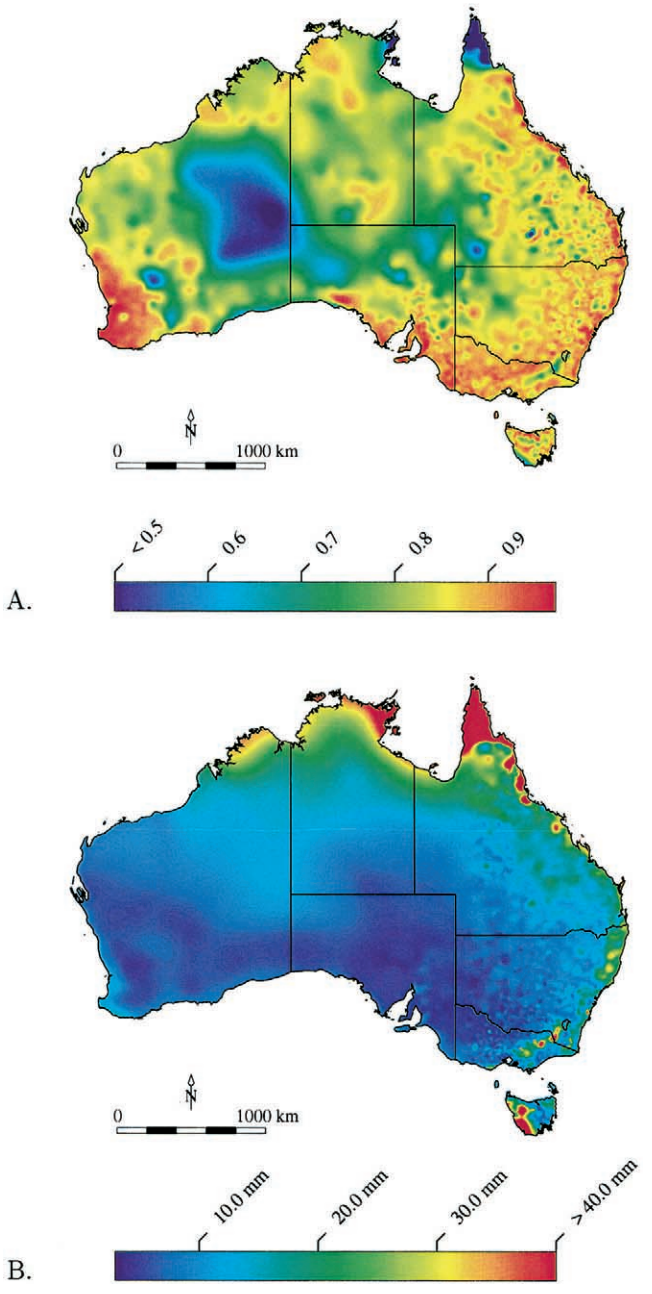

Fig. 17. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed monthly rainfall and cross validated estimates. All estimates were computed by ordinary kriging normalised rainfall using an exponential variogram with nugget=0.0, sill=2.5, range=1.5$\bar{d}$.

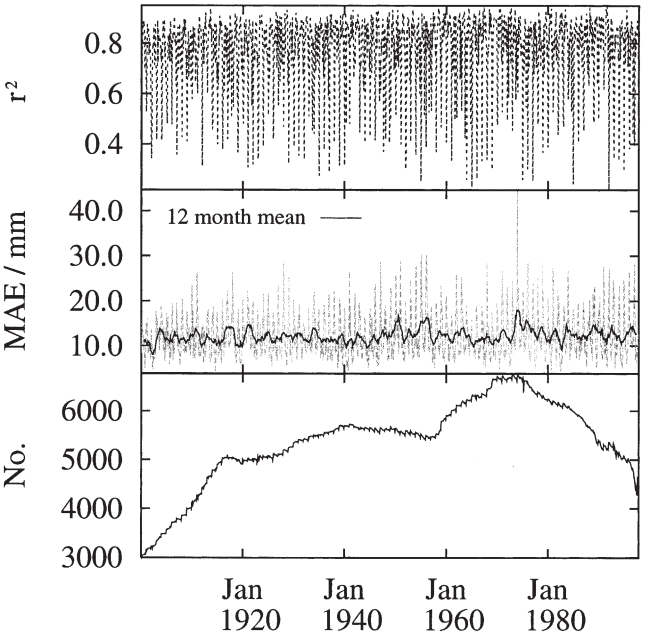

Fig. 18. Cross validated interpolation statistics for monthly rainfall. All estimates were computed by ordinary kriging normalised rainfall using an exponential variogram with nugget=0.0, sill=2.5, range=1.5$\bar{d}$.

