Spatio-Temporal Analysis Of Climatic Data Using Additive Regression Splines

Sharples, J.J. and M.F. Hutchinson

Centre for Resource and Environmental Studies, Australian National University,
E-Mail: jasons@cres.anu.edu.au

Keywords: *Spatio-temporal analysis; smoothing spline; rainfall; pan evaporation; climate change*

EXTENDED ABSTRACT

Environmental modelling often requires knowledge of the values of certain climatic variables at locations where no such information is available. When this is the case, one must rely on interpolated values derived from climatic data recorded at surrounding locations. The accuracy of the interpolated field, however, can often be critically dependent on the inclusion of additional predictor variables in the data models used to calculate the interpolated values. This is the case with precipitation, for example, as it is often influenced by the underlying topography. When interpolating precipitation data it is therefore desirable to include predictors such as elevation and topographic slope and aspect, in addition to those quantifying the data point locations, to achieve accurate precipitation surfaces. Furthermore, studies have shown that interpolation accuracy is improved by allowing for a spatially varying dependence on these topographic variables. Additional predictors will also be appropriate when analysing temporal trends in climatic data. Interpolation procedures that incorporate additional predictors in a spatially varying way can also be useful tools for analysing how the effects of certain predictors vary across the spatial extent of the region under consideration.

While it may be desirable to include several additional predictor variables in a data model, there are practical constraints that limit the feasibility of such an approach. A common problem that arises when analysing multivariate data is that interpolation methods are limited by the fact that estimating a d-variate function with no constraints on its structure, apart from smoothness, requires data sets of impractical size for larger values of d; a fact referred to as the *curse of dimension*. Consequently, interpolation based on higher dimensional data can be numerically expensive or completely impractical.

In many cases the interpolated surface required for application is two- or three-dimensional. This being the case, unconstrained interpolation based on higher dimensional data can produce more elaborate dependencies on the predictor variables than actually needed. It is therefore natural to employ a data fitting method that allows the incorporation of multiple predictors but bypasses the curse of dimension by identifying only the underlying two- or three-dimensional (spatial) dependencies.

Additive regression spline models may be thought of as extensions of linear regression models that incorporate spatially varying dependences on the predictor variables. Additive regression splines may also be thought of as special cases of tensor-product smoothing splines. As such, they enable robust spatio-temporal analysis of climatic data that depend on many variables, in a spatially varying way. Additive regression spline models also bypass the usual technical difficulties associated with interpolation of higher dimensional data sets.

In this paper we discuss the application of additive regression spline models in the analysis of climatic data that depend on many variables in a spatially varying way. We illustrate their use in two applications. The first uses additional predictor variables related to topographic slope and aspect to analyse the topographic modulation of Swiss daily rainfall. Spatial patterns of the direction and extent of orographic modulation are presented along with an analysis of the short-range correlation structure within the data. The second application uses polynomial functions of time as additional predictors to analyse spatio-temporal trends in Australian pan evaporation data collected between 1970 and 2003. Unlike other methods employed in the literature to analyse temporal trends in climatic variables, the methods presented here allow use of data from all stations, not just the serially complete ones. Estimates of the spatially disaggregated linear trend in annual pan evaporation arising from the first-order and fourth-order temporal models are presented.

1. INTRODUCTION

Spatio-temporal analysis of climatic data using interpolation methods is important for a number of applications in environmental sciences. For example, such interpolatory methods play a pivotal role in the assessment of the impacts of climate on agriculture, ecology, hydrology and tourism. Furthermore, because physically based models present forecasts in the form of gridded surfaces at resolutions ranging from tens to hundreds of kilometres, methods for interpolating climatic data have an important part to play in the calibration and validation of such models. Spatio-temporal interpolation methods are also becoming increasingly important in climate change research. Such interpolation methods are required to establish spatial patterns in trends of climatic variables so that the impacts of potential climate changes can be assessed. In this paper we discuss such a method and illustrate its use with two case studies.

