Predictive Uncertainty in Environmental Modelling

Gavin C. Cawley and Gareth J. Janacek

School of Computing Sciences

University of East Anglia

Norwich NR4 7TJ

United Kingdom

E-mail: gcc@cmp.uea.ac.uk

Malcolm R. Haylock

Climatic Research Unit

University of East Anglia

Norwich NR4 7TJ

United Kingdom

E-mail: M.Haylock@uea.ac.uk

Stephen R. Dorling

School of Environmental Sciences

University of East Anglia

Norwich NR4 7TJ

United Kingdom

E-mail: S.Dorling@uea.ac.uk

Abstract

Artificial neural networks have proved an attractive approach to non-linear regression problems arising in environmental modelling, such as statistical downscaling, short-term forecasting of atmospheric pollutant concentrations and rainfall run-off modelling. However, environmental datasets are frequently very noisy and characterised by a noise process that may be heteroscedastic (having input dependent variance) and/or non-Gaussian. The aim of this paper is to review existing methodologies for estimating predictive uncertainty in such situations, and more importantly to illustrate how a model of the predictive distribution may be exploited in assessing the possible impacts of climate change and to improve current decision making processes. The results of the WCCI-2006 predictive uncertainty in environmental modelling challenge are also reviewed, suggesting a number of areas where further research may provide significant benefits.

Keywords: Predictive uncertainty, environmental modelling, multi-layer perceptron, statistics

1 Introduction

Neural networks have been shown to provide a simple and flexible approach to a wide variety of non-linear regression problems arising in the environmental sciences. Some recent applications include statistical downscaling (Harpham & Wilby, 2005), water level-discharge modelling (Bhattacharya & Solomatine, 2005), river stage forecasting (Dawson et al., 2005) and air quality forecasting (Schlink et al., 2003). The presence of special sessions devoted to environmental sciences and climate modelling at IJCNN-2005 and IJCNN-2006 provides further evidence of the importance of this field of research. Environmental modelling problems are typically very noisy and often characterised by a noise process that is heteroscedastic (i.e. the variance of the noise process is input-dependent) and may also be non-Gaussian, for example the target data may be strictly non-negative or highly skewed. Conventional neural network regression techniques aim to estimate the conditional mean of the target data, via minimisation of a sum-of-squares error function. The aim of this paper is to demonstrate that practical benefits can be accrued by attempting to model the entire conditional distribution of the noise contaminating the data in addition to the conditional mean. For example, we may estimate the conditional variance of a Gaussian noise process, which may be achieved by training a second regression network to predict the squared residuals of the first (e.g. Nix & Weigend, 1994). The combined model provides a Gaussian *predictive distribution* indicating the relative plausibility of different values for the target function. The provision of a predictive distribution, instead of only the conditional mean, can be exploited in a number of ways:

- The predictive distribution implies a plausible interval (a.k.a. “error bars”) on all predictions, which in turn provide a valuable indicator of the reliability of the model.
- An estimate of the predictive distribution allows the estimation of the true *risk*, i.e. we may integrate the loss associated with all plausible outcomes, weighted by the probability of their occurrence.
- Where a neural network is used as one component within a much larger model, the un-

certainties associated with the inputs and outputs of each component, may be propagated through the model (e.g. via a Monte-Carlo simulation) so that all sources of uncertainty can be integrated over to obtain a moderated prediction.

- Often we are interested in predicting extreme events, especially the exceedance of some arbitrary threshold, for instance predicting episodes of poor air quality. By their very nature, extreme events are not modelled well by an estimate of conditional mean of the data, and so a conventional sum-of-squares model will consistently under-predict extreme events. However, given a full predictive distribution, we may at least estimate the *probability* of an extreme event by integrating the upper tail of the predictive distribution, even if the estimate of the conditional mean never exceeds the threshold.

Modelling predictive uncertainty in environmental data is also interesting from a machine learning perspective as the noise processes involved are often non-Gaussian and/or heteroscedastic, and so “off-the-shelf” solutions may not be entirely satisfactory, and thus there is significant scope for further research.

The remainder of this paper is structured as follows: Section 2 describes the four benchmark datasets used in the WCCI-2006 predictive uncertainty in environmental modelling challenge. Section 3 presents a variety of conventional statistical approaches, and discusses the deviations from the usual modelling assumptions often encountered in environmental modelling. Section 4 describes a simple methodology for estimating the predictive distribution based on methods developed by Peter Williams (Williams, 1991, 1995, 1996, 1998). Section 5 demonstrates that an estimate of the predictive distribution can be exploited to provide practical benefits for the end-user, via an illustrative (if a little contrived) example based on the estimation of insurance losses associated with flood hazards. The results of the WCCI-2006 Predictive Uncertainty in Environmental Modelling Competition, which aimed to stimulate research in this area, are presented in Section 6. Section 7 discusses some areas where further research may provide significant benefits. Finally, the work is summarised and conclusions draw in Section 8.

2 Datasets

In this section, we describe the four benchmark datasets that were used in the WCCI-2006 predictive uncertainty in environmental modelling challenge. These benchmarks are also used in this paper to demonstrate the importance of estimating predictive uncertainty, especially for datasets characterised by a non-Gaussian or heteroscedastic variance structure. These datasets are freely available from the challenge website (`http://theoval.cmp.uea.ac.uk/∼gcc/competition/`).

2.1 The `SYNTHETIC` Benchmark

A synthetic heteroscedastic regression problem, taken from Williams, 1996, provides a relatively small dataset that can be easily visualised for the purposes of model development and for illustrating the importance of predictive uncertainty. As the true conditional mean and variance functions are known, it is straight-forward to assess the quality of the model without direct access to the test data. The univariate input patterns, x, are drawn from a uniform distribution on the interval $(0,\ \pi)$, the corresponding targets, y, are drawn from a univariate Normal distribution with mean and variance that vary smoothly with x:

$$
\begin{aligned}
x_i &\sim \mathscr{U}(0, \pi), \\
y_i &\sim \mathscr{N}\left(\sin\left[\frac{5x}{2}\right] \sin\left[\frac{3x}{2}\right], \frac{1}{100} + \frac{1}{4}\left[1 - \sin\left[\frac{5x}{2}\right]\right]^2 \right).
\end{aligned}
$$

Figure 1 shows a plot of the synthetic benchmark dataset, along with indications of the true conditional mean and standard deviation. The heteroscedastic (input-dependent variance) nature of the data is clearly evident.

[Figure 1 about here.]

2.2 The `SO2` Benchmark

The `SO2` benchmark represents an atmospheric pollution forecasting problem, where the aim is to predict 24 hours in advance the SO_2 concentration in urban Belfast, based on meteorological conditions and current SO_2 levels (see Nunnari et al., 2004 for further details). The meteorological conditions are important in this case as the air pollution problem in urban Belfast is largely due to domestic (commonly coal-fired) heating, and so is at its worst during periods of cold weather. Also high atmospheric pressure and temperature inversions tend to cause stagnant conditions and consequently poor dispersion of atmospheric pollutants.

2.3 The `PRECIP` Benchmark

The `PRECIP` benchmark models a realistic statistical downscaling exercise, the aim of which is to predict the (scaled) precipitation for Newton Rigg, a relatively wet station in the North-West of the United Kingdom, using inputs representing large scale circulation features (see Cawley, Dorling, Jones, & Goodess, 2003; Haylock, Cawley, Harpham, Wilby, & Goodess, 2006 for further details). Figure 2 shows a histogram of the target data for the training set of the `PRECIP` benchmark, highlighting a number of unusual features of this dataset. Firstly, the data is non-negative (it would make little sense to talk of negative rainfall). Secondly, there is a large probability mass centred on zero, representing the proportion of days where no rainfall occurs. Rainfall presents an example of a *mixed* distribution, and is often modelled as separate occurrence and amount processes, where the probability of rainfall is given by, e.g. a logistic regression model, and the amount of rainfall given by a regression model fitted to the training data representing days where rainfall was actually observed.

[Figure 2 about here.]

2.4 The TEMP Benchmark

The TEMP benchmark problem is perhaps the most easily modelled of the real-world benchmark problems, and again represents a downscaling problem, where the aim in this case is to model the daily maximum temperature at the Writtle station in the South-East of the United Kingdom base on similar large scale circulation features as those used for the PRECIP benchmark. In this case, the data are reasonably well modelled by a conventional sum-of-squares regression model.

3 Conventional Statistical Approaches

The data encountered in environmental applications are typically characterised by deviations from the basic assumptions underpinning least-squares regression methods commonly employed. These may include a non-Gaussian noise process, heteroscedasticity (i.e. a noise process with input dependent variance), the data may not represent a truly independent and identically distributed sample (e.g. there may be temporal or spatial correlations). In this section, we review some methods from classical statistics used to counter some of these problems, illustrating each technique using examples based on the four benchmark datasets.

3.1 Background

Suppose we have a dataset $\mathscr{D} = \{(y_i,\, x_i)\}_{i=1}^{\ell}$, comprised of ℓ observations on a set of random variables $Y,\ X = (X_1,\, X_2, \ldots, X_d)$. Our interest lies in predicting the Y values, the *responses*, given the explanatory variables X, and so we seek to fit a predictive model to the available data. The best point predictor of Y is given by $\psi(X)$, the mean of the *conditional* distribution of Y given the X's. That is

$$\psi(X) = E\left[Y \mid X\right]$$

This prediction is optimal in the sense that it minimises the expected squared error, such that

$$E\left[(Y-\psi(X))^2\right] \leq E\left[(Y-\theta(X))^2\right]$$

for any function $\theta(\cdot)$. The conditional mean function $\psi(X)$ is called the *regression* function and we can show that in the case where our variables Y and X have a joint multivariate Normal distribution ψ is a linear function of the inputs (c.f. Rasmussen & Williams, 2006).