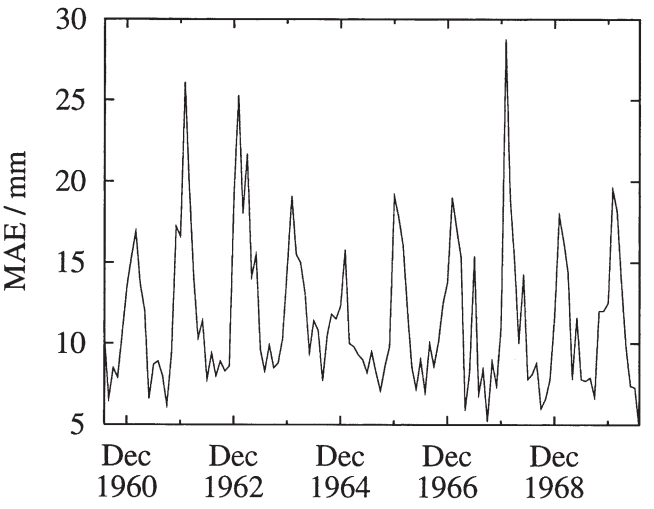

Fig. 19. Mean absolute error (MAE) between observed monthly rainfall data and cross validated estimates. The figure shows the seasonal effect of rainfall type on interpolation skill.

was computed using the observed and cross validated monthly rainfall data for all stations throughout the period 1900–1997. The figure indicates that low rainfall amounts are being accurately conserved, but there is a slight tendency to underestimate larger rainfall amounts.

Summary statistics for monthly rainfall are presented in Table 5. Error estimates such as those in Table 5 can be misleading if the data analysed contain a high proportion of zero rain values. To overcome this potential problem, statistics have been computed for the overall datasets which incorporate both zero and non-zero rain amounts, and also for the subset of stations which reported non-zero rain amounts. As can be seen from Table 5, the presence of zeroes increases the apparent accuracy of the rainfall interpolation.

The results of the spatial and temporal analyses for daily rainfall are shown in Figs. 21 and 22, respectively. As in the case of monthly rainfall, the temporal error statistics are highly variable and dominated by seasonal

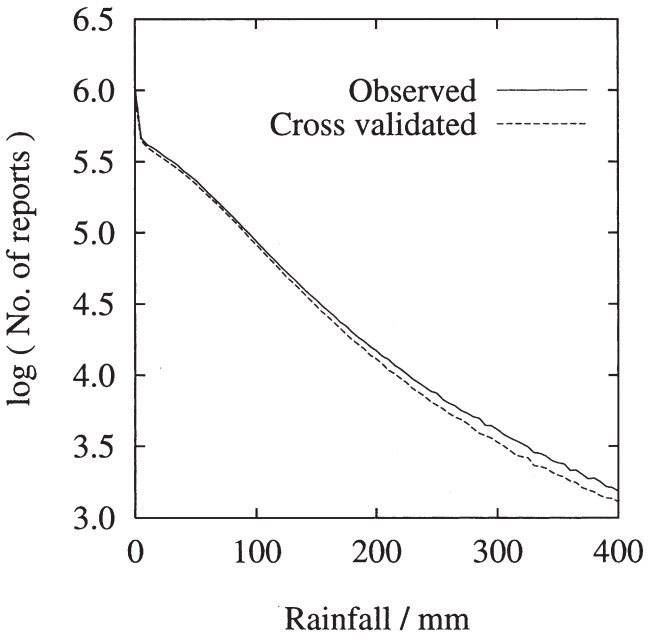

Fig. 20. Frequency histogram of observed and cross validated monthly rainfall data. Histogram computed using Δrain=5 mm. All estimates were computed by ordinary kriging using an exponential variogram with nugget=0.0, sill=2.5, range=2.0$\bar{d}$.

