Non-linear modelling of river levels using the Gamma test

Peter J Durrant and Antonia J Jones

Department of Computer Science, Cardiff University,

P.O. Box 916, Cardiff, CF24 3XF, Wales, UK,

Contact regarding publication should be made to:

Professor Antonia J Jones

Department of Computer Science, Cardiff University,

P.O. Box 916, Cardiff, CF24 3XF, Wales, UK

tel: +44 292 087 4812

fax: +44 292 087 4598

October 20, 2002

Abstract

We constructed non-linear predictive models for the River Kennet at Theale and the River Thames at Windsor using river and precipitation data from the Thames Valley region in the UK. Our approach used a novel non-linear data analysis technique called the Gamma test, combined with heuristic search techniques to provide a practical solution to the problem of constructing forward predictions of river levels and flows. Our three hour predictive model for the River Kennet at Theale calculated the level to a standard error of 1 cm, and our four hour predictive model for the River Thames at Windsor calculated the level to a standard error of 3cm.

The Gamma test is used to examine the relationship between inputs and outputs in numerical data-sets. It is used prior to modelling to estimate the variance of the output that cannot be accounted for by the existence of any smooth model based on the inputs, even though the model is unknown. This error variance estimate provides a target Mean Squared Error that any smooth non-linear model should attain on unseen data. Building a model with greater accuracy than the error variance indicated by the Gamma test will result in a model that has overtrained on the data set and which cannot generalize well for unseen data.

Keywords: Data analysis, Gamma test, modelling, prediction, Thames river, flood warning.

1 Introduction

In this paper we describe the application of a new non-linear analysis technique, called the *Gamma test*, to the problem of modelling and predicting river level and flow. This work, drawn from [Durrant, 2001], was a feasibility study for the *MAPFLOWS* (Modular Automated Prediction and Flood Warning System) project. We studied an area of the River Thames (UK) and a tributary, the River Kennet, with the goal of predicting the River Kennet at Theale and the

River Thames at Windsor. We used hourly rainfall, and river flow and level data gathered for one year at various points in the Thames Valley. This data was analysed using the Gamma test and modelled using neural networks.

We first briefly describe the Gamma test in section 2, then consider the rivers and surrounding area in section 3. In section 4 we describe the analysis of the data and in section 5 we show the modelling results.

2 The Gamma test

The Gamma test was first briefly reported in [Končar, 1997] and [Aðalbjörn Stefánsson et al., 1997], and later discussed and used in [Chuzhanova et al., 1998], [Guedes de Oliveira, 1999], [Tsui, 1999], [Tsui et al., 2002], [Durrant, 2001] and [Jones et al., 2002]. A formal proof was given in [Evans, 2001], [Evans and Jones, 2002a] and [Evans and Jones, 2002b].

The idea is quite distinct from earlier attempts at non-linear analysis. Suppose we have a set of input–output observations of the form

$$\{(\boldsymbol{x}_i, y_i) \,|\, 1 \leq i \leq M\} \tag{1}$$

where the inputs $\boldsymbol{x} \in \mathbb{R}^m$ are vectors confined to some closed bounded set $C \subset \mathbb{R}^m$ and, without loss of generality, the corresponding outputs $y \in \mathbb{R}$ are scalars.

Rather than pre-suppose some particular parametric form for the underlying non-linear model we suppose that it belongs to some *general class of functions*. In particular we suppose that the underlying relationship is of the form

$$y = f(x_1, \ldots, x_m) + r \tag{2}$$

where f is a suitably *smooth* and *unknown* function that maps the components of the input vector $\boldsymbol{x}$ to the output y and r is a stochastic variable which represents noise. We assume that the mean of the distribution of r is zero (since a constant bias can be subsumed into the unknown function f). Other reasonable restrictions are made, for example that the variance of r, $\mathrm{Var}(r)$, is bounded. Hence the domain of possible models is restricted to the *class of functions which have bounded first and second partial derivatives.*

Even though the underlying function f is unknown the Gamma test can estimate $\mathrm{Var}(r)$ directly from the data. This estimate, called the *Gamma statistic* and denoted by Γ, is calculated directly from the data in $O(MlogM)$ time. To compute Γ we calculate two quantities. First

