

## AN ANALYSIS OF 1990-2011 ONTARIO SURFACE AIR TEMPERATURES

DAVID R. BRILLINGER<sup>†</sup>

*Statistics Department, University of California, 367 Evans,  
Berkeley, CA 94720-3860, USA*

The paper's concern is the estimation of the average monthly temperature across a given region as a function of time. The region studied here is the Canadian province of Ontario and the time unit is month. Data for various stations and times were obtained from the Berkeley Earth website, ([www.berkeleyearth.org](http://www.berkeleyearth.org)). In the work a generalized additive model with random effects is employed that allows both spatial and temporal dependence. Handling variability in both space and time is basic.

### 1. Introduction

This study uses historical data to analyse Earth's surface temperatures as functions of time,  $t$ , and space,  $(x,y)$ . The particular focus is the monthly averages for the Canadian province of Ontario. These particular data evidence a wide variety of the complications common to surface temperature studies. The data employed were obtained from the Berkeley Earth project [2, 12, 13],

The temperatures are measured at stations in a regular temporal fashion and typically there are many temperature values at the same  $(x,y)$ . However these locations are distributed irregularly. Further, stations enter and leave as time passes so there are starts and stops and gaps in the temporal sequences. This may be seen in the right hand panel of Figure 1 below where the measurement times are shown as related to their latitude. The left hand pane provides the geographic locations of the stations. One sees that the measurements get scarcer as one moves north in the province.

Biases may be anticipated to be present in the results of equally weighted analyses if the spatial and temporal irregularities are not dealt with. The temporal case will be handled here, for now, by restricting consideration to stations not missing too many values. The spatial aspect will be dealt with by introducing

---

<sup>†</sup> Work partially supported by NSF grant DMS-100707157.

weights, a continuous form of poststratification. Generalized additive models with random effects are employed.

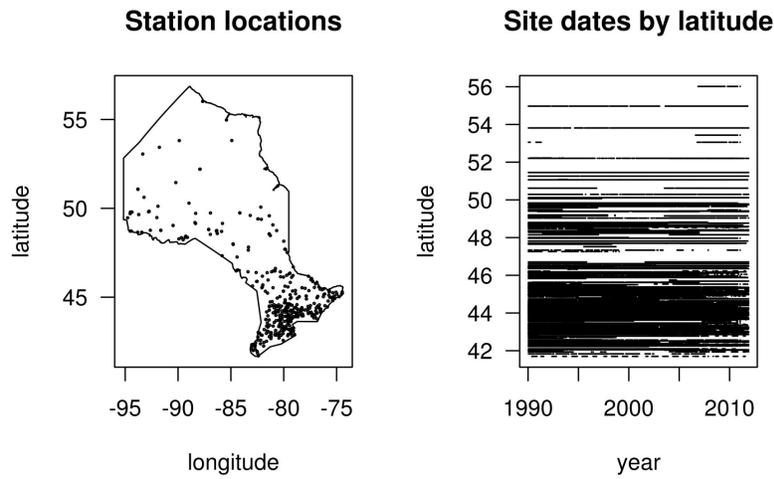


Figure 1. The left panel show the locations of 1990-2011 stations. The right panel highlights the stations' dates and related latitude.

The statistical package R, [11], is employed in the analyses throughout. The R-libraries employed included `mgcv`, `spatialkernel`, `RandomFields` and `splancs`. The library `mgcv` works with regression splines. It includes smoothing parameter estimation by cross-validation.

## 2. The data.

Figure 1 may be interpreted in general terms. Consider the left hand panel. The bottom and left hand side border on water, the Lakes Huron, Erie and Ontario. As one moves north in the province the temperature gets cooler. The winters and summers can both be extremely hot. The circles in the plot are the locations of the stations that existed at some time in the period 1990-2011. There are 440 of these. The panel shows a much greater density of stations in the south. The right panel is consistent with this. In addition it shows the temporal gaps in coverage of the province generally and the station monthly values specifically.

For this time period there are 264 data months, however December 2011 was missing throughout. A decision was taken to exclude any station missing more than 20 months at this stage.