Table 5
Summary statistics for the interpolation of monthly rainfall data

Statistic	Monthly rainfall	
	Rain+no rain	Rain
1900–1997		
RMSE (mm)	35.1	36.3
MAE (mm)	12.2	12.9
ME (mm)	−0.7	−0.5
r^2	0.77	0.77
1990–1997		
RMSE (mm)	36.6	36.9
MAE (mm)	12.5	13.1
ME (mm)	−1.0	−0.7
r^2	0.77	0.77

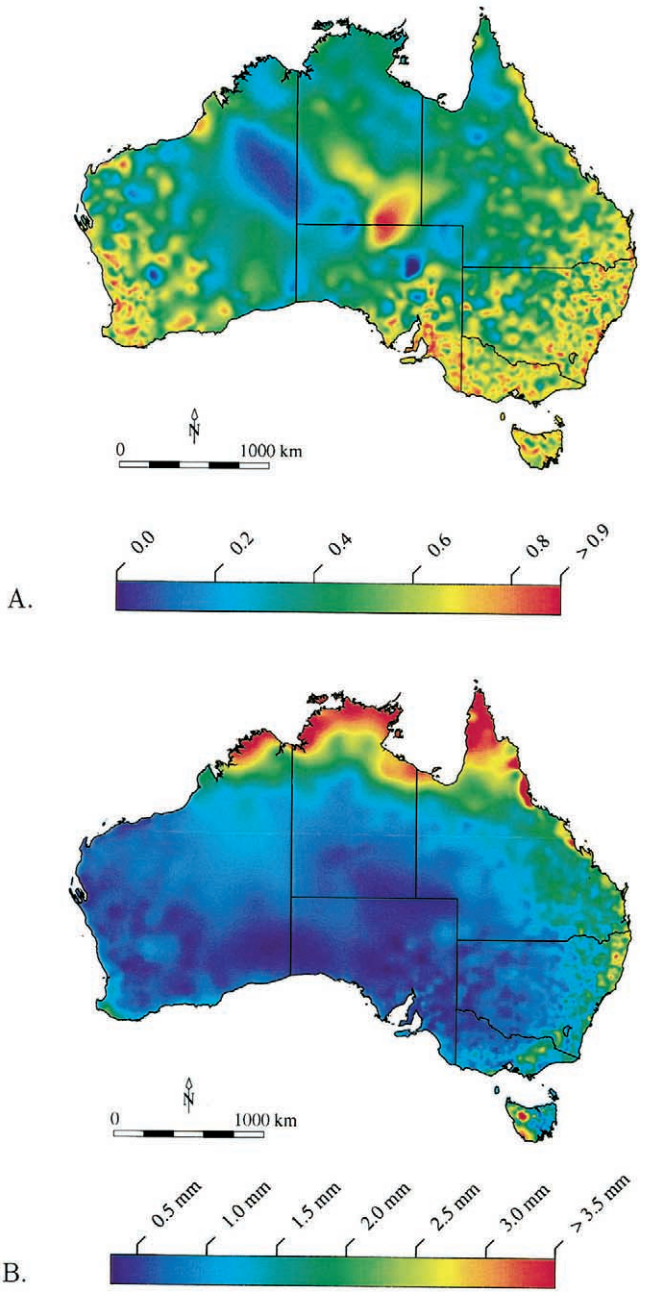

Fig. 21. Average coefficient of determination (r^2) (A) and mean absolute error (MAE) (B) between observed daily rainfall and cross validated estimates. All estimates were computed by ordinary kriging using an exponential variogram with nugget=0.0, sill=2.5, range=2.0$\bar{d}$.

fluctuations. The error statistics presented in Figs. 21 and 22 were computed using ordinary kriging of daily rainfall datasets. Identical error analyses were undertaken for daily rainfall derived by partitioning monthly rainfall onto individual days. Two methods were tested. First, the daily distribution at the cross validation site was assumed to be identical to the distribution at the nearest recording station. Second, the distribution was calculated by computing interpolated estimates of the daily rainfall at the validation site, for every day of the month. The procedure used for computing cross validated daily rainfall estimates from monthly data is summarised as follows.