A conventional statistical approach normally begins by assuming that the responses are realisations of a deterministic process, which can be represented by a linear function of the explanatory variables, corrupted by zero mean Gaussian noise, with a fixed variance, i.e.,

$$y_i = x_i^T \beta + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_d)$ is a vector containing the coefficients of the linear regression model (we assume that if an offset or bias term is required, it is implemented by adding a dummy explanatory variable with a fixed non-zero value). Very often it is more convenient to think of a vector of responses $y = (y_1,\ y_2,\ \ldots,\ y_n)^T$ in which case we have

$$y = X\beta + \varepsilon \tag{2}$$

where $X = \left\{x_{ij}\right\}_{i,j=1}^{i=\ell,j=d}$ is the $\ell \times d$ *design* matrix and $\varepsilon = (\varepsilon_1,\ \varepsilon_2,\ \ldots, \varepsilon_\ell)$. The regression coefficients are determined by minimising the sum-of-squares function

$$Q = (y - X\beta)^T (y - X\beta).$$

Using standard results we find the condition for a minimum is

$$X^T X\hat{\beta} = X^T y \quad \Longrightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y. \tag{3}$$

Assuming the errors are zero-mean ($E[\varepsilon] = 0$), and are uncorrelated with common variance σ^2 ($\mathrm{var}(\varepsilon) = \Sigma = \sigma^2 I$) and that the elements of the design matrix X are not stochastic, i.e. there is no uncertainty in the measurement of the explanatory variables, then $\hat{\beta}$ are also maximum likelihood estimates. While the model provides a good point prediction for the target variable, we often require an indication of the uncertainty of the predictions, in the form of a predictive distribution, i.e. the distribution of plausible values for the response variable given the vector of observed explanatory variables. In this case, a sensible choice would be

$$\hat{y}_i \sim \mathscr{N}(x^T \hat{\beta} | \hat{\sigma}^2),$$

where

$$\hat{\sigma}^2 = \frac{1}{N-d}(y - X\hat{\beta})^T (y - X\hat{\beta})$$

is the optimal unbiased estimate of the variance of the noise process. Of course in reality many of the problems we are interested in are concerned with non-normal data or violate some other regression assumption. In the remainder of this section, we will describe some measures that can be taken to deal with these deviations from the standard assumptions, with illustrations based on the benchmark datasets.

3.2 Transformations of the Response Variable

A non-Gaussian noise process is a fairly common characteristic of environmental datasets, where the response variable may be strictly non-negative (e.g. concentrations of atmospheric pollutants) or highly skewed, or both. One approach to this problem is to transform the response data, for example one might try a square root or logarithmic transformation. In many applications, the responses exhibit a conditional variance that is dependent on the conditional mean, for instance count data. In this case, the aim of a transformation is usually to *stabilize* the variance, i.e. make it constant. For example the arcsin transform $\arcsin\sqrt{x}$ can be shown to make the variance of Binomial variables constant. There are however rather more systematic approaches such as the

Box-Cox transform (Box & Cox, 1964)

$$y^{(\lambda)} = \begin{cases} (y_i^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}$$

In the regression case Box and Cox suggest using the value of λ which gives the maximum of the profile likelihood. Suppose y^{λ} is the transformed value of the response and assume that y^{λ} follows a Normal linear model with parameters β and σ^2 for some value of λ. Given a value of λ, we can estimate the linear model parameters β and σ^2 for transformed response y^{λ} not y. For the transformed response y^{λ}, the log-likelihood is

$$\ell(\beta, \sigma, \lambda) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}2\sum_{i=1}^{n}(y_i^{\lambda} - \mu_i)^2$$

and when we change variables to y the resulting log-likelihood is

$$\ell(\beta, \sigma, \lambda) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}2\sum_{i=1}^{n}(y_i^{\lambda} - \mu_i) + \sum_{i=1}^{n}(\lambda - 1)\log(y_i)$$

where the last term is Jacobian of the transformation. We can estimate β and σ^2 for any fixed value of λ by regressing the transformed response y^{λ} on the inputs. Substituting the resulting maximum likelihood estimates of β and σ^2 we obtain the profile log-likelihood

$$\ell_p(\lambda) = C - \frac{n}{2}\log\left(\hat{\sigma}_{\lambda}^2\right) + \sum_{i=1}^{n}(\lambda - 1)\log(y_i)$$

where C is a constant, not involving λ, It is common to work with

$$y_g^{\lambda} = y^{\lambda}\left(\prod_{i=1}^{n} y_i\right)^{-1/n}$$

In this case it is easy to see that

$$\ell_p(\lambda) = C - \frac{n}{2}\log\left(\hat{\sigma}_{\lambda}^2\right)$$

For a transformation we choose the λ that maximizes $\ell_p(\lambda)$ and would normally plot the profile likelihood to look at the region around the maximum.

An alternative to the Box-Cox form is the Yeo-Johnson power transformation (Yeo & Johnson, 2000). These transformations are defined as

$$\psi(\lambda, y) = \begin{cases} \left((y+1)^{\lambda} - 1\right)/\lambda & \lambda \neq 0,\ y \geq 0 \\ \left(1 - (1-y)^{2-\lambda}\right)/(2-\lambda) & \lambda \neq 2,\ y < 0 \\ -\log(-y+1) & \lambda = 2,\ y < 0 \end{cases}$$

If y is strictly positive, then the Yeo-Johnson transformation is identical to the Box-Cox transformation. When Y is strictly negative it is identical to the Box-Cox transformation of $(-y+1)$, with power $2-\lambda$. When y takes values of both sign then we have some problems in that we have differing transformations for different observations. As far as we are aware the difficulties encountered in this case are not resolved and in consequence we prefr Box-Cox

3.2.1 Example - the `SO2` Dataset

Consider the `SO2` data, the raw histogram of the responses looks very skewed, as shown in Figure 3, and we suspect non-normality of the noise process. Note that in this case the response variable represents the concentration of an atmospheric pollutant, and is strictly non-negative and thus a Gaussian noise process is clearly unreasonable unless the variance of the noise is negligible in comparison to the magnitude of the conditional mean. Note that we seek a transformation that makes the residuals of the model approximately normal, rather than merely the unconditional distribution of the responses, shown in Figure 3. Some of the responses are zero, presumably as the pollutant concentration was below the level that can be detected. If we replace these zeros with the value 1.5, which is half the smallest non zero reading, we can try the Box-Cox transformation. The profile likelihood is shown in Figure 4. The plot is peaked at a value very close to zero, giving a strong indication that we should use a logarithmic transform. The histogram of the transformed response variable is also shown in Figure 3, clearly a normal noise process is a more reasonable proposition for the transformed data. The predictive distribution must also be transformed, to account for the transformation of the response, in this case giving a log-normal predictive distribution.

[Figure 3 about here.]

[Figure 4 about here.]

3.3 The generalized linear model

An extension to the linear model, first proposed by Nelder in the mid 70's is known as generalized linear modelling (McCullagh & Nelder, 1989). The aim is to have a richer class of error distributions and to try and ensure that a linear combination of the predictors gives a reasonable model. We cannot deal with all possible distributions for the noise process (and therefore the predictive distribution) so we restrict ourselves to the exponential family. This is the family of distributions whose density or probability functions are of the form

$$f(y:\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\} \tag{4}$$

here θ and ϕ are parameters of the distribution, while $a(), b()$ and $c()$ are known functions. Many of the distributions encountered in environmental modelling belong to this family. Our interest is the value of the canonical parameter θ (commonly the mean μ), and we regard ϕ as a nuisance parameter. If we use standard distributional results it is easy to show that

$$E[y] = \mu = b^{'}(\theta) \tag{5}$$

and

$$\mathrm{var}(y) = b^{''}(\theta)a(\phi) \tag{6}$$

So we see that the variance is the product of two functions $b^{''}(\theta)$ and $a(\phi)$ which depends only on ϕ. We can write the variance function as a function of the mean μ, say $V(\mu)$. It is common for the function $a(\phi)$ to take the form $a(\phi) = \frac{\phi}{w}$, where ϕ, called the *dispersion parameter*, is constant over the data set and w is a known prior weight. The characteristics of some of the more important members of the exponential family are as follows: For the normal distribution,

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right],$$

so in this parametrization $a(\sigma) = \sigma^2$ and the nuisance parameter is σ. For the Gamma distribution, we have

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu y}{\mu}\right)$$

This parametrization gives $E[Y] = \mu$ and $\text{var}[Y] = \mu^2/\nu$, and it seems clear that $a(\nu) = \frac{1}{\nu}$. Lastly, for the binomial distribution,

$$f(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \quad y = 0, 1, \cdots, n$$

Here we have only one parameter π and the dispersion is one, that is $a(\phi) = 1$.

For the linear model we just equate the value of the predictor function, η, to the mean, μ,

$$\mu_j = \eta_j = \mathbf{x_j^T}\beta$$

For a generalized linear model we connect the mean and the predictor by a monotone *link* function $g(\cdot)$,

$$g(\mu_j) = \eta_j = \mathbf{x_j^T}\beta$$

While from the technical statistical view there is much to be said for the canonical link[1], there is no *a-priori* reason why this link is appropriate for a particular data set. The link function is chosen to ensure additivity and linearity of the explanatory variables. This choice is part of the modelling process. All the distributions that we commonly use have special canonical link functions, those

[1] Using the canonical link, the partial derivative of the negative log-likelihood with respect to η_i is given by $y_i - g(\eta_i)$, simplifying the optimisation procedure used to fit the model.

of interest to this study are set out in Table 1.

[Table 1 about here.]

3.3.1 Example - the Precipitation Dataset

The precipitation data is interesting in that the responses are the product of an occurrence process, which decides whether or not there is any rainfall on a particular day, and an amount process, which governs the amount of rainfall, given that some rainfall is observed. A reasonable approach to this model is to assign a probability, π, to the event that it rains and then consider the amount of rain, X, *given* it has rained, using separate models. We have now estimated our distribution with a probability $1-\pi$ at zero and a distribution $\pi f(x \mid \text{ rain })$ elsewhere. Our approach is therefore

1. To use a logistic model to estimate the probability of rain π. That is

   $$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta.$$

   Here the x_i^T are the inputs. This is a standard problem as the rain, no rain dichotomy gives us a series of Bernoulli trials.

