ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 167-172

Statistical Downscaling with Artificial Neural Networks

Gavin C. Cawley*, Malcolm Haylock‡, Stephen R. Dorling†, Clare Goodess‡ and Philip D. Jones‡

*School of Information Systems, University of East Anglia, Norwich, U.K. NR4 7TJ.

‡Climatic Research Unit, University of East Anglia, Norwich, U.K. NR4 7TJ.

†School of Environmental Sciences, University of East Anglia, Norwich, U.K. NR4 7TJ.

Abstract.
Statistical downscaling methods seek to model the relationship between large scale atmospheric circulation, on say a European scale, and climatic variables, such as temperature and precipitation, on a regional or sub-regional scale. Downscaling is an important area of research as it bridges the gap between predictions of future circulation generated by General Circulation Models (GCMs) and the effects of climate change on smaller scales, which are often of greater interest to end-users. In this paper we describe a neural network based approach to statistical downscaling, with application to the analysis of events associated with extreme precipitation in the United Kingdom.

1 Introduction

General circulation models are considered to provide the best basis for estimating future climates that might result from anthropogenic modification of the atmospheric composition (i.e., the enhanced greenhouse effect). However, output from these models cannot be widely or directly applied in many impact studies because of their relatively coarse spatial resolution. The mismatch in scales between model resolution and the increasingly small scales required by impacts (e.g., agriculture and hydrology) analyses can be overcome by downscaling. Two major approaches to downscaling, statistical and dynamical (the latter using physically-based regional climate models), have been developed and tested in recent years, and shown to offer good potential for the construction of high-resolution scenarios of future climate change [1–4]. Statistical down-

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 167-172

scaling methods are based on the application of relationships identified in the real world, between the large-scale and smaller-scale climate, to climate model output and on two major assumptions: first, that variables representing large scale atmospheric processes (such as sea level pressure, geopotential height and relative humidity) are more reliably simulated by climate models than variables describing the smaller scale dynamics (such as rainfall); and, second, that the relationships between the large-scale and regional/local scale variables remain valid in a changed climate. In this paper we present the initial results of a study comparing error metrics for training neural network models for statistical downscaling of daily rainfall at stations covering the north-west of the United Kingdom, with application to modelling extreme events.

2 Method

For this study, we adopt the familiar Multi-Layer Perceptron network architecture (see e.g. Bishop [5]). The optimal model parameters, $\boldsymbol{w}$, are determined by gradient descent optimisation of an appropriate error function, $E_{\mathcal{D}}$, over a set of training examples, $\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{N}$, $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$, $t_i \in \mathbb{R}$, where $\boldsymbol{x}_i$ is the vector of explanatory variables and t_i is the desired output for the i^{th} training pattern. The error metric most commonly encountered in non-linear regression is the sum-of-squares error, given by

$$E_{\mathcal{D}} = \frac{1}{2}\sum_{i=1}^{N}(y_i - t_i)^2, \tag{1}$$

where y_i is the output of the network for the i^{th} training pattern. In order to avoid over-fitting to the training data, however, it is common to adopt a regularised [6] error function, adding a term $E_{\mathcal{W}}$ penalising overly-complex models, i.e.

$$M = \alpha E_{\mathcal{W}} + \beta E_{\mathcal{D}}, \tag{2}$$

where α and β are regularisation parameters controlling the bias-variance trade-off [7]. Minimising a regularised error function of this nature is equivalent to the Bayesian approach which seeks to maximise the posterior density of the weights (e.g. [8]), given by $P(\boldsymbol{w} \mid \mathcal{D}) \propto P(\mathcal{D} \mid \boldsymbol{w})P(\boldsymbol{w})$ where $P(\mathcal{D} \mid \boldsymbol{w})$ is the likelihood of the data and $P(\boldsymbol{w})$ is a prior distribution over $\boldsymbol{w}$. The form of the functions $E_{\mathcal{D}}$ and $E_{\mathcal{W}}$ correspond to distributional assumptions regarding the data likelihood and prior distribution over network parameters respectively. The usual sum-of-squares metric (1) corresponds to a Gaussian likelihood,

$$E_{\mathcal{D}} = \frac{1}{2}\sum_{i=1}^{N}(y_i - t_i)^2, \quad \Longleftrightarrow \quad P(\mathcal{D} \mid \boldsymbol{w}) = \frac{1}{\sqrt{2\pi\beta^{-1}}}\exp\left\{-\frac{[t_i - y(\boldsymbol{x}_i)]^2}{2\beta^{-1}}\right\}$$

with fixed variance $\sigma^2 = 1/\beta$. For this study, we adopt the Laplace prior propounded by Williams [9], which corresponds to a L_1 norm regularisation

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 167-172

term,

$$E_{\mathcal{W}} = \sum_{i=1}^{W} |w_i|. \qquad \Longleftrightarrow \qquad P(w) = \frac{1}{2\beta} \exp\left\{-\frac{|w|}{\beta}\right\}$$

where W is the number of model parameters. An interesting feature of the Laplace regulariser is that it leads to pruning of redundant model parameters. From 2, at a minimum of M we have

$$\left|\frac{\partial E_y}{\partial w_i}\right| = \frac{\alpha}{\beta} \qquad w_i > 0, \qquad \left|\frac{\partial E_y}{\partial w_i}\right| < \frac{\alpha}{\beta} \qquad w_i = 0.$$

As a result, any weight not obtaining the data misfit sensitivity of α/β is set exactly to zero and can be pruned from the network.