1. Independently cross validated estimates of monthly rainfall were computed using the algorithm outlined above.
2. The daily distribution of monthly rainfall was computed using the two procedures just described.
3. The cross validated daily rainfall was computed by calculating the proportion of cross validated monthly rainfall that was received on the target date. This pro-

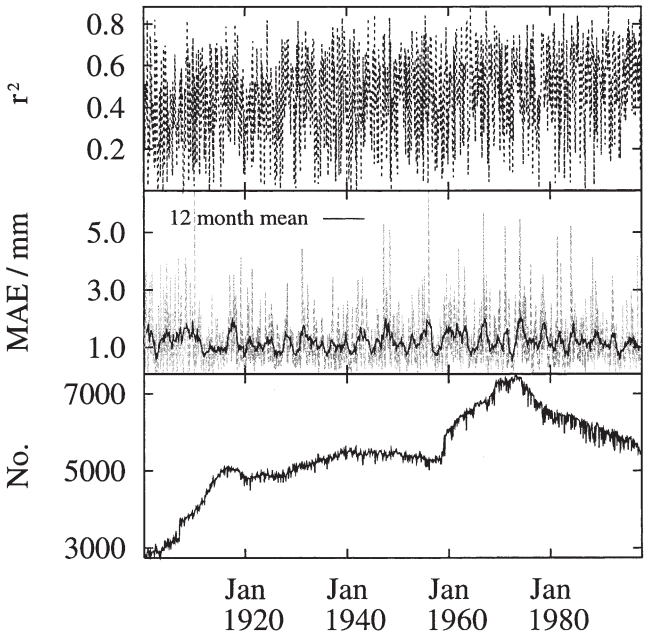

Fig. 22. Cross validated interpolation statistics for daily rainfall. All estimates were computed by ordinary kriging using an exponential variogram with nugget=0.0, sill=2.5, range=2.0$\bar{d}$.

portion was taken from the daily distribution computed in Step 2.

4. Error statistics were computed using the observed data and the cross validated estimates calculated in Step 3. All analyses, including the direct interpolation of daily data used an exponential variogram with zero nugget, sill=2.5, and the range set equal to 2.0$\bar{d}$. It was found that variations in the range factor had no effect on the computed statistics, which supports the observations of Dietrich and Osborne (1991).

Statistics comparing the performance of the three methods for interpolating or deriving daily rainfall data are presented in Table 6. The first algorithm which redistributed monthly rainfall using the daily distribution of the nearest neighbour was found to be least accurate and can be disregarded. The second algorithm, which effectively used daily rainfall interpolations to derive the daily distribution, returned error statistics which were comparable, albeit marginally worse, than the direct method, i.e. interpolating daily rainfall directly. As discussed in Section 2, the patched datasets have been constructed using the first algorithm which redistributes the monthly rainfall according to the daily distribution at the nearest neighbour. In light of the above statistics, the patched datasets will be updated using rainfall data that have been derived by partitioning monthly data according to the daily interpolations. This approach will be adopted in preference to interpolating daily data directly, even though the direct approach may have marginally better error statistics. The rationale for this is two-fold. First, the daily surfaces derived from monthly data will sum to the interpolated monthly total. In contrast, rainfall surfaces generated by direct interpolation of daily data will (in most cases) fail to conserve the interpolated monthly total. Second, monthly rainfall data can generally be assumed to be of higher quality than daily data for the reasons outlined earlier. This is because accumulated totals can be used, and the monthly total can be checked by summing the daily and/or accumulated quantities. This procedure has inherent problems as discussed in Section 2.2.2, but it does provide an additional level of error checking.

As in the case of monthly rainfall, a frequency histogram has been computed to detect systematic over- or underestimating of rain amounts. Fig. 23 indicates low rain amounts are accurately conserved, but as with monthly rainfall, the larger amounts are underestimated. The conservation of small rain amounts is of particular importance for many modelling applications as it indicates that these small values are not being spread into areas of zero rainfall.

Table 6
Summary statistics for the interpolation of daily rainfall data via three different methods

Statistic	Rainfall type					
	Daily (direct)		Derived (nearest)		Derived (neighbours)	
	Rain+no rain	Rain	Rain+no rain	Rain	Rain+no rain	Rain
1900–1997						
RMSE (mm)	3.9	–	4.8	–	4.1	–
MAE (mm)	1.2	–	1.3	–	1.2	–
ME (mm)	0.0	–	0.0	–	0.0	–
r^2	0.46	0.42	0.37	0.34	0.45	0.42
1990–1997						
RMSE (mm)	3.7	–	4.9	–	4.0	–
MAE (mm)	1.1	–	1.2	–	1.1	–
ME (mm)	0.0	–	0.0	–	0.0	–
r^2	0.53	0.49	0.43	0.40	0.51	0.47