2. ADDITIVE REGRESSION SPLINES

Suppose that we have n dependent data values z_i (i=1,…,n) recorded at spatial locations with longitude x_i and latitude y_i, and that we have K additional predictor variables at the corresponding locations denoted by p_{ij}, j=1,…,K. The (bivariate) additive regression spline model is given by (Sharples and Hutchinson, 2004)

$$z_i = f_0(x_i, y_i) + \sum_{j=1}^{K} p_{ij} f_j(x_i, y_i) + \varepsilon_i, \qquad (1)$$

where i ranges from 1 to n. Heuristically, we may think of the additive regression spline model as a linear regression model in which the constant regression parameters have been replaced by the bivariate functions f_0,…, f_K. These functions will be referred to as the additive components of the model. Note that, although we restrict attention in this paper to bivariate additive regression spline models, additive regression spline models in which the additive components are univariate, trivariate or higher dimensional functions are also possible.

The random error term ε_i in (1) includes both errors in the measurement of the dependent variable z_i as well as errors due to the failings of the model. The errors are assumed to be realisations of a zero mean normal random variable with constant unknown variance σ^2 and positive spatial correlation specified by a single unknown parameter a. It is assumed that the variance matrix of the random error terms may be written

$$E(\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}) = \sigma^2 V_a,$$

where E denotes expectation, $\boldsymbol{\varepsilon}=(\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ and V_a is the $n \times n$ positive definite correlation matrix depending on the pair-wise horizontal separations of the n data points and the unknown scale parameter a.

The unknown functions f_0, f_1,…, f_K are then estimated by the suitably smooth bivariate functions g_0, g_1,…,g_K that minimise

$$(\mathbf{z}-\hat{\mathbf{z}})^T V_a^{-1} (\mathbf{z}-\hat{\mathbf{z}}) + \sum_{j=1}^{K} \rho_j J_2(g_j) \qquad (2)$$

where $\mathbf{z} = (z_1, \ldots, z_n)^T$ is the vector of dependent data values and $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_n)^T$ is the vector of fitted values with

$$\hat{z}_i = g_0(x_i, y_i) + \sum_{j=1}^{K} p_{ij} g_j(x_i, y_i)$$

$J_2(g)$ is a measure of roughness defined in terms of the second-order partial derivatives of the function g.

The non-negative smoothing parameters ρ_1,…, ρ_K determine a balance between how well the fitted additive regression spline model reproduces the dependent data values and the smoothness of the additive component functions. In practice the smoothing parameters are determined by appealing to standard methods such as minimising the generalised cross validation (GCV) or maximising the generalised maximum likelihood (GML). For more information on smoothing parameter selection see Wahba, 1985; 1990, Gu and Wahba, 1991.

As stated above, additive regression spline models can be viewed as special cases of tensor-product smoothing splines (Gu and Wahba, 1993a; 1993b). However, unlike tensor-product splines, additive regression splines can be implemented via a relatively simple extension of the methods used to derive standard thin-plate smoothing splines. This can be done without appeal to the underpinning reproducing kernel structure that is usually associated with tensor-product splines (Sharples and Hutchinson, 2004).

3. SPATIO-TEMPORAL ANALYSIS OF CLIMATIC DATA

In this section we provide examples of the application of additive regression splines to climatic data. In particular, we illustrate their use in the analysis of daily precipitation data collected over the

Swiss Alps and in the analysis of trends in Australian annual pan evaporation data.

3.1. Topographic enhancement of precipitation

The spatial distribution of rainfall over a region is often strongly related to the shape of the underlying topography (Spreen, 1947; Barry, 1992; Smith, 1979). Precipitation patterns can be affected by topography through a variety of processes. The most obvious examples are the wet regions found on the windward side and the corresponding dry, rain-shadow regions in the lee of mountain ranges that are subject to dominant prevailing winds. Other processes facilitating the topographic enhancement of precipitation include the formation of orographic wave clouds and induced atmospheric instability resulting from uplift and condensation (Smith, 1979).

The slope of the topography and its orientation with respect to the direction of dominant prevailing winds are important factors in determining where and by how much rainfall is enhanced. The effects of these features of the topography are also likely to be spatially variable. To properly account for features in the spatial distribution of rainfall such as rain-shadows, it therefore makes sense to include variables quantifying topographic slope and aspect in spatial interpolation methods, in addition to those quantifying spatial location. Moreover, it is desirable to allow for spatially varying dependencies on these variables over the region of interest.