2. We now look at days with rain and model the amounts using a Gamma generalized linear model. We choose the Gamma as it appears to provide a reasonable fit to the data and we are advised by those who study precipitation that a Gamma distribution is appropriate. Here we have used a log link as it gives a better model, so

   $$\eta_i = \log(\mu_i) = x_i^T\beta.$$

   The error distribution being Gamma, that is

   $$f(y,\mu,\nu) = \frac{1}{\Gamma(\nu)y}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left(-\frac{\nu y}{\mu}\right).$$

3.4 Dispersion Models

It is possible to modify our approaches to both the normal regression and the generalized model case to allow for non-constant (*heteroscedastic*) variances. The normal case uses an algorithm suggested by Aitken, 1987. If we have d predictors x and a response y we can, as a model assume

$$E[y|x] = \beta^T x \quad \text{and} \quad \mathrm{var}(y|x) = \exp(\gamma^T x).$$

where the coefficients of the models of the conditional mean and conditional variance, β^T and γ^T respectively, are estimated separately. The fitting procedure alternates between updates of the model of the conditional mean, by fitting β via weighted least-squares with weights $\exp(-\gamma^T x)$, and updates of the model of the conditional variance, fitting a Gamma model to the square residuals r with scale factor 2. The algorithm begins by assuming constant variance, i.e. $\gamma = 0$, and is repeated until convergence is obtained. In the generalised linear model case we can follow a similar procedure. Suppose we write $\mu_i = E[y_i]$ for the expectation of the ith response. Then we know from the structure of generalized linear models that $\mathrm{Var}(y_i) = \phi_i V(\mu_i)$ where V is the variance function and ϕ_i is the dispersion of the i^{th} response. We now assume the linked linear models

$$g(\mu_i) = x_i^T \beta \qquad \text{and} \qquad h(\phi_i) = z_i^T \gamma$$

with linear and exponential link functions, $g(z) = z$ and $h(z) = \exp(z)$ respectively. The parameters β are estimated as for a standard generalized linear model. The parameters γ are estimated by way of a dual generalized linear model, in which the deviance components of the ordinary generalized linear model appear as responses. The estimation procedure alternates between one iteration for the mean submodel and one iteration for the dispersion submodel until we have convergence.

3.4.1 Example: The Synthetic Dataset

Looking at the synthetic data, as shown in Figure 1, it is immediately apparent that we have non-constant variance and that the conditional mean is a non-linear function of the explanatory variable. In this case, the variance of the target distribution is not a simple function of the mean and so a Box-Cox transformation is not helpful here. However, the noise process is Gaussian, albeit heteroscedastic, and so Aitken's dispersion model is highly appropriate. As the conditional mean and variance of the response variable are non-linear functions of the single explanatory variable, we fit a polynomial model by augmenting the input vector such that $x = (x,\ x^1,\ x^2,\ \ldots,\ x^7,\ 1)$, including a bias term. The fitted conditional means and variances are shown in Figure 5.

[Figure 5 about here.]

3.5 Quantile Regression

An alternative approach to modelling predictive uncertainty seeks to model the predictive distribution directly rather than assume a particular parametric form. This can be achieved by forming a predictive model estimating the quantiles of the target distribution. Just as minimising the sum-of-squares error leads to estimation of the conditional mean of the target distribution, a linear model fitted via minimisation of the sum-of-absolute errors,

$$\frac{1}{\ell}\sum_{i=1}^{\ell}|y_i - \hat{y}_i|$$

can be shown to estimate the conditional median of the responses. This approach can be generalised to provide the q^{th} quantile, by differentially weighting positive and negative errors, according to

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\psi(\hat{y}_i, y_i)$$

where

$$\psi(\hat{y}, y) = \begin{cases} (1-q)(\hat{y}-y) & \text{if } \hat{y} > y \\ q(y_i - \hat{y}) & \text{otherwise} \end{cases}$$

This approach is known as quantile regression (Koenker, 2005).

3.5.1 Example : The `SYNTH` Benchmark

Figure 6 shows (a) the conditional median and (b) conditional deciles ($q = 0.1, 0.2, \ldots, 0.9$) of the synthetic benchmark dataset. While the conditional median is useful in that it provides a more robust estimate of central tendency (in the sense that the mean absolute error is less sensitive to outliers than the mean squared error), the conditional deciles provide a useful (if somewhat coarse) indication of the shape of the target distribution.

[Figure 6 about here.]

4 Modelling Predictive Uncertainty with Neural Networks

In this section, we outline a neural network approach to modelling predictive uncertainty in environmental applications, based on a sound Bayesian methodology developed by Williams (Williams, 1991, 1995, 1996, 1998). For this study, we adopt the familiar Multi-Layer Perceptron network architecture (see e.g. Bishop, 1995). The optimal model parameters, w, are determined by gradient descent optimisation of an appropriate error function, $E_{\mathscr{D}}$, over a set of training examples, $\mathscr{D} = \{(x_i, t_i)\}_{i=1}^{N}$, $x_i \in \mathscr{X} \subset \mathbb{R}^d$, $t_i \in \mathbb{R}$, where 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

$$E_{\mathscr{D}} = \frac{1}{2} \sum_{i=1}^{N} (y_i - t_i)^2, \tag{7}$$

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 (Tikhonov & Arsenin, 1977) error function, adding a term $E_{\mathscr{W}}$ penalising overly-complex models, i.e.

$$M = \alpha E_{\mathscr{W}} + \beta E_{\mathscr{D}}, \tag{8}$$

where α and β are regularisation parameters controlling the bias-variance trade-off (Geman, Bienenstock, & Doursat, 1992). 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. MacKay, 1992b; Neal, 1996), given by

$$P(w \mid \mathscr{D}) \propto P(\mathscr{D} \mid w)P(w),$$

where $P(\mathscr{D} \mid w)$ is the likelihood of the data and $P(w)$ is a prior distribution over w. The form of the functions $E_{\mathscr{D}}$ and $E_{\mathscr{W}}$ correspond to distributional assumptions regarding the data likelihood and prior distribution over network parameters respectively. The usual sum-of-squares metric (7), corresponds to a Gaussian likelihood,

$$P(\mathscr{D} \mid w) = \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left\{-\frac{[t_i - y(x_i)]^2}{2\beta^{-1}}\right\}$$

with fixed variance $\sigma^2 = 1/\beta$. Here, we adopt the Laplace prior propounded by Williams, 1995, which corresponds to a L_1 norm regularisation term,

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

where W is the number of model parameters (i.e. weights). An interesting feature of the Laplace regulariser is that it leads to the automatic pruning of redundant model parameters. From (8), 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 a data misfit sensitivity of α/β is set exactly to zero and can be pruned from the network.

4.1 Eliminating Regularisation Parameters

The hyper-parameters α and β can be estimated by maximising the evidence (MacKay, 1992b) or alternatively may be integrated out analytically (Buntine & Weigend, 1991; Williams, 1995). Here we take the latter approach; the posterior distribution of the parameters is given by

$$p(w) = \int p(w|\alpha)p(\alpha)d\alpha. \tag{9}$$

Assuming the Laplace prior, the prior distribution over the weights of the network, conditioned on the regularisation parameter α, is given by,

$$p(w|\alpha) = Z_{\mathscr{W}}(\alpha)^{-1}\exp\{-\alpha E_{\mathscr{W}}\} \tag{10}$$

where the necessary normalising constant is given by

$$Z_{\mathscr{W}}(\alpha) = \left(\frac{2}{\alpha}\right)^{W}. \tag{11}$$

Substituting equations (10) and (11) into equation (9), adopting the (improper) uninformative Jeffreys prior, $p(\alpha) = 1/\alpha$ (Jeffreys, 1939), and noting that α is strictly positive,

$$p(w) = \int_0^\infty 2^{-W}\alpha^{W-1}\exp\{-\alpha E_{\mathscr{W}}\}d\alpha.$$

Using the Gamma integral, $\int_0^\infty x^{\nu-1}e^{-\mu x}dx = \frac{\Gamma(\nu)}{\mu^\nu}$ (Gradshteyn & Ryzhic, 1994, equation 3.384), we obtain

$$p(w) = \frac{\Gamma(W)}{(2E_{\mathscr{W}})^W}.$$

Taking the negative logarithm and omitting irrelevant constant terms,

$$-\log p(w) = W \log E_{\mathscr{W}}. \tag{12}$$

Applying a similar treatment to the data misfit term (assuming a sum-of-squares error), we have

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

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:

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

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 are not normally regularised. This approach provides a sound basic approach to non-linear regression using multi-layer perceptron networks, with Bayesian regularisation to prevent over-fitting *and* automatic selection of an appropriate network architecture as a result of the Laplace prior. As the regularisation parameters are integrated out analytically, the user need only select the initial number of hidden layer units, and more importantly an appropriate data misfit term that represents any available prior knowledge regarding the form of the noise process contaminating the data. Using a large number of hidden units in the initial network helps to avoid local minima, as it increases the probability of starting with hidden layer units that approximate useful features of the data, but of course increases training time. A useful rule of thumb is to experiment with the initial size of the hidden layer until the Laplace prior on average prunes approximately half of them away. This ensures that the hidden layer will be large enough, without undue computational expense. Otherwise, the network architecture is automatically determined by the Bayesian regularisation scheme.

4.2 Choice of Data Misfit Term

In this paper, we are concerned with modelling predictive uncertainty, and so rather than simply estimating the conditional mean of the target data, we seek to construct a model such that the output specifies the entire predictive distribution. A sensible first step in solving an inference problem is to select an appropriate likelihood function to describe the statistical properties of the target data (c.f. MacKay & Gharahmani, 2005). The training criterion for the neural network should then be based on the negative logarithm of a parametric likelihood function, that incorporates any distributional assumptions regarding the noise process suggested by our prior knowledge of the data. In order to obtain a predictive distribution, we simply construct a network with one output for each of the parameters of this likelihood.

The most basic likelihood, providing a measure of the data misfit, used in this study, assumes a heteroscedastic (input dependent variance) Gaussian noise process, i.e.

$$E_{\mathscr{D}} = \sum_{i=1}^{\ell} \left\{ \log \sigma(x_i) + \frac{[\mu(x_i) - t_i]^2}{2\sigma^2(x_i)} \right\}. \tag{13}$$

Note the multi-layer perceptron network now has two output units, one giving the conditional mean of the target distribution, $\mu(x)$, as before, and an additional unit giving the conditional standard deviation, $\sigma(x)$. A linear activation function is used in the output unit corresponding to $\mu(x)$, and an exponential activation function for the unit corresponding to $\sigma(x)$, to enforce strictly positive estimates of conditional variance. This approach provides two advantages: Firstly the estimates of conditional variance provide error bars, indicating the uncertainty of model predictions (Nix & Weigend, 1994, 1995; Williams, 1996). Secondly the output of the model now completely specifies the target distribution, so the regularisation parameter β is no longer necessary. This data-misfit term is appropriate for regression on temperature data, where a Gaussian noise process is intuitively reasonable, but where the variability in temperature as well as the expected temperature may depend on, for example, the time of year.

The concentration of atmospheric pollutants provides an example of a type of data where a more complex likelihood may be appropriate. Clearly a pollutant concentration cannot be negative, and the uncertainty in predictions is likely to be skewed upward. A common ploy would be to implement a log-normal likelihood, by simply taking the logarithm of the target data and employing the data misfit given in (13).