To deal with the remaining missing values an elementary analysis of variance model was employed, specifically

$$Y_{jk} = \alpha_j + \beta_k + \varepsilon_{jk} \quad (1)$$

with  $j$  indexing years and  $k$  months. It was and fit by least squares. This provided estimates of the December 2011 values in particular. Here  $Y$  is temperature,  $\alpha$  a year effect,  $\beta$  a month effect and  $\varepsilon$  an error. The locations remained irregularly spread, as may be seen in Figure 4 to come. The effects of this will be ameliorated by employing weights, a form of continuous post-stratification, The fit was by simple least squares and the estimates  $a_j + b_k$  then employed as needed. This led to a complete array of 264 by 81 "temperatures". The 81 cases will be referred to as the basic subset of stations. The 264 time points were all spaced 1 month apart

The data for one of the stations, Toronto Centre, is provided in the top panel of Figure 2. The dominant feature in the graph is the annual 12 month seasonal. It provides the vast majority of the variability in the series. The monthly average temperature may be seen to range from about -25 degrees centigrade to about +20 during the period 1990 to November 2011.

### 3. Models and preliminary analyses. .

The model (1) provided an elementary method to infer missing values. Often one seeks regular arrays of data as various statistical methods have such in mind, however more complex methods could be employed. The missing time observations were dealt with from the start. Figure 1 right hand panel shows the importance of this. At the outset stations with more than 20 months missing were excluded

As an illustrative analysis we consider the 1990 to 2011 data for Toronto Centre with the December 2011 value filled in. Next consider the following stochastic model

$$Y(t) = \alpha(\text{year}) + \beta(\text{month}) + \varepsilon(t) \quad (2)$$

where year and mon (month) are factors and  $\varepsilon(\cdot)$  is a stationary autoregressive series of order 1. This is the model Wood considers in Section 6.7.2 of his book [14]. He provides R-commands and results for an analysis of daily Cairo, Egypt temperature for a stretch of 3780 days of data starting 1 January 1995. He employs the R-library mgcv and presents a variety of results. The fitting procedures employed are based on an assumption of gaussian  $\varepsilon$ . The series  $\{\varepsilon(t)\}$  is assumed to be a stationary AR(1) and is treated as an additive random effect

in the model. The functions  $\alpha$  and  $\beta$  are assumed to be smooth. The actual fitting is via the R-function `gamm(.)` employing maximum likelihood. The functions  $\alpha$  and  $\beta$  are represented as splines. Estimation of numbers and locations of knots is included in the function `gamm(.)`.

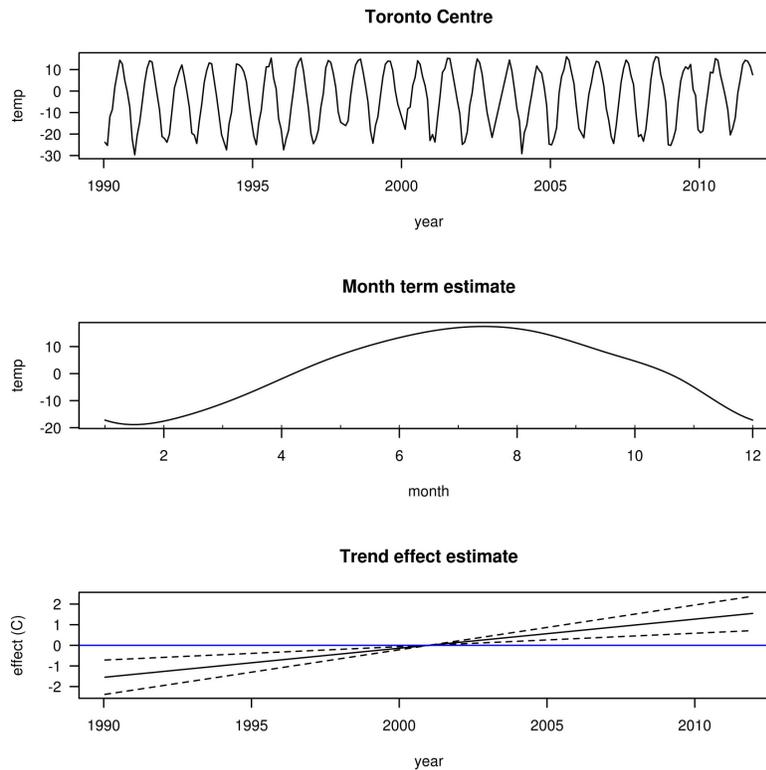


Figure 2. Top panel: Toronto Centre monthly temperatures. Middle panel: estimated seasonal effect Bottom panel: Toronto trend effect estimate and associated uncertainties..