To incorporate the effects of topographic slope and aspect into interpolation procedures we use the two horizontal components p and q, of the unit normal to an appropriately scaled digital elevation model (DEM). If the scaled DEM is characterised as the graph of the function $h(x,y)$ then the horizontal components of the unit normal are given by

$$p = \frac{-h_x(x,y)}{\sqrt{1+\left|\nabla h(x,y)\right|^2}}, \quad q = \frac{-h_y(x,y)}{\sqrt{1+\left|\nabla h(x,y)\right|^2}}$$

where h_x and h_y are the partial derivatives of the scaled DEM elevation $h(x,y)$ with respect to x and y respectively, and $\nabla h = (h_x, h_y)$ is the gradient vector field of the scaled DEM. Equivalently,

$$p = -\cos\alpha \sin\theta, \quad q = -\sin\alpha \sin\theta$$

where α is the topographic aspect angle and θ is the angle of steepest slope. The variables p and q, unlike topographic slope and aspect themselves, are continuous functions of horizontal location x, y. They are small for mild slopes, where topographic interaction with rain bearing atmospheric flows should be slight, and largest for steep regions where topography can have a significant influence on atmospheric flows. The use of both p and q permits the direction of interaction to be determined directly from the precipitation data without reference to the prevailing wind field (Hutchinson 1998b).

The bivariate additive regression spline model, incorporating the additional predictor variables p and q is given by

$$r_i^{\frac{1}{2}} = f_0(\mathbf{x}_i) + p_i f_p(\mathbf{x}_i) + q_i f_q(\mathbf{x}_i) + \varepsilon_i \qquad (3)$$

with i=1,…, n. Here r_i denotes the precipitation total recorded at the i-th data location $\mathbf{x}_i = (x_i, y_i)$. The square root transformation is applied to the precipitation totals to remove the natural skewness in rainfall data and to reduce interpolation error (Hutchinson, 1998a).

The method of maximising GML is used to determine the appropriate degree of smoothing of the functions f_0, f_p and f_q as well as the appropriate values of the two parameters σ^2 and a defining the random error (Wahba, 1985; Wang, 1998). We estimate the short-range correlation structure with three different candidate model: Exponential, Markov and Gaussian. Using $s_{ij} = a^{-1} \|\mathbf{x}_i - \mathbf{x}_j\|$ to denote the scaled separation of pairs of data points, the three models are given by

$$\begin{aligned} \text{Exponential:} \quad & [V_a]_{ij} = \exp(-s_{ij}) \\ \text{Markov:} \quad & [V_a]_{ij} = (1+s_{ij})\exp(-s_{ij}) \\ \text{Gaussian:} \quad & [V_a]_{ij} = \exp(-s_{ij}^2) \end{aligned}$$

The additive regression model (3) has been applied to daily precipitation totals collected over Switzerland on 8 May 1986. These data were used as the basis for the Spatial Interpolation Comparison 1997 (Dubois, 1998). The model (3) was fitted to a subset of 367 data points, whose locations are shown in figure 1 along with a 10 km resolution DEM of the region.

Of particular interest is the inferred effective wind field associated with the model (3). This two-dimensional vector field is defined as

$$\mathbf{w}(x,y) = \left(f_p(x,y), f_q(x,y)\right) \qquad (4)$$

and describes both the direction and magnitude of the orographic enhancement of the daily rainfall

totals. The fitted inferred effective wind field can be seen in figure 2.

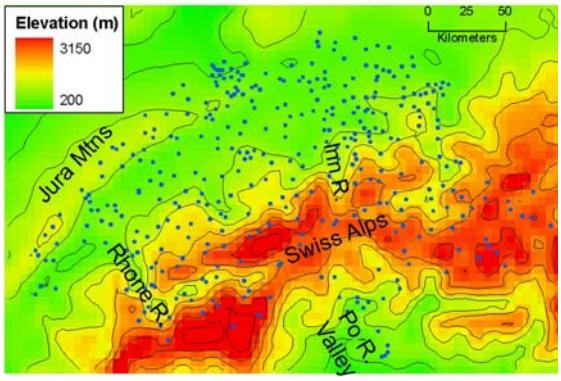

Figure 1. Locations of the 367 rain gauges overlaid on a 10km resolution DEM of the Swiss Alps

Figure 2 indicates clear topographic forcing in relation to the dominant synoptic conditions. The day's rainfall was associated with an extensive low-pressure system centred on Ireland. There is orographic enhancement on the southern face of the Alps, indicating channelling up the Po Valley and orographic enhancement on the north-western faces of both the Jura Mountains and the Alps, with channelling up and beyond the Rhone and Inn River Valleys (refer to figure 1 for the location of these subregions). This is in broad agreement with the precipitation climatology produced by Frei and Schär (1998).