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2 \quad \text{where} \quad 1 \le k \le p \tag{3}$$

where $N[i,k]$ denotes the index of the kth nearest neighbour to $\mathbf{x}_i$, and $|.|$ denotes Euclidean distance (typically $p = 10$). Thus $\delta_M(k)$ is the mean square distance to the kth nearest neighbour. Second

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} \left(y_{N[i,k]} - y_i\right)^2 \tag{4}$$

Here $y_{N[i,k]}$ is the output value corresponding to the kth nearest neighbour of $\boldsymbol{x}_i$. Finally we perform linear regression of the pairs $(\delta_M(k), \gamma_M(k))$ $(1 \le k \le p)$ and return the constant term of this line as the noise variance estimate Γ. The gradient A of the regression line is also returned, as this gives an indication of model complexity. It can be shown [Evans and Jones, 2002b, Evans and Jones, 2002a] that

$$\Gamma \to \mathrm{Var}(r) \quad \text{as} \quad M \to \infty \tag{5}$$

where the convergence is in probability. A useful graphical realization is the *Gamma scatter plot* obtained by plotting all points $(|\boldsymbol{x}_{N[i,k]} - \boldsymbol{x}_i|^2, \frac{1}{2}(y_{N[i,k]} - y_i)^2)$ and overlaying the regression line, see for example Figure 9.

The Gamma test is a non-parametric analysis technique and the results apply regardless of the particular methods used to subsequently build a model of f.

If, for example, the Gamma statistic is large (compared with the variance of y) then it is unlikely that a smooth model exists to map the inputs to the output, but if the Gamma statistic is close to zero then this becomes more likely. We can standardize the result by considering $\Gamma/\mathrm{Var}(y)$, which returns a scale invariant noise estimate which normally lies between zero and one[1], which we call V_{ratio}.

We determine the reliability of the Γ statistic by running a series of Gamma test for increasing M, to establish the size of data set required to produce a stable asymptote. This is known as an *M-test*.

It is obviously pointless to attempt to fit the model beyond the stage where the MSE on the training data is smaller than $\mathrm{Var}(r)$. This will cause *over-fitting*, and although the resulting model may perform virtually perfectly on the training data it will give poor prediction results on previously unseen data.

- Thus one problem of model construction solved by the Gamma test is at what point to cease training.

We can also use the M-test to decide how much data we require to build a model with a mean squared error which approximates the estimated noise variance.

However, the utility of the Gamma test goes beyond merely estimating the noise variance and providing a criterion for ceasing training: we can also use it to select the most predictively useful input variables. There are $2^m - 1$ possible selections of inputs that we might use to build a model. Provided m is not too large, say $m \leq 20$, we can run a Gamma test for each possible selection. (For larger m we can use a heuristic search technique, such as a genetic algorithm.) The selection of inputs which provides the Gamma statistic closest to zero is

[1]We say V_{ratio} is 'normally' in the range [0, 1] because if $\mathrm{Var}(r)$ is equal to (or close to) zero (or M is too small), it can happen that Γ derived from the algorithm is negative (in which case we might replace Γ by $|\Gamma|$ for our estimate of $\mathrm{Var}(r)$), similarly it is possible that $\Gamma > \mathrm{Var}(y)$.

then selected as the best. This is called a *full embedding* search.

In this study we used the software implementation[2] *winGamma* described in [Durrant, 2001]. We can quickly analyse continuous real valued data and determine the extent to which the data can be modelled by a smooth input-output model (bounded first and second partial derivatives). The software first computes the Gamma statistic, which provides an estimate for the target mean squared error (to be achieved on unseen data) by any smooth model built from the data. Using a variety of non-linear modelling tools the system can then quickly and efficiently construct such a model, using the Gamma statistic as a stopping criterion to prevent overtraining.

Once precipitation has occurred the process of runoff, although highly complex in any particular catchment area, is completely determined by physical and hydraulic processes, geomorpholocal processes, boundary and initial conditions, and any system parameter such as gating openings. Thus in many respects downstream water flow/level prediction is an important application ideally suited to the algorithms incorporated into *winGamma*. This is principally because once precipitation has occurred the entire water transport process to the sites for which prediction is required is essentially determined by a smooth (albeit complex) process with lags.