Modelling frontal precipitation data requires a more sophisticated statistical model, and is often modelled using a Gamma distribution (Stern & Coe, 1984) or a mixture of exponentials (Woolhiser & Pegram, 1979). In this paper we adopt the hybrid Bernoulli/Gamma error metric proposed by Williams, 1998. The distribution of the amount of precipitation, X, is modelled by

$$P(X > x) = \begin{cases} 1 & \text{if } x < 0 \\ \alpha\Gamma\left(\nu, \frac{x}{\theta}\right) & \text{if } x \geq 0 \end{cases} \tag{14}$$

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 $\alpha(x_i)$ and the scale, $\theta(x_i)$, and shape, $\nu(x_i)$, 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 their respective constraints.

5 Exploiting Predictive Uncertainty

Environmental modellers are commonly interested in the impacts of extreme events, for example the impact of changes in future climate on local rainfall and subsequently on the flood hazard in susceptible catchments. 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 (Hewitson & Crane, 1996; Wilby et al., 1998; Giorgi & Mearns, 1999; Zorita & Storch, 1999). 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, based on the historical record. 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 order to estimate the impacts of changes in future climate on flood hazard, the predictions of a general circulation model are downscaled to provide predictions of future precipitation patterns, which in turn are processed by a hydrological model to assess the effect of changes in rainfall patterns on water-levels in the river fed by the catchment being studied. In this example, we will consider a fictitious catchment[2] in which there is a flood hazard if the three-day total precipitation is in excess of 35 cm. Figure 7 shows a plot of the financial loss associated with flood events as a function of the three-day total precipitation; the loss is modelled as a a constant component that is incurred whenever the river is unable to contain the run-off, and a component that reflects the additional damage resulting from increasingly severe flood events.

[Figure 7 about here.]

Figure 8 shows the three-day total precipitation time series for the study catchment area for the

[2] The results are actually based on downscaled predictions for a real precipitation time series data from Newton Rigg, a rather wet station in the North West of the United Kingdom.

period 1979-1993. Note that many of the apparent dry spells are caused by missing data in the historical record rather than the absence of precipitation and are not included in the analysis. The measured loss for the observed time series is 49.02 units.

[Figure 8 about here.]

Figure 9 shows the predicted three-day total precipitation based on a conventional neural network downscaling model trained to estimate the conditional mean of the target distribution. The network was trained on two segments of the precipitation time series spanning the periods 1961–1978 and 1994–2000. Note that the conditional mean systematically under-predicts the extreme rainfall events, as the predictive distribution is highly skewed. As a result, the predicted loss according to the simple neural network downscaling model is only 8.22 units, which severely under-estimates the true loss.

[Figure 9 about here.]

A second neural network downscaling model was trained, this time using the hybrid Bernoulli/Gamma data misfit term (14). In this case, the model has three outputs, one supplies an estimate of the probability of rainfall and two that define a Gamma distribution modelling the plausibility of different amounts of rainfall. As this model provides a full probabilistic prediction, it is possible to generate synthetic precipitation time series, using the neural network as a conditional weather generator model. In order to infer the expected loss associated with the flood hazard, a Monte Carlo simulation is conducted using 100,000 synthetic precipitation time series generated by the network. Figure 10 shows a histogram of the measured losses from the Monte Carlo simulation, clearly the actual loss of 49.02 units is plausible, given the prediction distribution of loss. The expected loss, via Monte-Carlo integration, is 70.72 units, which is much closer to the recorded loss.

[Figure 10 about here.]

While this example is deliberately somewhat contrived, it does demonstrate that a probabilistic characterisation of the uncertainty of model predictions can be exploited in impact studies, especially where the principal focus lies on the implications of extreme events, which by their very nature are not modelled well by the conditional mean. The integration over sources of uncertainty also provides the results in a format that is well suited to the needs of end-users, such as government institutions or the insurance industry. Clearly the distribution of plausible losses is exactly the information required by such users for well-informed policy-making and forward planning.

6 The Predictive Uncertainty in Environmental Modelling Challenge

The WCCI-2006 predictive uncertainty in environmental modelling challenge consisted of one `SYNTHETIC` benchmark dataset and three real-world environmental datasets `PRECIP`, `SO2` and `TEMP`. The format of the competition was based closely on the regression problems of the earlier Pascal predictive uncertainty challenge. The negative log-likelihood of the test data was used as the performance criterion for the final ranking of submissions, as it is the natural measure of the fit of a distribution to a set of data. Two standard methods were available for describing the predictive distribution for each pattern, the mean and variance of a Gaussian predictive distribution, or a set of quantiles, allowing the definition of an arbitrary predictive distribution[3]. An unusual feature of the competition is that the competitors had the option of suggesting alternate forms for specifying the predictive distribution (as the likelihood can be described in any number of parametric forms). A mixture Gaussian option was added at a late stage in the competition in response to a request from one of the competitors. The target data for all three of the real-world environmental benchmark datasets are (finely) quantised, for example precipitation data is only measured to the nearest 0.1 mm. In principle it would therefore be possible to make the negative log-likelihood arbitrarily low by specifying the predictive distribution (via quantiles) as a set of delta functions centred on the quantised values. This technique was employed by some entries to the original Pascal predictive uncertainty challenge. In order to prevent this, the minimum

[3] The computation of the negative log likelihood using these representations is discussed in detail on the website for the original Evaluating Predictive Uncertainty Challenge, `http://predict.kyb.tuebingen.mpg.de/pages/evaluation.php`

allowable width of the quantiles (and similarly the variances of the individual components of a mixture Gaussian predictive distribution) were limited to match the quantisation interval used.

6.1 Reference Submissions

Three baseline models were submitted for each dataset, which gave a fixed predictive distribution for all patterns: `Baseline #1` - fixed Gaussian predictive distribution specified via the unconditional mean and variance of the target data, `Baseline #2` - fixed Gaussian distribution specified as a set of quantiles and `Baseline #3` - fixed predictive distribution specified by quantiles representing the empirical distribution of the target data. A fourth baseline model was created for the `SO2` benchmark, giving a fixed predictive distribution for all patterns based on a Gaussian mixture model of five components, fitted using the standard Expectation Maximisation (EM) algorithm (Dempster, Laird, & Rubin, 1977), as implemented by the NETLAB package (Nabney, 2004). In addition to these baseline models, neural network models were also submitted for each benchmark, the training procedure used is described in Section 4. A heteroscedastic Gaussian data mis-fit term (13) was used for the `SYNTHETIC` and `TEMP` benchmarks, a heteroscedastic log-normal term for the `SO2` benchmark and the hybrid Bernoulli/Gamma term (14) term for the `PRECIP` benchmark. In order to avoid training difficulties due to local minima of the cost function, 20 models were trained in each case, with randomly initialised weights, and the model giving the lowest value for the regularised loss retained. These models provide an indication of the “minimum” and “competitive” levels of performance for each benchmark.

Table 2 shows the negative log of the estimated density of the true labels (NLPD), i.e. the negative log-likelihood, of the training and test sets of the `SYNTHETIC` benchmark for selected entries. It can be seen that many of the entries were able to make clear improvements in modelling the predictive distribution over the baseline models, with the best models approaching the performance of the optimal “ground truth” model used to generate the data. However, the `SYNTHETIC` benchmark is relatively straight-forward, the only unusual feature being the heteroscedasticity of the noise process. The best models all include an explicit model of the variance of the target distribution.

[Table 2 about here.]

Table 3 shows the results obtained by leading entries to the challenge. In this case the use of a sum-of-squares model, with a normal predictive distribution, performs very poorly as it is unable to account for the peak in the true probability density function caused by the days on which no rainfall was observed.

[Table 3 about here.]

Table 4 shows the test set mean-squared error and negative log likelihood for selected models over the SO2 benchmark. Clearly this is the noisiest of the benchmark datasets, and while some reduction in the mean-squared-error is possible, it is difficult to produce a model that improves on the baseline models in terms of the quality of the predictive distribution.

[Table 4 about here.]

In this case a heteroscedastic Gaussian noise process is a reasonable assumption. Table 5 shows the test set MSE and NLPD statistics for selected models, in almost all cases the models significantly improve on the baseline models in terms of the NLPD.

[Table 5 about here.]

The final standings in the competition, decided by mean NLPD score over the three environmental datasets, are shown in Table 6. The overall winner is Markus Harva. This criterion is a natural choice as it represents the joint likelihood of all of the data, given the models.

[Table 6 about here.]

7 Areas for Further Research

7.1 Inherent Bias in the Conditional Variance

It is well known that estimates of the conditional variance are likely to be significantly biased. If the model of the conditional mean over-fits the data, this reduces the apparent local noise density, and so error bars based on the conditional variance will be unrealistically narrow. This problem has previously been addressed via Bayesian approaches (Bishop & Qazaz, 1996; Goldberg, Williams, & Bishop, 1998), and by the use of leave-one-out cross-validation (Cawley, Talbot, Foxall, Dorling, & Mandic, 2004). However these approaches are currently only suitable for relatively small scale applications, with only a few thousands of training patterns. Further research is needed to develop large scale algorithms suitable for environmental applications, where much larger amounts of data are typically available.

7.2 Incorporating the Uncertainty in the Model Parameters

In this paper we have reviewed the use of maximum-likelihood based loss functions for neural networks, which allow us to incorporate prior knowledge regarding the uncertainty in model predictions due to the inherent noise process contaminating the data. Another important source of uncertainty lies in the uncertainty due to the estimation of the model parameters from a finite sample of data. It seems likely that a better model of the predictive distribution might be obtained by including this effects of the uncertainty in the model parameters, e.g. via the Laplace approximation (MacKay, 1992a, 1992b) or via Markov-Chain Monte Carlo methods (Neal, 1996).

7.3 The Form of the Predictive Distribution

While expert knowledge is sometimes available regarding the form of the noise process contaminating the data, it would be useful also to have a data-driven approach, where the form of the noise process is also inferred from the training data. The mixture density network (Bishop, 1994), where the outputs of the model specify the components of a Gaussian mixture model of the predictive distribution, represents the most basic approach. The warped Gaussian Process, (Snelson, E., & Ghahramani, 2004), in which the observation space is transformed so as to be well modelled as a Gaussian process, represents a more recent approach.

8 Conclusions

In this paper we have demonstrated that a model of the predictive distribution can be exploited in studies of the impacts of changes in future climate, via a somewhat contrived, but nevertheless illustrative example. An on-line competition has been organised in an attempt to promote research on methods for estimating the uncertainty inherent in statistical predictions. The results demonstrate that this is a difficult topic, where standard approaches do not yield uniformly good results. We hope that the competition has gone some way to highlight an area where further research is likely to produce practical benefits in the analysis of environmental data.