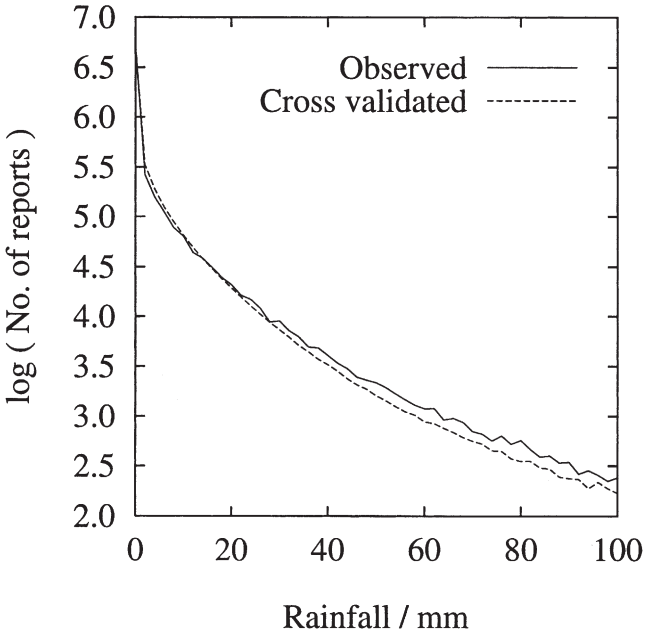

Fig. 23. Frequency histogram of observed and cross validated daily rainfall data. Histogram computed using Δrain=2 mm. All estimates were computed by ordinary kriging using an exponential variogram with nugget=0.0, sill=2.5, range=2.0$\bar{d}$.

4. Discussion

In the previous sections we outlined the procedures and algorithms used to construct a database of observed and interpolated data. The datasets described are in widespread use and have assisted in the development of many research, educational and managerial applications which were previously restricted by the difficulty in obtaining complete and continuous datasets. The AussieGRASS project (Carter et al., 2000) is one such example. Given a comprehensive array of agrometeorological data, AussieGRASS uses a pasture simulation model to estimate various soil, water and plant parameters. The model uses interpolated surfaces for spatial simulation on a continental scale, and the Patched Point Datasets are used for point modelling at discrete locations.

Applications such as AussieGRASS have driven the need for database development, and they also provide the motivation for continued error checking and data analysis. In Section 3 we presented an analysis of the interpolation error to enable users to assess the data quality. While such analyses are of fundamental importance from a user's perspective, they can also highlight difficult issues regarding the integrity of observational datasets. To illustrate this point, we will examine the spatial error maps for mean sea level pressure and daily rainfall.

The mean sea level pressure map (Fig. 8) indicates the region of highest interpolation error is centred about a recording station located in central Australia (Station 013017). Outliers such as this can arise if incorrect coordinates are specified for the station's location. To resolve this, the latitude, longitude and elevation were checked against high resolution maps and subsequently found to be correct. The coordinates of the two nearest stations were also checked, and found to be correct. To determine if the apparent interpolation error was a consistent phenomenon, and not the spurious result of one or two data errors, the interpolation error was examined as a function of time. Fig. 24 shows the difference between the cross validated estimates and the observational data for the analysis period. This figure was constructed from cross-validated estimates for 489 dates and indicates that the interpolation algorithm systematically overestimates mean sea level pressure at the location of the suspect recording station. Since the error is reasonably constant over a period of 41 years, it is unlikely that the error is due to human error in reading the instrument. Consistent errors such as this may be due to a variety of factors, such as inappropriate siting of recording stations, and these problems have been previously documented (Peck, 1997).

The spatial error map computed for daily rainfall (Fig. 21) shows a region in central Australia where the coefficient of determination (r^2) between the observed and cross validated estimates exceeds 0.9. This result appears to be unusually high, given the station density in this region and the lower correlation displayed by the surrounding areas. If this result were due to the presence of a large proportion of zero rainfall reports, the surrounding areas would exhibit a similar pattern. This is obviously not the case. Error statistics computed through cross validation may reveal regions of low interpolation error if stations in the relevant area had duplicated data records. Duplicate records may arise through either operator error when observational data are entered into the database maintained by the Bureau, or it may be due to observer error at recording stations. The data records for