3 River level and flow prediction

River levels and flows can be forecast using indirect or direct methods. The indirect methods initially involves prediction of runoff either through a rainfall-runoff model or by routing the flow observed at an upstream gauge to the desired location downstream. The predicted runoff is later con-

[2] *winGamma*^{TM} is available under licence from the Department of Computer Science, Cardiff University, Wales UK.

verted to a water level by use of a rating curve. The rainfall-runoff models, for example [Kitadinis and Bras, 1980a, Kitadinis and Bras, 1980b] and [Georgakakos, 1986a, Georgakakos, 1986b], require knowledge of underlying hydrology and the establishment of many rain gauges together with a good telemetry system. Routing techniques are more useful when the travel time is longer and the downstream flow is low or controlled. For direct prediction of water levels, statistical correlation techniques have been employed [Mutreja et al., 1987].

These techniques can provide accurate results in a reasonably wide range of circumstances. However, the empirical models need careful construction for each particular catchment area and it would be hard to envisage a general purpose adaptive system that could proceed from such a basis.

[Thirumalaiah and Deo, 1988] used neural networks as a pattern recognition technique for river stage forecasting in the Godavari Basin (India) and their results showed that adaptive modelling for level prediction is quite practical. More recently [Wright and Dastorani, 2001] gave an account of a neural network based approach for ungauged catchment peak flow prediction.

Our approach will also be adaptive and use neural networks for non-linear modelling of the river system. It is worth noting that the same basic techniques might be applied to the management of hydroelectric reservoirs. An example might be the operation of a reservoir with an uncontrolled inflow but which has the means of regulating the outflow. If advance information regarding the inflow is available then the reservoir can be operated, for example by a rule based system, so as to optimize electricity production or minimize downstream flood damage.

3.1 The Thames Valley

The river system data used in this analysis was measured in the Thames river basin above Windsor, see Figure 1. The data was provided by the UK Envi-

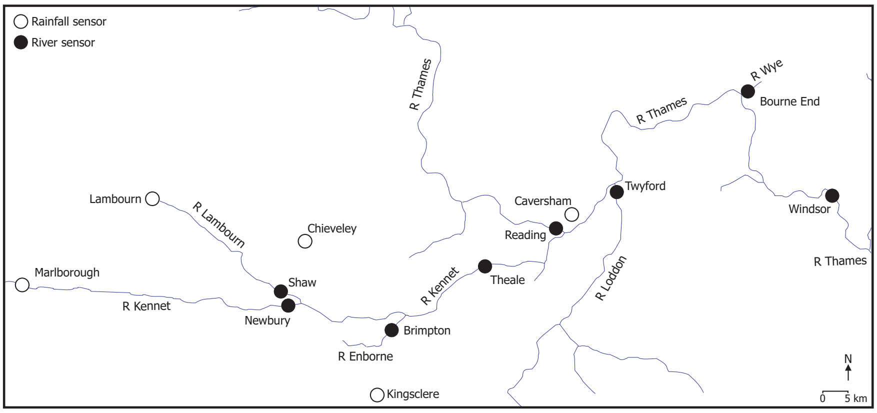


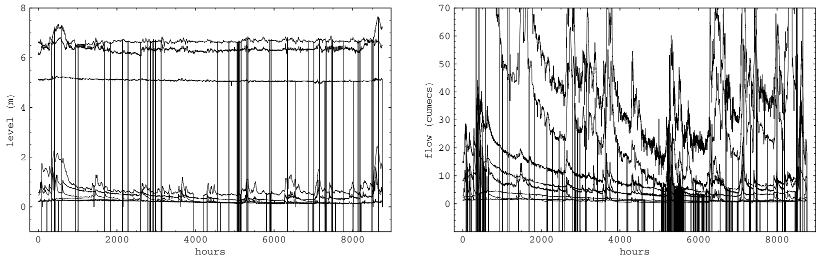
Figure 1: The Thames study area.

ronment Agency. It consists of flow rate (cubic metres per second) and level readings (metres) for the rivers at Newbury, Shaw, Brimpton, Theale, Reading, Twyford, Bourne End and Windsor. The rainfall, measured in mm/hour, at five sites in the region was also collected. All of the measurements were collected hourly over the year from 10am on 1st January 1999 to 9am on 1st January 2000. The river and rainfall sensor positions are marked on the map. The general direction of flow is from west to east.