Acknowledgments

Many thanks are due to Joaquin Quinonero Candela, Carl Rasmussen and Yoshua Bengio for organising the original (and very successful) Evaluating Predictive Uncertainty Challenge, and to Vladimir Cherkassky, Julio Valdes, Vladimir Krasnopolsky and Dimitri Solomatine for organising the special session on Applications of Learning and Data-Driven Methods to Earth Sciences and Climate Modeling at IJCNN-2005. These two events provided the impetus (and the format) for this competition. Thanks to the anonymous reviewers for their helpful comments. Thanks are also due to Radford Neal, Iain Murray and Markus Harva for their helpful feedback on the format of the competition and to Peter Williams for provoking our initial interest in modelling predictive distributions. The competition would not have been possible without Nicola Talbot, who implemented the competition website and assisted in the typesetting of the manuscript. Lastly, many thanks to all those who took part!

References

Aitken, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, *36*(3), 332–339.

Bhattacharya, B., & Solomatine, D. P. (2005, January). Neural networks and M5 model trees in modelling water level-discharge relationship. *Neurocomputing*, *63*, 381–396.

Bishop, C. M. (1994). *Mixture density networks* (Tech. Rep. No. NCRG/94/004). Neural Computation Research Group, Aston University.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bishop, C. M., & Qazaz, C. S. (1996, July 16–19). Bayesian inference of noise levels in regression. In C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, & B. Sendhoff (Eds.), *Proceedings of the international conference on artificial neural networks (icann-96)* (Vol. 1112, pp. 59–64). Bochum, Germany: Springer.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, *26*(2), 211–246.

Buntine, W. L., & Weigend, A. S. (1991). Bayesian back-propagation. *Complex Systems*, *5*, 603–643.

Cawley, G. C., Dorling, S. R., Jones, P. D., & Goodess, C. (2003, April 23–25). Statistical downscaling with artificial neural networks. In *Proceedings of the european symposium on artificial neural networks (esann-2003)* (pp. 167–172). Bruges, Belgium.

Cawley, G. C., Talbot, N. L. C., Foxall, R. J., Dorling, S. R., & Mandic, D. P. (2004, March). Heteroscedastic kernel ridge regression. *Neurocomputing*, *57*, 105–124.

Dawson, C. W., See, L. M., Abrahart, R. J., Wilby, R. L., Shamseldin, A. Y., Anctil, F., et al. (2005, 31 July - 4 August). A comparative study of artificial neural network techniques for river stage forecasting. In *Proceedings of the ieee international joint conference on neural networks (ijcnn '05)* (Vol. 4, pp. 2666–2670).

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

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

Giorgi, F., & Mearns, L. O. (1999). Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research*, *104*, 6335–6352.

Goldberg, P., Williams, C., & Bishop, C. (1998). Regression with input-dependent noise: A Gaussian process treatment. In M. Kearns, M. Jordan, & S. Solla (Eds.), *Advances in neural information processing systems* (Vol. 10, pp. 493–499). Cambridge, MA: MIT Press.

Gradshteyn, I. S., & Ryzhic, I. M. (1994). *Table of integrals, series and products* (fifth ed.; A. Jeffrey, Ed.). Academic Press.

Harpham, C., & Wilby, R. L. (2005, October). Multi-site downscaling of heavy daily precipitation occurrence and amounts. *Journal of Hydrology*, *312*(1–4), 235–255.

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L., & Goodess, C. (2006). Downscaling heavy precipitation over the UK: A comparison of dynamic and statistical methods and their future scenarios. *International Journal of Climatology* (in press).

Hewitson, B. C., & Crane, R. G. (1996). Climate downscaling: Techniques and application. *Climate Research*, *7*, 85–95.

Jeffreys, H. S. (1939). *Theory of probability*. Oxford University Press.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

MacKay, D. J. C. (1992a). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447.

MacKay, D. J. C. (1992b). A practical Bayesian framework for backprop networks. *Neural Computation*, *4*, 448–472.

MacKay, D. J. C., & Gharahmani, Z. (2005). *Comments on 'maximum likelihood estimation of intrinsic dimension' by Levina and Bickel.* `http://www.inference.phy.cam.ac.uk/mackay/dimension`.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (second ed., Vol. 37). Chapman & Hall.

Nabney, I. (2004). *NETLAB: Algorithms for pattern recognition*. Springer.

Neal, R. M. (1996). *Bayesian learning for neural networks*. New York: Springer.

Nix, D. A., & Weigend, A. S. (1994). Estimating the mean and variance of the target probability distribution. In *Proc., int. conf. on neural networks* (Vol. 1, pp. 55–60).

Nix, D. A., & Weigend, A. S. (1995). Learning local error bars for nonlinear regression. In *Advances in neural information processing systems* (Vol. 7, pp. 489–496). MIT Press.

Nunnari, G., Dorling, S. R., Schlink, U., Cawley, G., Foxall, R., & Chatterton, T. (2004, October). Modelling SO_2 concentration at a point with statistical approaches. *Environmental Modelling and Software*, *19*(10), 887–905.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., et al. (2003, July). A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Envrionment*, *37*(23), 3237–3253.

Snelson, E., E., R. C., & Ghahramani, Z. (2004). Warped gaussian processes. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems* (Vol. 16, pp. 337–344). Cambridge, MA: MIT Press.

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

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. New York: John Wiley.

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., et al. (1998). Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, *34*, 2995–3008.

Williams, P. M. (1991, February). *A Marquardt algorithm for choosing the step-size in backpropagation learning with conjugate gradients* (Cognitive Science Research Paper No. CSRP-229). University of Sussex, Brighton, U.K.

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

Williams, P. M. (1996). Using neural networks to model conditional multivariate densities. *Neural Computation*, *8*, 843–854.

Williams, P. M. (1998). Modelling seasonality and trends in daily rainfall data. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems - proceedings of the 1997 conference* (Vol. 10, pp. 985–991). MIT Press.

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

Yeo, I.-K., & Johnson, R. A. (2000, December). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959.

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

List of Figures

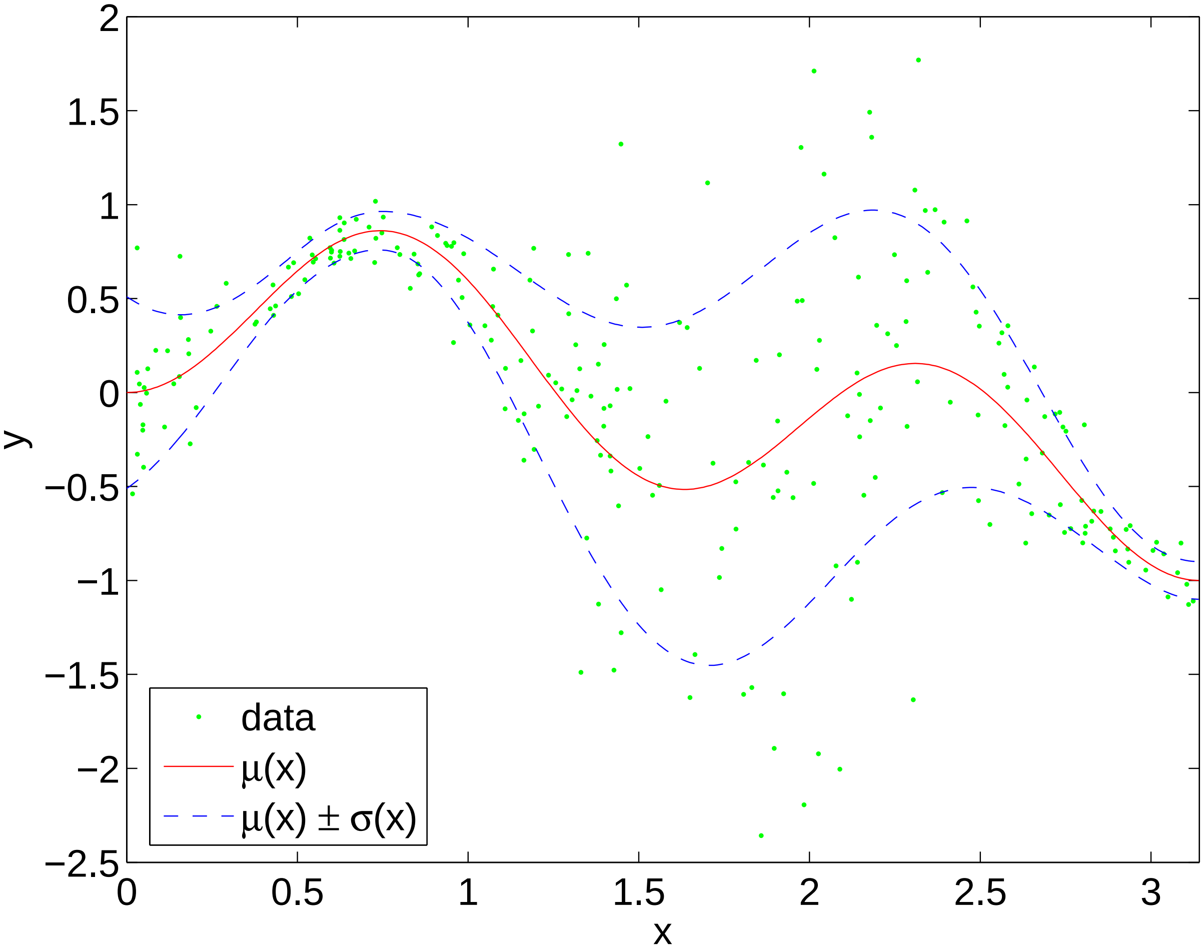


Figure 1: Plot of the training data for the `SYNTHETIC` benchmark dataset, along with an indication of the true conditional mean, $\mu(x)$ and conditional standard deviation, $\sigma(x)$.