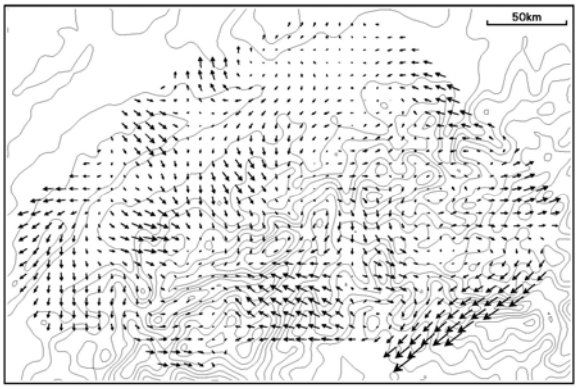

Figure 2. Inferred effective wind field **w**(*x*,*y*), overlaid on an 8km resolution DEM.

Figure 3 shows the fitted exponential, Markov and Gaussian models of the short-range correlation. All three models indicate substantial short-range correlation out to a separation of approximately 3-4 km.

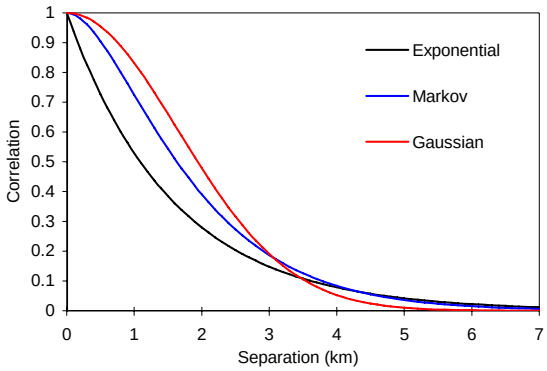

Figure 3. Fitted short-range error correlation models.

3.2. Spatio-temporal trends in Australian pan evaporation

Understanding the terrestrial water balance is important for a variety of applications including those encountered in agriculture, hydrology and ecology. Perceived changes in the terrestrial moisture balance, whether they are caused by anthropogenic influences or arise as a consequence of the natural variability of climatic systems, have the potential to affect the way we manage environmental systems, the way we formulate policy and the way we implement climate change mitigation and adaptation strategies.

An important component of the terrestrial moisture balance is the evaporative demand of the atmosphere. In practice the evaporative demand of the atmosphere, known as potential evaporation, is taken to be proportional to the amount of water that evaporates from a standardised class-A pan. Analyses of pan evaporation records across the globe suggest an overall decrease in pan evaporation over the last few decades. This apparent global decrease in pan evaporation does not seem to support the common belief that global warming will cause terrestrial evapotranspiration to increase. Reconciling this so-called 'pan evaporation paradox' has been the subject of several recent papers (Peterson *et al*., 1995; Chattopadhyay and Hulme, 1997; Brutsaert and Parlange, 1998; Thomas, 2000; Golubev *et al*., 2001; Roderick and Farquhar, 2002, 2004; Linacre, 2004; Liu *et al*., 2004; Hobbins *et al*., 2004).

To properly understand the driving factors behind changes in pan evaporation and the implications of these changes, robust analyses are required to effectively filter out the noise in the pan evaporation data and establish clear spatial patterns in pan evaporation trends. Moreover, since relatively few pan evaporation stations have complete records, the

analytical methods employed must also be able to deal with the fragmentary records of most pan evaporation stations. Additive regression spline models appear to be good candidates for meeting these requirements.