3.2 Data Pre-processing

The data was first scanned for sensor malfunctions, missing values or data entry errors. A plot of the raw data in Figure 2 over the entire period gives a general indication of data quality for the river level and flow measurements. In some cases these data errors continue for significant intervals. Their cause is unknown, but could be due to sensor failure or periods of routine maintenance.

It was apparent that the data would have to be cleaned to reduce the effect of faulty sensor readings prior to analysis or modelling. We identified 'incorrect measurements' as being those that were significantly different from the previous



(a) The raw river level data. (b) The raw river flow rate data.

Figure 2: The raw river level and flow rate data. Faulty sensor readings are indicated on the graphs by the plunging vertical lines.

reading taken at the same site. A simple thresholding algorithm[3], designed to operate in real-time when future sensor values would not be known, was used to correct obviously faulty readings by replacing them with their last known reliable value. This routine is effective with time-series values that can be expected to change relatively smoothly over time. It is obviously inappropriate for the rainfall measurements, where the values were effectively discontinuous, and we did not attempt to adjust the rainfall data.

An illustrative graph of the cleaned level data is given in Figure 3. Similar graphs were obtained for the cleaned flow data. In most cases where there were single missing values the data cleaning procedure provided a simple and effective approximation. A disadvantage of this technique occurs when a string of missing values are assigned the last valid measured value. For long strings this algorithm would in all likelihood produce increasingly inaccurate approximations.

3.3 Sensor consistency

Apart from the issue of data errors, we also discovered an issue regarding the accuracy and/or reliability of sensor readings which, although not obviously in error, nevertheless produced some puzzling results when correlated with other

[3]An appropriate threshold was set individually for each time series.

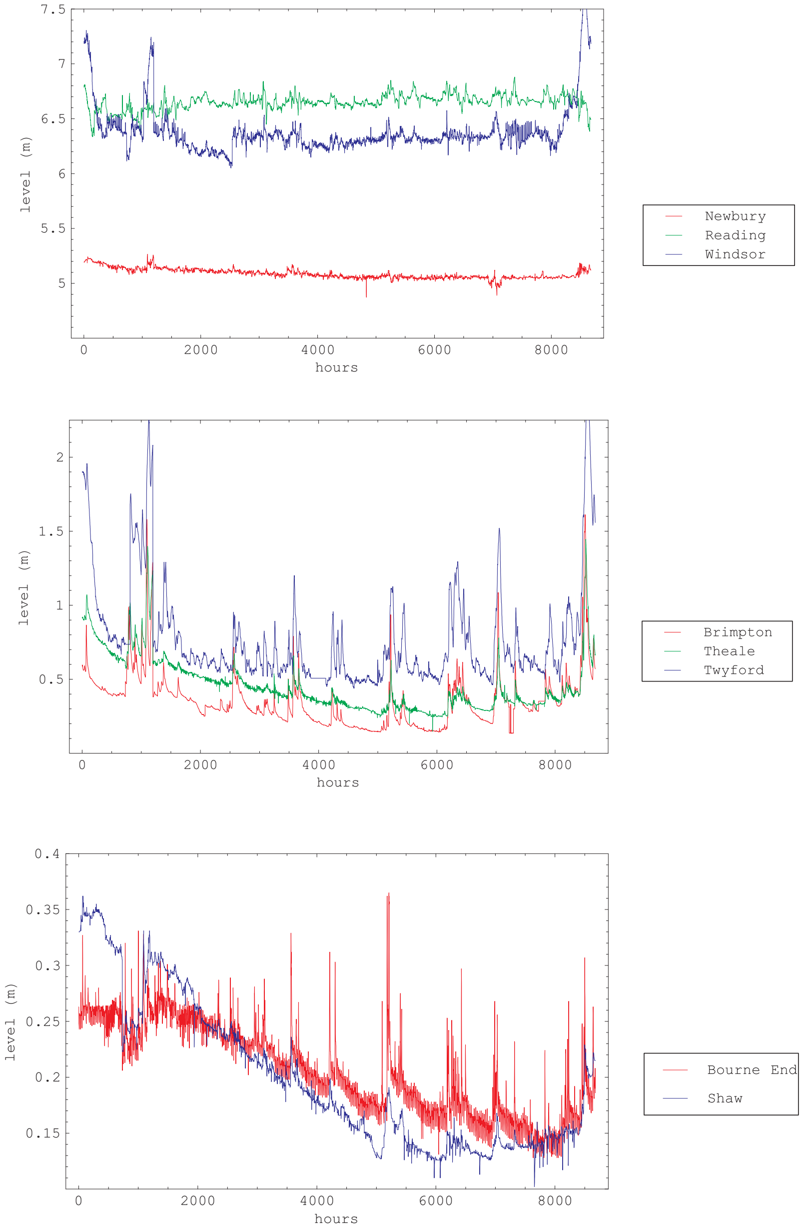


Figure 3: Cleaned river level data at the various sites.