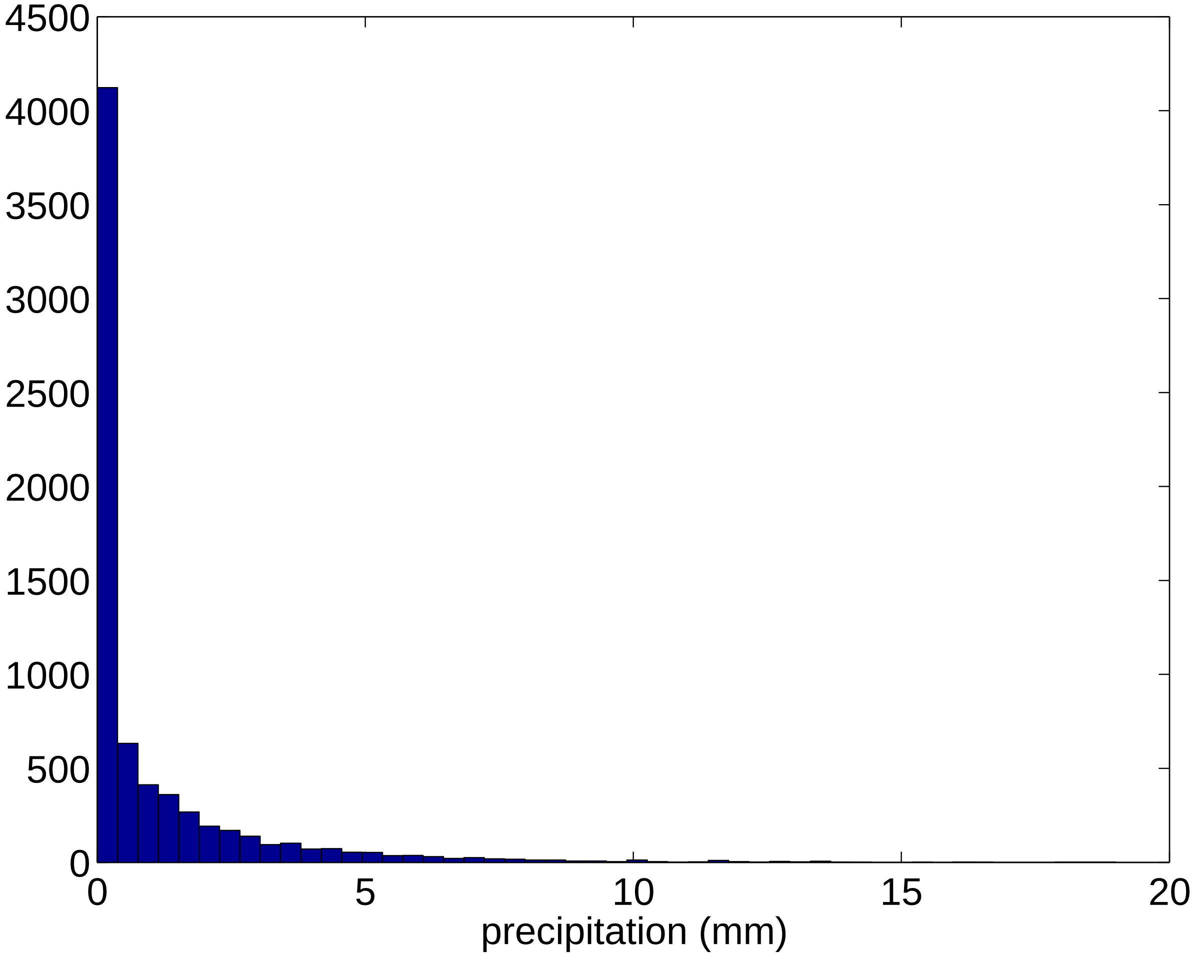


Figure 2: Histogram of the target data for the training set of the `PRECIP` benchmark dataset.

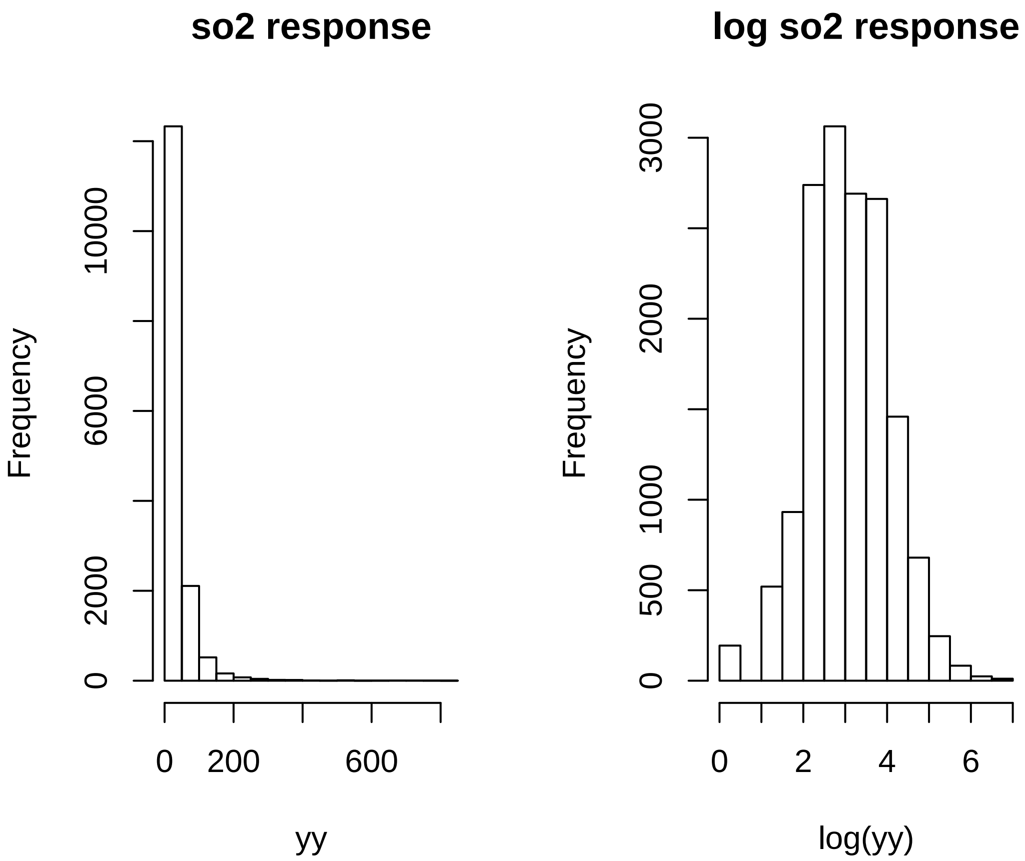


Figure 3: Histograms of the raw and transformed response variable for the SO_2 dataset.

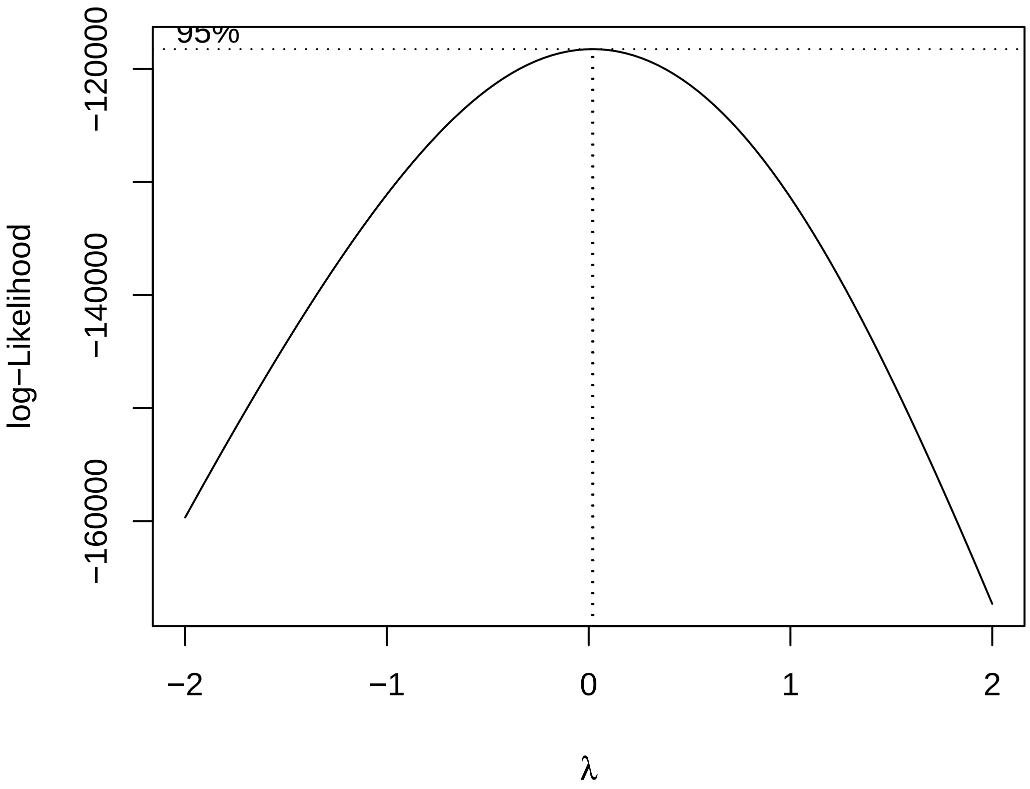
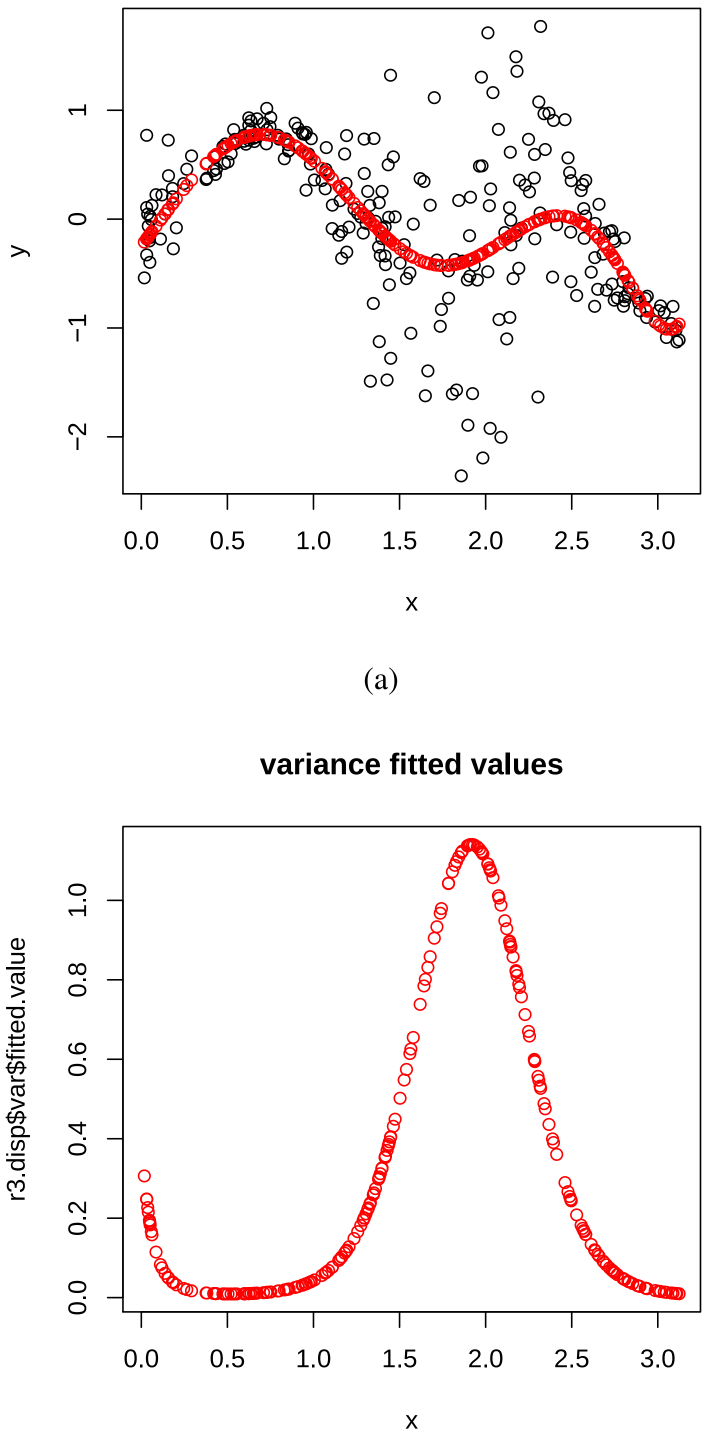