The pan evaporation data used in these analyses were obtained from the Australian Bureau of Meteorology (BoM). The BoM maintains a standardised class A pan network that provides reasonable coverage over most of the Australian continent. We used data from the years 1970 to 2003 since before 1970 the pan evaporation network was not considered reliable enough for analysis. A single datum consisted of the twelve monthly pan evaporation measurements taken at a particular station over a single year. For those years when a particular station didn't have any missing monthly pan evaporation measurements, annual pan evaporation totals were ascribed to the station by summing the twelve monthly data values.

In some instances pans were not fitted with bird screens to guard against animal predation until after 1970. The study conducted by van Dijk (1985) indicated that the presence of bird screens had the effect of reducing pan evaporation totals by 4-8% over the four different stations used in the study, with an average reduction of 7%. Hence rather than trying to account for homogeneity problems associated with installation of bird screens, pan evaporation totals recorded at stations not equipped with a bird screen were omitted from the analyses.

Over the period of 1970-2003, data from 450 pan evaporation stations were analysed. These analyses were based upon 6306 annual pan evaporation totals. The locations of the 450 stations can be seen in figure 4, while the number of active stations for each of the years 1970-2003 can be seen in figure 5.

To estimate spatial patterns in pan evaporation trends we use additive regression spline models incorporating temporal polynomials as additional predictors. In particular, we employ the Legendre polynomials, the first three of which are (Abramowitz and Stegun, 1972):

$$P_0(x)=1, \quad P_1(x)=x, \quad P_2(x)=\tfrac{1}{2}(3x^2-1)$$

with $-1 \le x \le 1$. Legendre polynomials are used instead of the standard polynomial basis functions because they are orthogonal with respect to the L^2-inner product (Kreyszig, 1978). This fact means that the temporal predictors defined by the Legendre polynomials are approximately independent. This ensures that models using these predictors are stable in the sense that dependencies on lower order temporal predictors are minimally affected by the inclusion of higher order temporal predictors in the model.

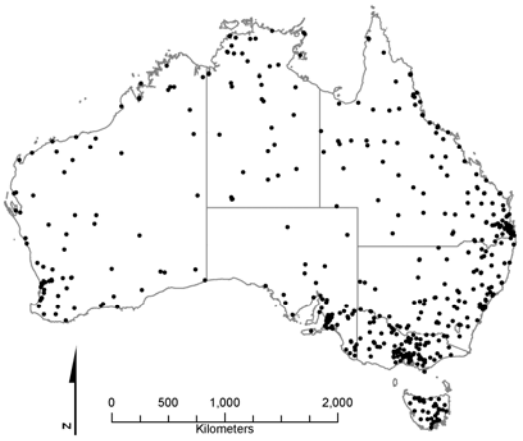

Figure 4. Locations of class A pan evaporation stations used in this study.

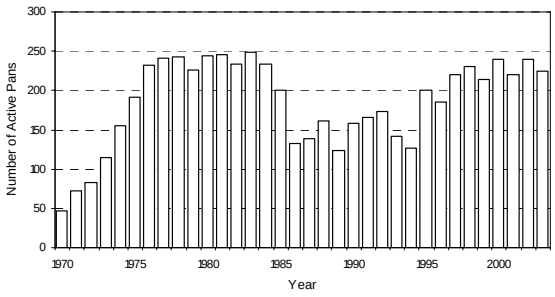

Figure 5. Number of active pan evaporation stations for each of the years 1970-2003.

To estimate the spatially disaggregated linear trend in pan evaporation, we consider the additive regression spline model

$$E_i^{\text{pan}} = f_0(x_i, y_i) + P_1(t_i) f_1(x_i, y_i) + \varepsilon_i \qquad (5)$$

with i=1,…, n. Here t_i is a transformed temporal coordinate. If we use T_i to denote the year (1970-2003), then the transformed temporal coordinate is given by

$$t_i = (2T_i - 3973)/33 \in [-1,1].$$

Considering the sparsity of the pan evaporation data network, we make the assumption of no spatial correlation in the errors, i.e. we take the covariance matrix $V_a = I$, the identity matrix. The appropriate degree of smoothing of the functions f_0 and f_1 and the appropriate value of the variance σ^2 are determined by minimising the GCV.
Fitting the additive regression spline model given by (5) to annual pan evaporation data and taking the

time derivative (with respect to the years) yields a spatially varying linear temporal trend estimate given by $2f_1(x,y)/33$ mm a^{-1}. The graph of this function can be seen in figure 6. The results confirm a decrease in annual pan evaporation for most parts of the continent but interestingly, suggest that the region encompassing most of Queensland, western New South Wales and northwest South Australia has experienced an increase in annual pan evaporation. Calculating the average of the trend surface in figure 6, we obtain an estimate for the continental trend in annual pan evaporation of –1.78 mm a^{-1}.