2.1 Eliminating Regularisation Parameters

To avoid the need for a lengthy search for the optimal regularisation parameters, α and β, they are integrated out analytically [9]. Adopting the (improper) uninformative Jeffreys prior, $p(\alpha) = 1/\alpha$ [10], applying the same treatment to the data misfit term (assuming a sum-of-squares error) and taking the negative logarithm (omitting irrelevant additive terms), we have

$$L = \frac{1}{2} N \log E_{\mathcal{D}} + W \log E_{\mathcal{W}}.$$

For a network with more than one output unit, it is sensible to assume that each output has a different noise process (and therefore a different optimal value for β). It is also sensible to assign hidden layer weights and weights associated with each output unit to different regularisation classes so they are regularised separately. This leads to the training criterion used in this study:

$$L = \frac{N}{2} \sum_{i=1}^{O} \log E_{\mathcal{D}}^{i} + \sum_{j=1}^{C} W_j \log E_{\mathcal{W}}^{j},$$

where O is the number of output units, C is the number of regularisation classes (groups of weights sharing the same regularisation parameter) and W_j is the number of non-zero weights in the j^{th} class. Note that bias parameters should not be regularised.

2.2 Choice of Data Misfit Term

In addition to the usual sum-of-squares error metric, which corresponds to the implicit assumption of a Gaussian noise process, we intend to investigate other data misfit terms corresponding to more realistic assumptions regarding the actual noise process, for example frontal precipitation is often modelled using a Gamma distribution [11] or a mixture of exponentials [12]. In this paper we begin by evaluating the hybrid Bernoulli/Gamma error metric proposed by

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 167-172

Williams [13]. The distribution of the amount of precipitation, X, is modelled by

$$P(X > x) = \begin{cases} 1 & \text{if } x < 0 \\ \alpha\Gamma\left(\nu, \frac{x}{\theta}\right) & \text{if } x \geq 0 \end{cases} \tag{3}$$

where $0 \leq \alpha < 1$, $\nu > 0$, $\theta > 0$ and $\Gamma(\nu, z)$ is the (upper) incomplete Gamma function, $\Gamma(\nu, z) = \Gamma(\nu)^{-1} \int_z^\infty y^{\nu-1} e^{-y} dy$. The model is then trained to approximate the conditional probability of rainfall $\alpha(\boldsymbol{x}_i)$ and the scale, $\theta(\boldsymbol{x}_i)$, and shape, $\nu(\boldsymbol{x}_i)$, parameters of a Gamma distribution modelling the predictive distribution of the amount of precipitation. Logistic and exponential activation functions are used in output layer neurons to ensure that the distributional parameters satisfy the constraints given previously.

3 Results

Artificial neural networks were then trained, using sum-of-squares and hybrid Bernoulli/Gamma data misfit terms, to model daily precipitation time series from 12 stations across the north-west of the United Kingdom, covering the period from Jan 1^{st}1960 to Dec 31^{st} 2000. The input to the model consisted of a set of 28 variables describing regional climatic conditions, for instance atmospheric pressure, temperature and humidity, extracted from the NCEP/NCAR reanalysis dataset [14]. An average over the predictions of twenty networks is taken in each experiment in order to provide some degree of robustness to the presence of local minima in the cost function. A simple two-fold cross-validation scheme was employed in assessing the performance of each cost function for each station, where the data was partitioned into contiguous sets describing the years $1960 - 1980$ and $1981 - 2000$. Tables 1 and 2 show the results obtained using sum-of-squares and hybrid Bernoulli/Gamma data misfit terms, for five test statistics for each of the twelve stations. The first statistic measures the root-mean-squared error (RMSE), giving a general indication of the ability of a model to reproduce the observed precipitation time series. The hybrid Bernoulli/Gamma model out-performs the sum-of-squares metric for every station (although the difference in performance is small). The remaining four statistics relate to the ability of the model to predict the occurrence of extreme precipitation. We define an extreme event as the occurrence of rainfall at levels above the 90^{th} or 95^{th} percentile of entire time-series for a given station. The probability of an excedance can be calculated by simply integrating the upper tail of the predictive distribution above the appropriate threshold level. It is then appropriate to measure the ability of the model to identify extreme events using the cross-entropy and area under the ROC curve statistics (since the misclassification costs are unknown). Again the hybrid Bernoulli/Gamma models out-perform the more conventional sum-of-squares metric for all statistics, for all stations.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 167-172

Table 1: Results for sum-of-squares data misfit term.