Some results of carrying out the fitting of model (2) are shown in Figure 2. The top panel, already discussed, provides the data themselves. The seasonal effect is clear and meant to be handled by  $\beta$ . The middle panel provides its estimate. That effect is well known and understood to a degree. The quantity of topical interest is  $\alpha$ , where climate change might lurk. It is interesting that the estimate of  $\alpha$  here is approximately linear in time. The bowtie lines in the figure are one way to display 2 standard error variations. It has in mind prediction

purposes. One of the outer limits of the bounds is almost on top of the zero line. There remain the data from the other stations to possibly clarify the situation.

#### 4. Results.

In the next analyses trend effect terms are estimated for each of the 81 stations and then the trends merged. These trend curves are plotted in Figure 3. The curve for Toronto Centre appearing in Figure 2 is one of them. There is considerable scatter but a central core of curves heading from the lower left to the upper right.

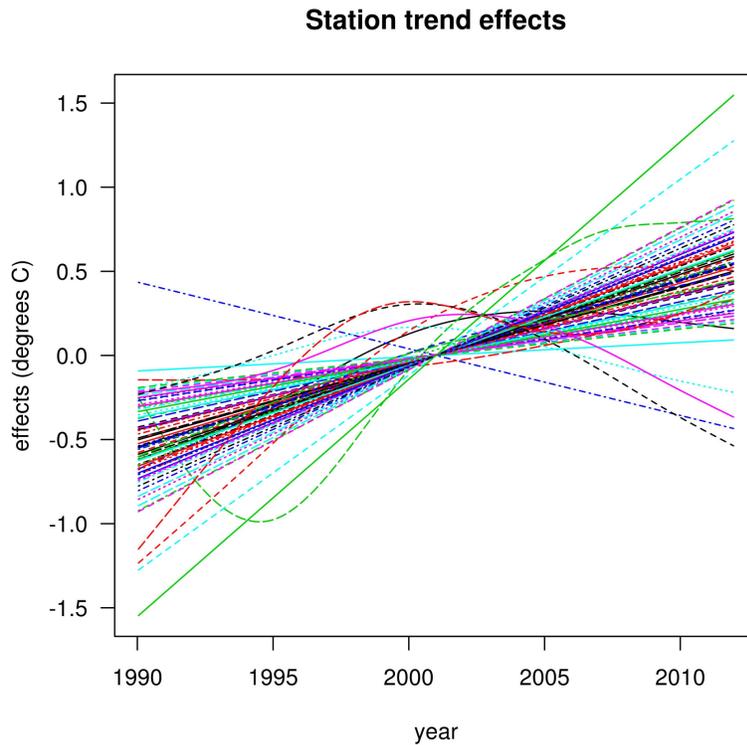


Figure 3. The 81 individual estimated station trends effects superposed

To make inferences one needs measures of uncertainty and these need to take note of temporal and spatial variability. In the figure one sees outliers highlighting the need for a robust/resistant method of merging the 81 curves

The next step is to average the values at each time point, however spatial bias is a serious concern. One debiasing approach is to weight the stations in a region inversely to the density of stations there. Pertinent weights were determined as follows: the locations of the stations are viewed as a planar point process and its intensity function estimated by kernel smoothing. This was realized by the function "lambdahat" of the R-library "spatialkernel". If the estimated density at the location  $(x,y)$  is  $d(x,y)$  the weight at  $(x,y)$  is  $w(x,y) = 1/d(x,y)$ . For the data concerned the array is not changing for the time period of interest. The rates are displayed in Figure 4 via the diameter of circles which are centered at the various array locations.. As anticipated the weights get larger as one moves to the north of the province where stations are scarce.