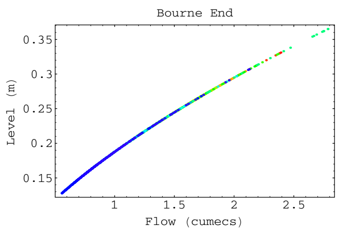
sensor readings taken at the same site. For example, if we correlate flow and level readings taken at the same site at the same time we might expect to see a simple functional relationship, in which increased flow produces a non-linear increase in level. Figure 4 shows Flow-Level correlation plots for the various sensor sites shown in the map of Figure 1.

We can see that in some instances (River Wye at Bourne End, River Enbourne at Brimpton, River Lambourn at Shaw, and to some extent River Kennet at Theale) our expectations are confirmed. However, in other instances (River Kennet at Newbury, River Thames at Reading, River Loddon at Twyford, and River Thames at Windsor) the flow-level correlations show very unpredictable and widely differing scatter plots and are open to interpretation. For example, in Figure 4(d) examining the River Thames at Reading we might suspect progressive sensor drift. Indeed closer inspection of the data (the point colour changes progressively through the spectrum from red to blue over the calendar year) shows that the different coloured 'lines' visible on the scatter plot often occur at intervals throughout the year. An alternative explanation might be that the river was dredged or altered periodically and this fact was reflected in changes to the flow-level relationship. However, other examples, such as the River Kennet at Newbury or the River Thames at Windsor are less easy to interpret.

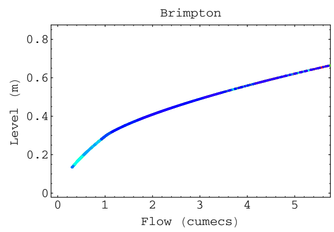
The inspection and cleaning of the raw data in conjunction with the interpretation of the Flow-Level correlation plots indicates that there is a serious issue of data accuracy and consistency to be addressed by the responsible agencies.

3.4 The rainfall data

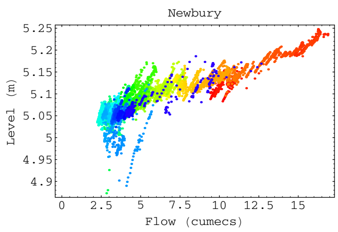
The raw data for the five hourly rainfall sites over the data period of one calendar year is shown in Figure 5 and illustrates that the rainfall measurements were relatively stochastic and discontinuous.



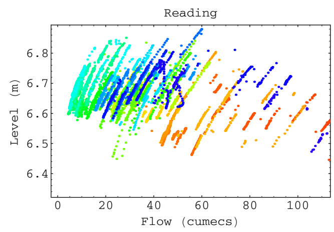
(a) River Wye at Bourne End.



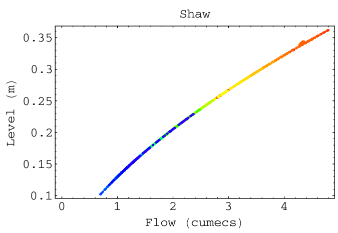
(b) River Enbourne at Brimpton.



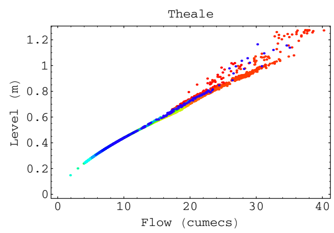
(c) River Kennet at Newbury.



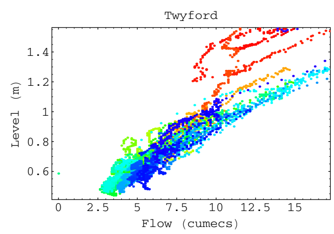
(d) River Thames at Reading.