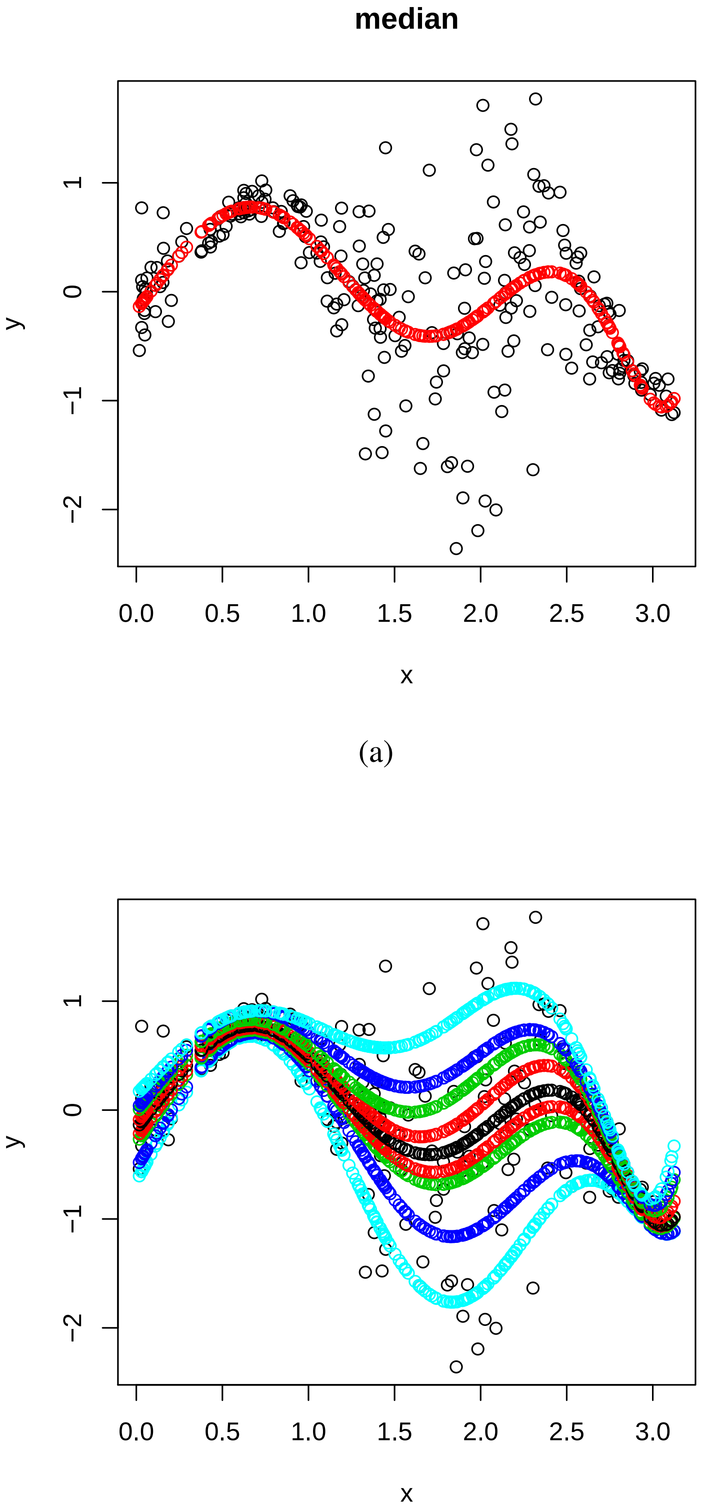
Figure 4: Profile likelihood for the SO2 benchmark, using the Box-Cox transformation to improve normality.



(a)

(b)

Figure 5: Fitted values for the conditional mean and variances of the synthetic benchmark using Aitken's dispersion model.



(a)

(b)

Figure 6: Fitted values for the conditional median and deciles of the synthetic benchmark using quantile regression.

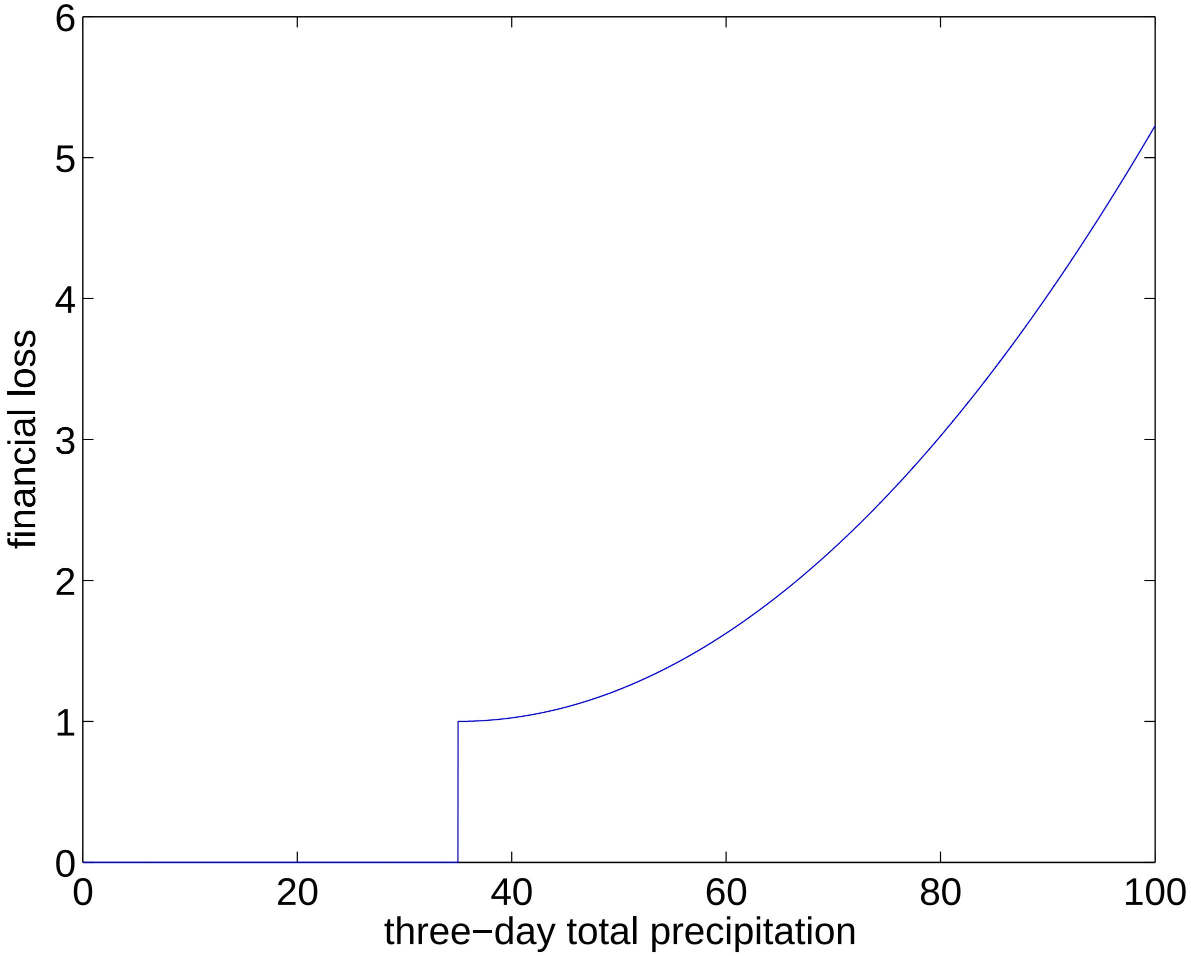


Figure 7: Financial loss associated with flood events in a susceptible catchment as a function of the three-day total precipitation.

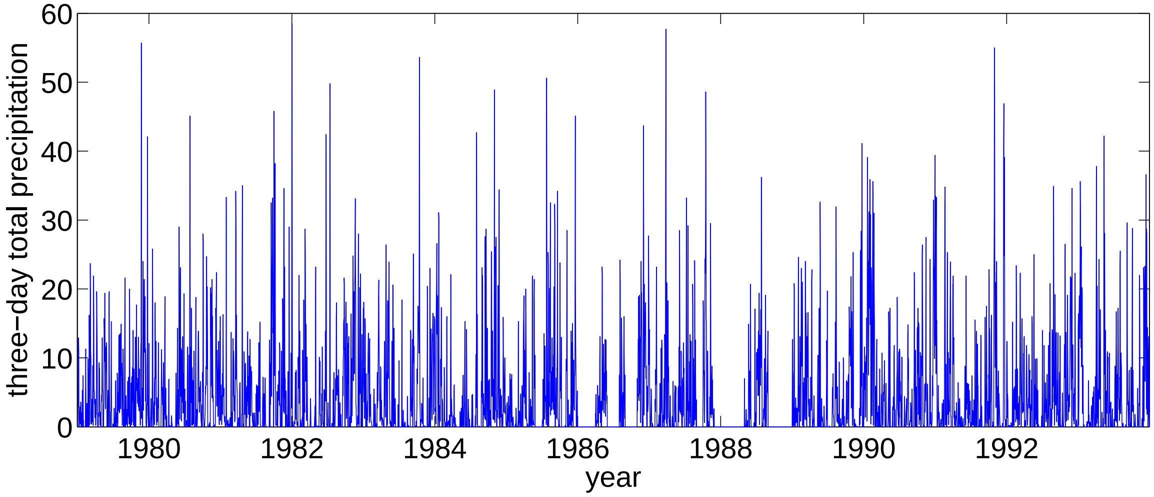


Figure 8: Three-day total precipitation time series for a catchment area susceptible to flooding.