**Weights - subset of 81 stations**

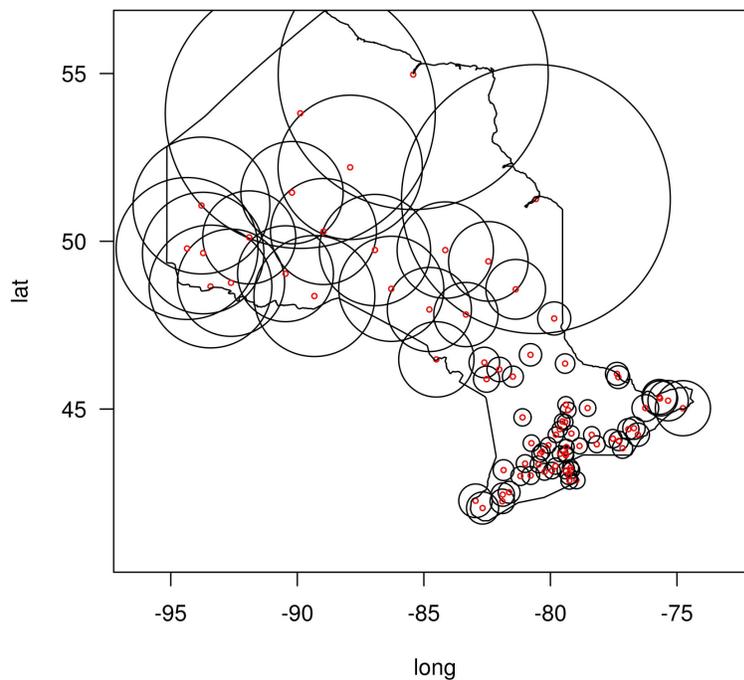
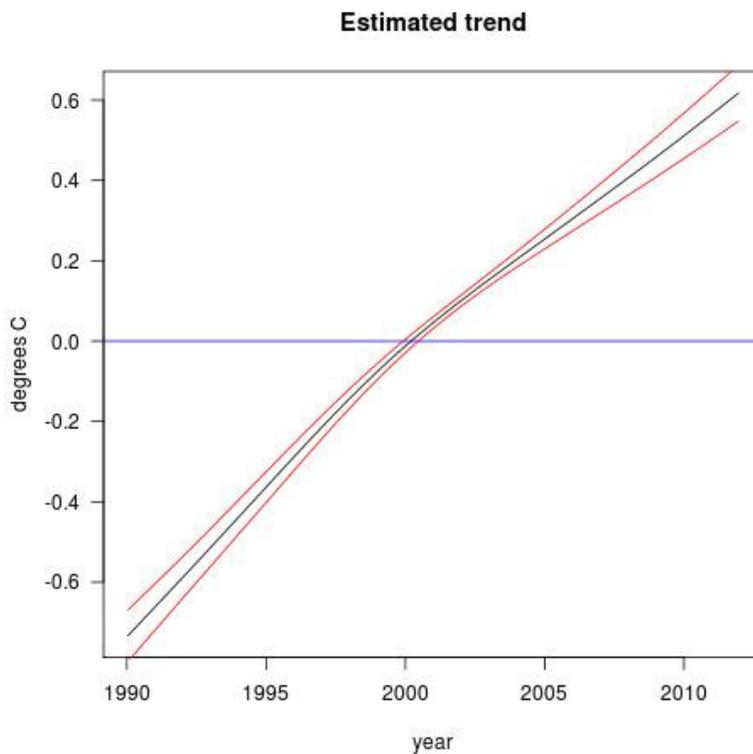


Figure 4. The weights employed in the merging work. There is one for each of the 81 stations station. The diameter of the circle is proportional to the density at the station. Also shown are the locations of the observation stations in Ontario..

For the moment a simple weighted average of the 81 estimated trend effects is employed. The weights are those of Figure 4 and the resulting curve is the central one in Figure 5. Assuming the 81 values at a given time point are statistically independent there is the classic formula for the variance of a weighted mean and an estimate for it. The bounding curves in the Figure are the approximate  $\pm 2$  s.e. limits.



**Figure 5.** Individual station trend effects with weights inversely proportional to station density. There are 2 s.e. bounds.

Looking at Figure 5 it is hard not to infer that there is an increasing trend for this province and time period. However serial dependence has to be a concern as

it may be affecting the estimated standard error. Perhaps the AR(1) didn't go far enough. However there is a multiplier that may be motivated and applied to standard errors after a least squares fit to take some note of remaining temporal dependence. In the present case it is estimated as 1.4015. The source of this number is the average of the first 10 periodogram values divided by the average of all, see for Example p. 127 in Hannan [8]. Including it will not change Figure 5 a great deal..

### 5. Model assessment.

The library `mgcv` has a function that provides results for some common methods of assessing fit, in the case of independent observations. It leads to Figure 6 here.