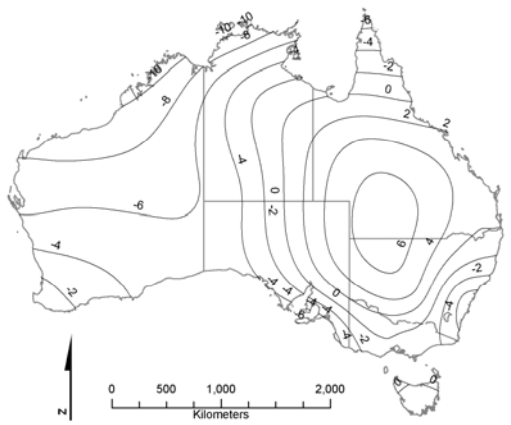

Figure 6. Estimated linear trend in annual pan evaporation. Contours show values in units of mm a^{-1}.

Estimation of higher order polynomial trends can be accomplished in a similar fashion after adding the appropriate higher order temporal Legendre polynomials to the set of additional predictors. The GCVs arising from the ensuing spline were seen to stabilise when the temporal variability was modelled using Legendre polynomials up to order four. This indicates that the most appropriate additive regression spline model incorporates temporal effects up to fourth-order:

$$E_i^{pan} = f_0(x_i, y_i) + \sum_{j=1}^{4} P_j(t_i)\, f_j(x_i, y_i) + \varepsilon_i, \quad (6)$$

with i=1,…, n. This assertion could also be confirmed by an analysis of deviance. The appropriate degree of smoothing of the functions f_0, f_1,…, f_4 and the appropriate value of the variance σ^2 are again determined by minimising the GCV.

The time derivative of the model (6) is now a trivariate function that varies with respect to space and time. The function f_1, present in the model (6), also provides an estimate of the spatially varying linear trend component of pan evaporation. This function, as estimated from the annual pan evaporation data, is displayed in figure 7.

The similarity in the spatial structure of the functions shown in figure 6 and figure 7 attests to the robustness of the methods employed. The magnitudes of the trends in figure 7, however, are smaller than those in figure 6 since the higher order components in the model (6) account for some of the temporal variability. Taking the average of the trend surface in figure 7, we obtain another estimate for the continental linear trend in pan evaporation of –1.81 mm a^{-1}, which is practically identical to the average obtained from the surface in figure 6.

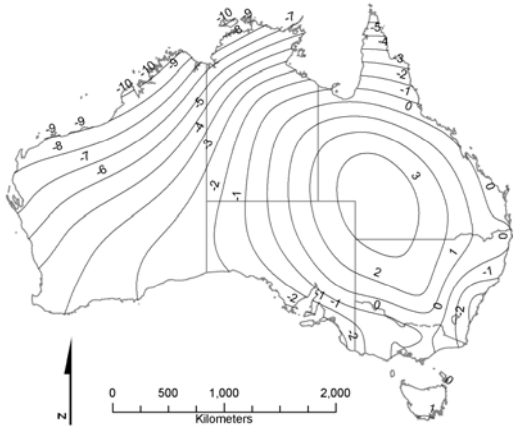

Figure 7. Linear component of the trend in annual pan evaporation as estimated by the fourth order temporal model. Contours show values in units of mm a^{-1}.

4. CONCLUSIONS

Additive regression spline models appear to be a practical option for analysing spatially varying effects of several predictors on observed climatic data. They are attractive from the point of view of overcoming curse of dimension problems associated with the analysis of noisy multivariate data. Moreover, their implementation involves a straightforward extension of existing standard thin-plate spline methodology.