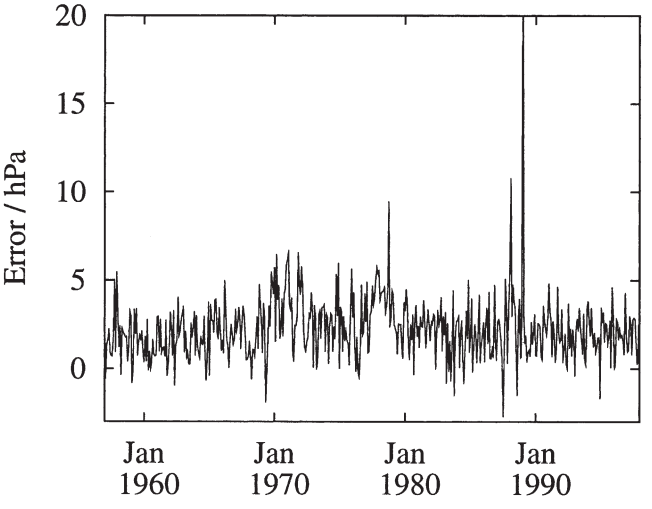

Fig. 24. Cross validated interpolation error for mean sea level pressure data recorded at station 013017. Error statistic shown is (cross valid. est.−obs.).

the three stations contributing to the anomalous region in Fig. 21 were checked for duplicate entries. Observational data were available for approximately 49 years for all three stations, and there were no duplicate records in the entire dataset. The error statistics computed for this region appear to be correct.

The algorithms used for the interpolation of daily and monthly rainfall were presented in Sections 2.2.2 and 3.3. In those sections we noted that the kriging parameters were set to enforce exact interpolation, or in other words, accurate reproduction of the observed data. Exact interpolation of rainfall data is normally considered naive for a number of reasons (Hutchinson, 1993). In particular, by constraining the surface to pass through the data points, one is assuming zero measurement error and the absence of localised microscale variation in the observed field. For rainfall data this is an unrealistic assumption, but one which has been imposed by application restraints (Carter et al., 2000). The introduction of alternative algorithms for interpolating rainfall data will be implemented as resources permit. This will then allow database clients to select the rainfall interpolation scheme best suited to their requirements.

5. Conclusion

A comprehensive archive of Australian rainfall and climate data has been described. The archive was constructed to facilitate research and managerial tasks requiring hydrometeorological data. Prior to the availability of such a database, individuals requiring data had to expend considerable resources to develop their own datasets from observational data. This task is both time consuming and in many cases made difficult by the fact that climate data may never have been recorded at a site within an acceptable distance of the desired location. These problems have been overcome by the development of an extensive network of patched and gridded datasets which are publicly available on the Internet. Interested readers should consult the URL http://www.dnr.qld.gov.au/silo for further information.

High resolution gridded surfaces were generated by spatial interpolation of observed daily data. Patched datasets were constructed at a set of point locations using the available observational data. Where observed data were unavailable, interpolated estimates were substituted. The use of interpolated data necessitates an accurate estimate of the interpolation error. Spatial and temporal error analyses were presented which will assist database clients in assessing the reliability of results derived from the data.

The data type and time period supported by the database have been summarised in Table 1, and the locations of recording stations with patched datasets are shown in Fig. 1. While the database currently contains data relevant only to Australia, the methodology used in its construction could be readily adapted to other countries.

Acknowledgements

The authors gratefully acknowledge the Australian Bureau of Meteorology for the provision of climate data. This work was supported by the Queensland Department of Natural Resources, the Land and Water Resources Research and Development Corporation and the Rural Industries Research and Development Corporation. Many helpful discussions with Dr Mike Hutchinson have been of great benefit, for which we are most grateful. The authors would like to thank Ken Day and Dr Graeme Hammer for assistance in proofreading the manuscript.

References

Bennett, R.J., Haining, R.P., Griffith, D.A., 1984. Review article: The problem of missing data on spatial surfaces. Annals of the Association of American Geographers 74, 138–156.

Bohren, C.F., Albrecht, B.A., 1998. Atmospheric Thermodynamics. Oxford University Press, New York.

Carter, J., Flood, N., Danaher, T., 1996. Development of data rasters for model inputs. In: Brook, K. (Ed.), Development of a National Drought Alert Strategic Information System, vol. 3. Queensland Department of Natural Resources. Final report on QPI20 to LWRRDC.