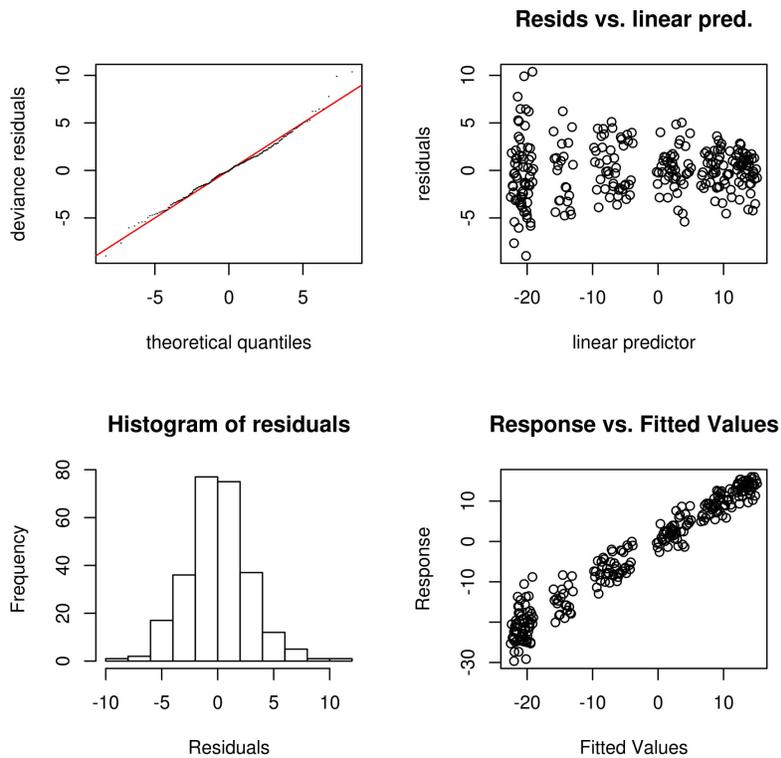


Figure 6. The output produced by the `mgcv` function `gam.check`.

On examining these statistics, the marginal gaussianity is suggested in the two left hand panels. This is important because the fitting and inference techniques employed are motivated by that assumption. The bottom right hand panel suggested that the responses are approximately linearly related to the fitted values. That would relate to how much of the variability of the temperatures was due to seasonal variability. The top right panel gives reason for further research. It reflects that the variability around the fit is greater at the low temperatures. One could deal with this by introducing further weights, but that is for future research.

Autocorrelation in the series  $\varepsilon$  of model (2) needs to be studied. The assumption of the independence of the error term,  $\varepsilon$ , values for the curves in Figure 3 entered into the computation of the standard error bounds. Even though these bounds are marginal temporal dependence is a concern in their derivation.

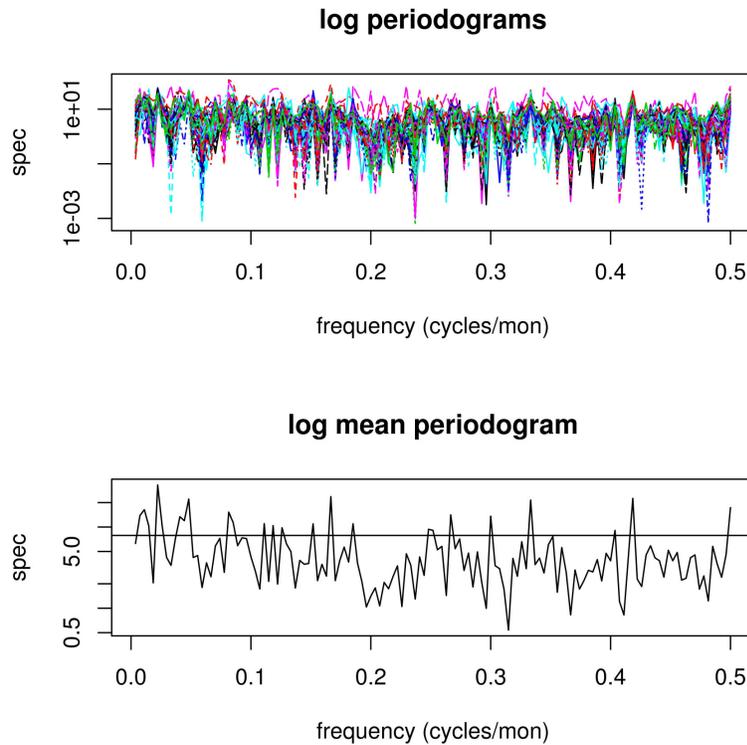


Figure 7. The top panel shows 81 periodogram replicates. The lower provides the average of the 81. Both figures are graphed on "log paper".