The results presented in section 3.1 suggest that additive regression spline models, with short-range error correlation, are a promising option for elucidating multiple topographic dependencies of precipitation data. The models can be considered to separate physical processes, as embodied in the topographic variables p and q, from the spatial variation of the effects of these processes, as embodied in the functions f_p and f_q. The topographic variables can be replaced by known functions of these variables if these functions are known to be closer to the controlling process. Inclusion of short-range correlation in the models significantly reduced overall model complexity, enabling robust

calibration of the precipitation model from limited data

Additive regression spline models also facilitate robust analyses of the spatio-temporal variability of climatic data, as illustrated in section 3.2. They permit analysis of spatio-temporal trends based on data from all stations, no matter how short their records, and thus make use of many more stations than just the serially complete ones that are usually the focus of climatic trend analyses.

The additive regression spline models permit spatial disaggregation of the overall trends in climatic data. Obtaining a clear picture of the spatial patterns in climatic trends helps us understand the connections between various climatic factors and forcing agents. It can also provide insight into the effects of anthropogenic climate change.

5. REFERENCES

Brutsaert, W. and Parlange, M.B., 1998. Hydrological cycle explains the evaporation paradox. *Nature* 396, 30.

Chattopadhyay, N. and Hulme, M., 1997. Evaporation and potential evapotranspiration in India under conditions of recent and future climate change. *Agricultural and Forest Meteorology* 87, 55-73.

Dubois, G., (1998) Spatial Interpolation Comparison 97: Foreword and Introduction. *Journal of Geographical Information and Decision Analysis* 2(2): 1-10.

Frei, C. and Schär, C., (1998) A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology* 18: 873-900.

Gu, C. and Wahba, G., (1991), Minimising GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing* 12: 383-398.

Gu, C. and Wahba, G., (1993a), Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society Series B* 55: 1-23.

Gu, C. and Wahba, G., (1993b), Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *Journal of Computational and Graphical Statistics* 2: 353-368.

Hobbins, M.T., Ramirez, J.A. and Brown, T.C. 2004. Trends in pan evaporation and actual evapotranspiration across the conterminous U.S.: Paradoxical or complementary? *Geophysical Research Letters* 31, L13503, doi:10.1029/2004GL019846

Hutchinson, M.F., (1998a) Interpolation of rainfall data with thin plate smoothing splines I: two dimensional smoothing of data with short-range correlation. *Journal of Geographical Information and Decision Analysis* 2(2): 153-167.

Hutchinson, M.F., (1998b) Interpolation of rainfall data with thin plate smoothing splines II: analysis of topographic dependence. *Journal of Geographical Information and Decision Analysis* 2(2): 168-185.

Kreyszig, E., 1978. *Introductory Functional Analysis with Applications*. Wiley, New York.

Linacre, E.T., 2004. Evaporation trends. *Theoretical and Applied Climatology* 79, 11-21.

Liu, B., Xu, M., Henderson, M. and Gong, W., 2004. A spatial analysis of pan evaporation trends in China, 1955-2000. *Journal of Geophysical Research* 109, D15102, doi:10.1029/2004JD004511.

Peterson, T.C., Golubev, V.S. and Groisman, P.Y., 1995. Evaporation losing its strength. *Nature* 377, 687-688.

Roderick, M.L. and Farquhar, G.D., 2002. The cause of decreased pan evaporation over the past 50 years. *Science* 298, 1410-1411.

Roderick, M.L. and Farquhar, G.D., 2004. Changes in Australian pan evaporation from 1970 to 2002. *International Journal of Climatology* 24, 1077-1090.

Sharples, J.J. and Hutchinson, M.F., (2004), Multivariate spatial smoothing using additive regression splines. *ANZIAM Journal* 45 (E): C676-C692.

Thomas, A., 2000. Spatial and temporal characteristics of potential evapotranspiration trends over China. *International Journal of Climatology* 20, 381-396.

van Dijk, M.H., 1985. Reduction in evaporation due to the bird screen used in the Australian class A pan evaporation network. *Australian Meteorological Magazine* 33, 181-183.

Wahba, G., (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* 13: 1378-1402.

Wang, Y., (1998) Smoothing spline models with correlated errors. *Journal of the American Statistical Association* 93: 341-338.