Station	RMSE	AUROC$_{90}$	XENT$_{90}$	AUROC$_{95}$	XENT$_{95}$
Appleby Castle	3.7444	0.8667	3320.73	0.8911	1991.08
Carlisle	3.6433	0.8436	3737.49	0.8604	2318.02
Douglas	4.9440	0.8494	3447.95	0.8559	2274.04
Haydon Bridge	3.5732	0.8235	4123.04	0.8497	2320.32
Keele	3.6136	0.8354	3864.68	0.8485	2428.67
Loggerheads	4.2811	0.8360	4035.75	0.8511	2411.00
Lyme Park	3.9568	0.8552	3365.60	0.8724	2144.30
Morecambe	4.2112	0.8674	3320.85	0.8809	2148.81
Newton Rigg	3.8326	0.8745	3440.64	0.8860	2224.12
Pen Y Ffridd	4.6962	0.8484	3490.81	0.8814	2114.63
Ringway	3.7558	0.8352	3432.96	0.8531	2106.55
Slaidburn	5.5720	0.8933	3045.82	0.9097	1953.09

Table 2: Results for hybrid Bernoulli-Gamma misfit term.

Station	RMSE	AUROC$_{90}$	XENT$_{90}$	AUROC$_{95}$	XENT$_{95}$
Appleby Castle	3.6934	0.8732	3085.66	0.8943	1881.61
Carlisle	3.6106	0.8504	3542.59	0.8661	2191.44
Douglas	4.9033	0.8548	3318.07	0.8611	2118.66
Haydon Bridge	3.5417	0.8324	3752.74	0.8546	2300.11
Keele	3.5929	0.8391	3695.27	0.8500	2329.65
Loggerheads	4.2519	0.8415	3704.52	0.8535	2336.00
Lyme Park	3.9354	0.8572	3285.96	0.8748	2026.60
Morecambe	4.1741	0.8708	3216.94	0.8818	2018.83
Newton Rigg	3.7755	0.8813	3226.18	0.8891	2063.19
Pen Y Ffridd	4.6787	0.8489	3370.69	0.8822	1988.07
Ringway	3.7066	0.8411	3242.77	0.8580	2016.91
Slaidburn	5.5389	0.8985	2895.57	0.9116	1795.75

4 Summary

In this paper we have described the use of multi-layer perceptron networks in statistical downscaling of daily precipitation at a network of stations covering the north-west of the United Kingdom. Furthermore, we have demonstrated that the use of an error metric incorporating realistic distributional assumptions results in consistently higher performance than is obtained using the more conventional sum-of-squares error metric. Further work is in progress to intercompare the performance of a wider range of realistic error metrics, such as a mixture of exponential or Gamma distributions.

ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 167-172

5 Acknowledgements

This work was supported by the European Commission under the Fifth Framework Thematic Programme Energy, Environment and Sustainable Development as part of the STARDEX project (STAtistical and Regional dynamical Downscaling of EXtremes for European regions, Contract no: EVK2-CT-2001-00115, `http//www.cru.uea.ac.uk/projects/stardex`) and by the Royal Society (research grant RSRG-22270). The UK precipitation data were provided by the British Atmospheric Data Centre. The 28 predictor variables were provided by Rob Wilby, Kings College London.

References

[1] B. C. Hewitson and R. G. Crane. Climate downscaling: Techniques and application. *Climate Research*, 7:85–95, 1996.

[2] R. L. Wilby, T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, 34:2995–3008, 1998.

[3] F. Giorgi and L. O. Mearns. Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research*, 104:6335–6352, 1999.

[4] E. Zorita and H. von Storch. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, 12:2474–2489, 1999.

[5] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[6] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems.* John Wiley, New York, 1977.

[7] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[8] D. J. C. Mackay. A practical Bayesian framework for backprop networks. *Neural Computation*, 4:448–472, 1992.

[9] P. M. Williams. Bayesian regularisation and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

[10] H. S. Jeffreys. *Theory of Probability.* Oxford University Press, 1939.

[11] R. D. Stern and R. Coe. A model fitting analysis of daily rainfall data (with discussion). *Journal of the Royal Statistical Society A*, 147:1–34, 1984.

[12] D. A. Woolhiser and G. G. S. Pegram. Maximum likelihood estimation of Fourier coefficients to describe seasonal variation of parameters in stochastic daily precipitation models. *Journal of Applied Meteorology*, 18:34–42, 1979.

[13] P. M. Williams. Modelling seasonality and trends in daily rainfall data. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems - Proceedings of the 1997 Conference*, volume 10, pages 985–991. MIT Press, 1998.

[14] E. Kalnay *et al.* The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, March 1996.