Carter, J.O., Hall, W.B., Brook, K.D., McKeon, G.M., Day, K.A., Paull, C.J., 2000. Aussie GRASS: Australian grassland and rangeland assessment by spatial simulation. In: Hammer, G., Nicholls, N., Mitchell, C. (Eds.), Applications of Seasonal Climate Forecasting in Agricultural and Natural Ecosystems — the Australian experience. Kluwer Academic Press, Netherlands, pp. 329–349.

Colquhoun, J.R., 1965. A limited test of methods of pressure reduction to mean sea level. Australian Meteorological Magazine 49, 28–39.

Creutin, J.D., Obled, C., 1982. Objective analyses and mapping techniques for rainfall fields: an objective comparison. Water Resources Research 18, 413–431.

Dedieu, G., Deschamps, P.Y., Kerr, Y.H., 1987. Satellite estimation of solar irradiance on the surface of the Earth and of surface albedo using a physical model applied to Meteostat data. Journal of Climate and Applied Meteorology 26, 79–87.

Dietrich, C.R., Osborne, M.R., 1991. Estimation of covariance parameters in kriging via restricted maximumlikelihood. Mathematical Geology 23, 119–135.

Hounam, C.E., 1961. Evaporation in Australia. A critical survey of the network and methods of observation together with a tabulation of the results of observations. Technical Report, Bureau of Meteorology, Bulletin 64.

Hutchinson, M.F., 1993. On thin plate splines and kriging. In: Tarter, M.E., Lock, M.D. (Eds.), Computing and Science in Statistics, vol. 25. Interface Foundation of North America, University of California, Berkeley, pp. 55–62.

Hutchinson, M.F., 1995. Interpolating mean rainfall using thin plate smoothing splines. International Journal of Geographical Information Systems 9, 385–403.

Hutchinson, M.F., Richardson, C.W., Dyke, P.T., 1993. Normalisation of rainfall across different time steps. In: Management of Irrigation

and Drainage Systems, vol. 9. Irrigation and Drainage Division, ASCE, US Department of Agriculture, pp. 432–439.

Hsu, K-L., Gao, X., Sorooshian, S., Gupta, H.V., 1997. Precipitation estimation from remotely sensed information using artificial neural networks. Journal of Applied Meteorology 3, 1176–1190.

Hsu, K-L., Gupta, H.V., Gao, X., Sorooshian, S., 1999. Estimation of physical variables from multichannel remotely sensed imagery using a neural network: Application to rainfall estimation. Water Resources Research 35, 1605–1618.

Isaaks, E.H., Srivastava, R.M., 1989. Applied Geostatistcs. Oxford University Press, New York.

Kuligowski, R.J., Barros, A.P., 1998. Using artificial neural networks to estimate missing rainfall data. Journal of the American Water Resources Association 34, 1437–1447.

Lam, N.S-N., 1983. Spatial interpolation methods: A review. The American Cartographer 10, 129–149.

Letestu, S. (Ed.), 1973. International Meteorological Tables, 3rd ed. Meteorological Organization, Geneva.

List, R.J. (Ed.), 1966. Smithsonian Meteorological Tables, 6th ed. Smithsonian Institution, Washington.

Manual of Barometry. United States Weather Bureau, Washington.

Meinke, H., 1996. Improving wheat simulation capabilities in Australia from a cropping systems perspective. PhD thesis, van de Landbouwuniversiteit te Wageningen.

Mills, G., Weymouth, G., Jones, D., Ebert, E.E., Manton, M., Lorkin, J., Kelly, J., 1997. A National Objective Daily Rainfall Analysis System. Bureau of Meteorology Research Centre, Melbourne, Australia.

Peck, E.L., 1997. Quality of hydrometeorological data in cold regions. Journal of the American Water Resources Association 33, 125–134.

Richardson, C.W., 1977. A model of stochastic structure of daily precipitation over an area. Colorado State University Hydrology Paper No. 91, 45 pp.

Stidd, C.K., 1973. Estimating the precipitation climate. Water Resources Research 9, 1235–1241.

Tanner, C.B., Sinclair, T.R., 1983. Efficient water use in crop production: Research or Re-Search? In: Taylor, H.M., Jordan, W.R., Sinclair, T.R. (Eds.), Limitations to Efficient Water Use in Crop Production, Chapter 1A. American Society for Agronomy, pp. 1–27.

Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia.

Wahba, G., Wendelberger, J., 1980. Some new mathematical methods for variational objective analysis using splines and cross validation. Monthly Weather Review 108, 1122–1143.