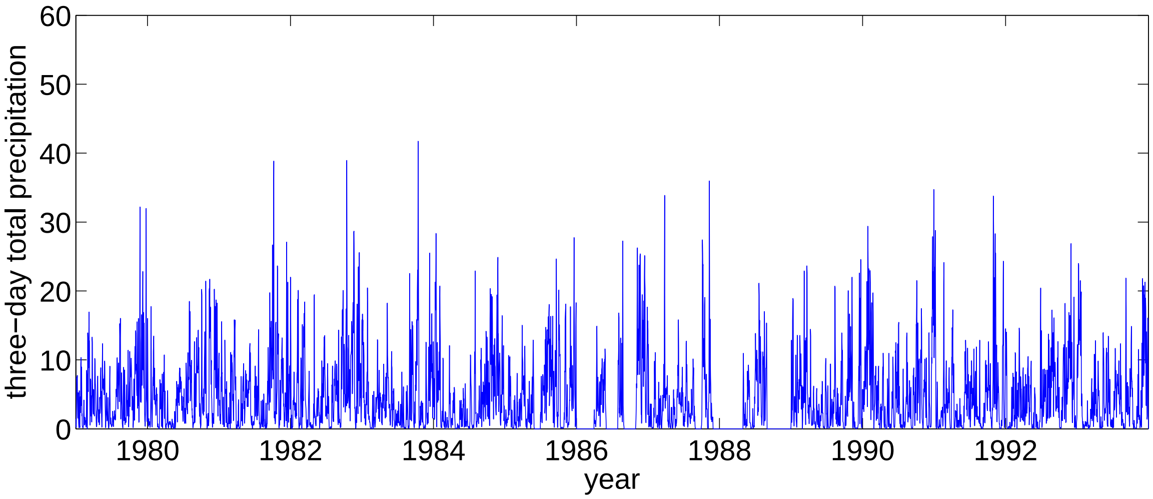


Figure 9: Predicted three-day total precipitation time series for a catchment area susceptible to flooding, using a neural network providing the conditional mean of the target distribution.

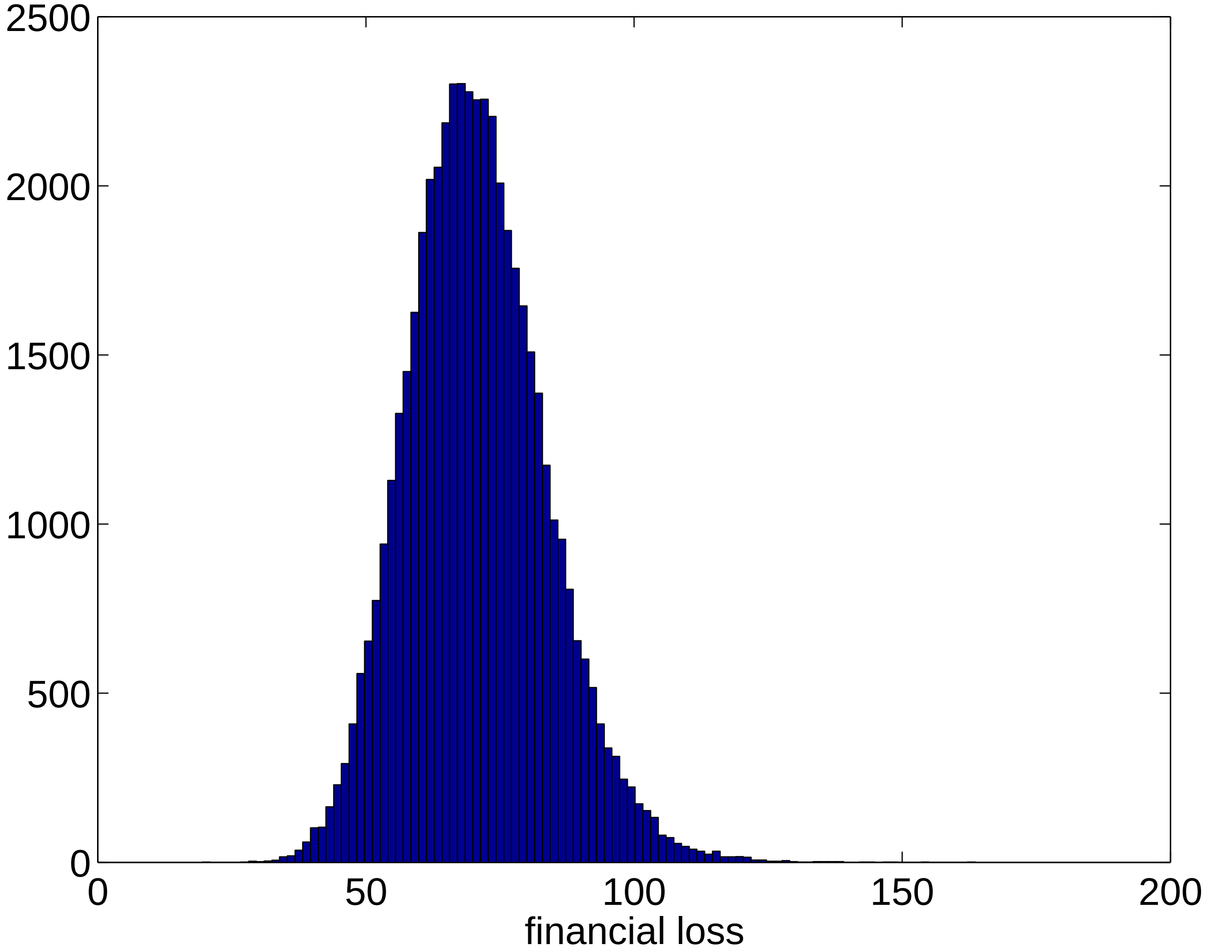


Figure 10: Distribution of expected financial loss associated with flood events in a susceptible catchment.

List of Tables

Table 1: Canonical links for common distributions

Distribution	Link
Normal	$\eta = \mu$
Poisson	$\eta = \log(\mu)$
Binomial	$\eta = \log\{\pi/(1-\pi)\}$
Gamma	$\eta = \mu^{-1}$
Inverse Gaussian	$\eta = \mu^{-2}$

Table 2: Training and test set negative log-likelihood statistics for the SYNTHETIC benchmark.

Name	**Method**	**Train NLPD**	**Test NLPD**
Reference	ground truth	0.3333	0.3489
Harva	varmlp (MoG)	0.3251	0.3858
Cawley	MLP	0.3083	0.4046
Kurogi *et al.*	CAN2 ensemble + CV	0.2236	0.4304
Boardman	Support Vector Regression	0.4150	0.4745
Nikulin	CM+GbO	0.3590	0.4805
Bagnall	YJ	1.0081	1.0313
Reference	Baseline #1	1.1064	1.1357
Reference	Baseline #2	1.1104	1.1374
Reference	Baseline #3	0.7923	1.2324

Table 3: Test set mean-squared error (MSE) and negative log-likelihood (NLPD) statistics for the `PRECIP` benchmark.

Name	Method	MSE	NLPD
Cawley	MLP	0.6305	-0.5095
Harva	varmlp	5.4493	-0.2792
Reference	Baseline #1	1.0002	-0.1772
Takeuchi	Kernel QR	0.6109	0.7469
Bagnall	YJ	2.1072	1.1139
Nikulin	CM+GbO	0.6539	1.2724
Boardman	Support Vector Regression	0.6441	1.6055
Reference	Baseline #2	1.0001	2.0346
Reference	Baseline #3	1.0001	2.0496
Kurogi *et al.*	CAN2 ensemble + CV + hetero + quantile	0.6465	3.0982

Table 4: Test set mean-squared error (MSE) and negative log-likelihood (NLPD) statistics for the `SO2` benchmark.

Name	**Method**	**MSE**	**NLPD**
Cawley	MLP	0.7985	4.2550
Harva	varmlp	0.8333	4.3702
Reference	Baseline #4	1.0000	4.4964
Reference	Baseline #1	1.0001	4.4968
Nikulin	CM+GbO	0.8576	4.6162
Bagnall	YJ	1.7598	4.7578
Boardman	Support Vector Regression	0.8396	5.0897
Reference	Baseline #3	1.0000	5.1655
Reference	Baseline #2	1.0000	5.2181
Takeuchi	Kernel QR	0.6884	6.0425
Kurogi *et al.*	CAN2 ensemble + CV + hetero + quantile	0.7807	11.0063

Table 5: Test set mean-squared error (MSE) and negative log-likelihood (NLPD) statistics for the TEMP benchmark.

Name	**Method**	**MSE**	**NLPD**
Snelson	Sparse pseudo-input Gaussian process (SPGP)	0.0661	0.0348
Cawley	MLP	0.0693	0.0530
Kurogi *et al.*	CAN2 ensemble + CV + hetero + quantile + outlier	0.0681	0.0591
Boardman	Support Vector Regression	0.0709	0.0760
Nikulin	CM+GbO	0.0729	0.1076
Bagnall	Linear Regression	0.077432	0.136235
Harva	varmlp	0.0925	0.2015
Whittley	QuantLin	24.9839	0.6251
Reference	Baseline #1	1.0000	1.3004
Reference	Baseline #2	1.0000	1.4151
Reference	Baseline #3	1.0000	1.4177
Takeuchi	Kernel QR	0.0965	24.7922

Table 6: Final standings in the competition - the overall winner, decided by mean NLPD score, is Markus Harva.

Name	**PRECIP**	**SO_2**	**TEMP**	**Mean**
Cawley	-0.5095	4.2550	0.0530	1.2661
Harva	-0.2792	4.3702	0.2015	1.4308
Nikulin	1.2724	4.6162	0.1076	1.9987
Bagnall	1.1139	4.7578	0.1362	2.0026
Boardman	1.6055	5.0897	0.0760	2.2571
Kurogi *et al.*	3.0982	11.0063	0.0591	4.7212
Takeuchi	0.7469	6.0425	24.7922	10.5272
Whittley	∞	∞	0.6251	∞
Snelson	∞	∞	0.0348	∞