Assessing temporal independence of time series residuals is a classic time series problem. Spectrum analysis is one way to address this issue. Figure 7 shows the periodograms of the 81 individual residual time series and of their averaged series. Generally there appears to be some slopping down as one moves from the left to right. The AR(1) term was meant to handle this. This occurrence motivates the estimation of a multiplier

The bottom curve of Figure 7 shows the average of the periodograms in the top panel. The horizontal line is as reference level. The plot is on log paper. There has to be some consideration of spatial dependence. The variogram is a classic parameter to study. An isotropic form will be estimated here. It may be considered an average of a non-isotropic form. Figure 8 provides the results obtained via the R-function "EmpiricalVariogram" of the R-library RandomFields. This particular function accepts replicates. There are two panels in the figure, one for the residual series and the other for the temperature series. The y-axis scales in the two are quite different. This goes along with the inclusion of a monthly term in the model reduces the variability greatly.

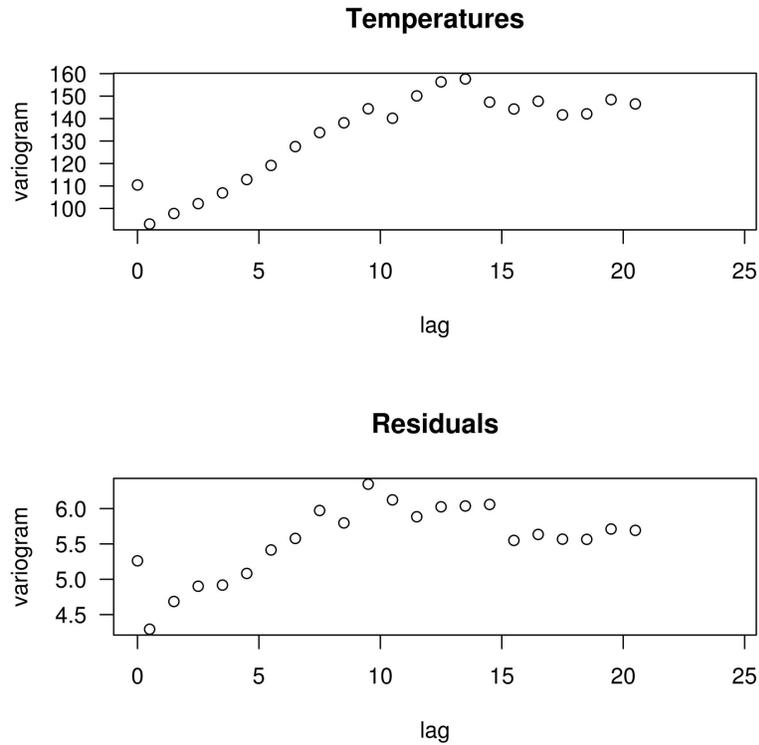


Figure 8. Estimated variograms, assuming isotropy, for the temperatures and residual series respectively. The individual date series are treated as replicates.

...

Unfortunately no indication of uncertainty is provided. One might use some form of simulation or seek a multiplier, as in the temporal case. Figure 8 suggests that spatial dependence extends some distance, e.g. up to lag 12. There was no use of utm coordinates, or of some other coordinate system in terms of km, so the lag units are confused. With the isotropic condition the correlation surface is ellipsoidally shaped. There is also the outlier at lag 0. It could be an artifact of the assumption of isotropy, or result from a mixture of populations being present.

Lastly, one of the referees emphasized that the seasonal shape would be changing as one moved from the south to the north of the province. This effect has been dealt with in a sense by working with the individual station series.

## **6. Discussion/Conclusions.**

This paper has presented an approach. Various limitations have turned up during its development. It is intended to address these in future work.

Time series development has often involved taking a statistic developed for the independent case and seeing its properties under a time series model. One can point to the work of Anderson [1], Grenander and Rosenblatt [7], Hannan [8], with mean levels given by regression models and their study of the OLS estimate in the presence of stationary errors. Spatial-temporal dependence does appear to exist. If wished Gauss-Markov methods could be invoked to increase the efficiency of the estimates and to obtain more appropriate standard error values.

In practice the approach to studying spatial-temporal dependency has been to employ a model including some at the mean level and then check residuals for dependency hoping they turn out to be close to white noise. Here AR(1) modelling was introduced at the outset, following the result in Wood [14]. Yet temporal dependency beyond that appeared to exist. In consequence a "Correction factor" was introduced. In the spatial case weights were introduced to attempt to reduce the spatial dependency, but an estimated variogram suggested that such dependency remained.

There are limitations of the present analysis. A key one is the non-general use of robustness methods. However a small early study of introducing biweights was carried out and the resulting figures did not change a lot. It needs also to be noted that temperature series from neighboring provinces/states were not included in the analyses and logically should be. Other explanatory, such as topography and station elevation might lead to improvements of the estimates. so too might interaction terms. One can mention month with latitude,

Other substantial contributions to spatial-temporal data modelling and analysis appear in: Cressie and Wikle [4], Dutilleul [5], Gelfand et al., [6]. and Le and Zidek [10], Craigmile and Guttorp [3], Jones and Vecchia [9].

In conclusion, part of the intention of preparing this paper was to alert the statistical community to the presence of Berkeley Earth. Its data and the methods they are developing are going to provide a goldmine for statistical researchers. Members of that team have begun to exploit the data and preprints on the website, [www.berkeleyearth.org](http://www.berkeleyearth.org). This present paper shows that some "shovel ready" statistical software is available to employ. Having said that please note that this paper is not to be seen as a Berkeley Earth report

### **Acknowledgments.**

I thank the Berkeley Earth Team, in particular R. Muller for inviting me to be involved, R. Rhode for developing and providing the data set employed this paper, C. Wickham setting me up to use the statistical package R for some initial analyses..

Others to thank include: M. Rooney who provided me with a listing of an R-program he had developed for spatial temporal data, H. Preisler of the US Forest Service for important statistical discussions, Roger Byre of the Berkeley Department of Geography, R. Lovett, P. Spector, and R. Spence of the Berkeley Department of Statistics assisted with the data management. The Referees' gentle but firm comments altered the analyses performed substantially.

### **References**

1. T. W. Anderson. The Statistical Analysis of Time Series. Wiley, New York. (1958).
2. Berkeley Earth, [www.berkeleyearth.org](http://www.berkeleyearth.org), (2012)
3. P. F. Craigmile and P. Guttorp, Space-time modelling of trends in temperature series. *J. Time Ser. Anal.* 32,378-395 (2011).
4. N. Cressie and S. H. Holan. Special Issue: Time Series in the Environmental Sciences, *J. Time Series Analysis* 32, 337-338 (2011).
5. P. Dutilleul. Spatio-Temporal Heterogeneity. Cambridge, Cambridge. (2011).
6. A. E. Gelfand, P. J. Diggle, M. Fuentes and P. Guttorp,. Handbook of Spatial Statistics. CRC Press, Boca Raton (2010).
7. U. Grenander and M. Rosenblatt, Statistical Analysis of Stationary Time Series Wiley, New York (1957).
8. E. J. Hannan, Time Series Analysis. Methuen, London (1960).
9. R. H. Jones and A. D. Vecchia Fitting continuous ARMA models to unequally spaced spatial data. *J. Amer. Stat. Assoc.* 88, 947-954 (1993).

10. N. D. Le and J. V. Zidek. Statistical Analysis of Environmental Space-Time Processes. Springer, New York. (2006).
11. R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org>, (2012).
12. R. Rohde, J. Curry, D. Groom, R. Jacobsen, R. A. Muller, S. Perlmutter, A. Rosenfeld, C. Wickham, J. Wurtele. <http://berkeleyearth.org/pdf/berkeley-earth-santa-fe-robert-rohde.pdf> (2011)
13. R. Rohde, J. Curry, D. Groom, R. Jacobsen, R. A. Muller, S. Perlmutter, A. Rosenfeld, C. Wickham, J. Wurtele. Berkeley Earth temperature averaging process. <http://berkeleyearth.org/pdf/berkeley-earth-averaging-process.pdf> (2011).
14. S. N. Wood, Generalized Additive Models. Chapman & Hall, (2006).