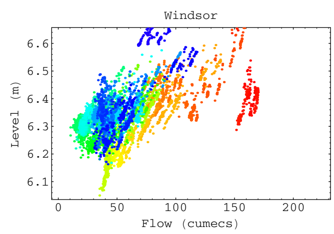
(e) River Lambourn at Shaw.



(f) River Kennet at Theale.



(g) River Loddon at Twyford.



(h) River Thames at Windsor.

Figure 4: The flow-level correlations of the river data measured at each sensor site. The hue of the points indicates the time of measurement (the colours change progressively through the spectrum: red points were measured at the start of the period and blue points at the end).

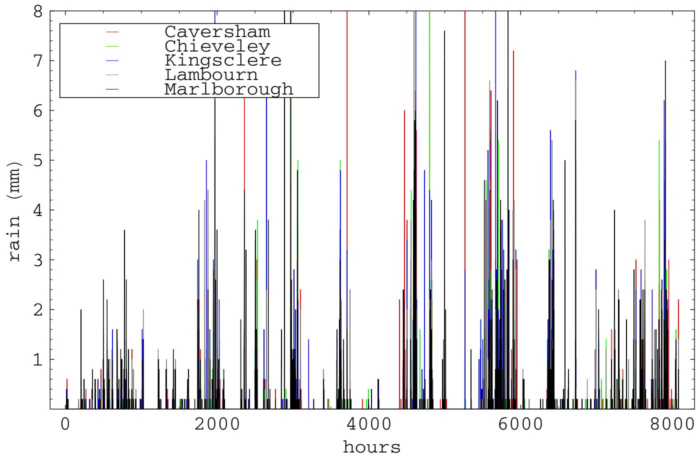


Figure 5: The rainfall monitored at the sites shown in Figure 1.

Prolonged periods of rain will saturate the catchment area and fill underground reservoirs, resulting in a change in the runoff dynamics where more water arrives in the tributaries and main watercourses, and it arrives more rapidly.

To incorporate these long term effects into the model and to investigate their relevance, the rainfall measurements were aggregated over different time intervals: 24-hours, 7-days and 28-days.

4 Data Analysis

Given this data one could build predictive models for both level and flow, but we report here on level models. Examination of the regional map in Figure 1 shows that two models can be sensibly constructed from the data measured at the marked sensor sites:

- Theale area model:

  Theale predicted using river measurements from Newbury, Shaw and Brimpton together with rainfall statistics.

- Windsor area model:

  Windsor predicted using river measurements from Theale, Reading, Twyford, Bourne End together with rainfall statistics.

Here 'rainfall statistics' indicates some combination of lagged and aggregated rainfall measurements.

Since the river Kennet at Theale flows into the higher reaches of the Thames this choice enables us to investigate whether a modular system could be built using the output of one model as an input into another. Thus the Windsor area model allows us to use either the real data measured at Theale, or the predicted river levels from the first model. In this particular case, because of the location of the sensors, both models will give a 4 hour prediction and so there is actually no advantage. In general, such a modular structure would enable longer prediction times: the tradeoff is the utility of longer prior warning against an initially reduced level of accuracy.

4.1 Normalization of data

Since different data types such as flow, level and rainfall were in different units and over significantly different range scales, all data was normalized prior to analysis by mapping the mean to zero and the standard deviation to $\frac{1}{2}$. In general the process of normalization attempts to initially equalize the relative numerical significance between the input variables and to aid the analysis routines, especially in the absence of any prior knowledge regarding input variable relevance.

Normalizing the data will produce a different set of near neighbour relationships compared to those for the un-scaled data. However, any two metrics on a Euclidean space are equivalent to within a constant, so a Gamma test analysis on normalized data will not affect the *asymptotic* nature of the Gamma statistic. Normalization can however affect the *rate* of convergence of the Gamma statistic and the *quantity* of data required to produce a model of given quality.

4.2 Determining the lags

One simple measure of the performance of a model on unseen data for which the measured outputs are known is the mean-squared error (MSE) over the test data. If $\{y_i : i \in U\}$ is a previously unseen set of measured values of an output and $\{\hat{y}_i : i \in U\}$ is a set of predictions for y_i then the mean-squared error of the predictions is given by

$$MSE = \frac{1}{|U|} \sum_{i \in U} (\hat{y}_i - y_i)^2 \tag{6}$$

The MSE is not an ideal measure in all respects. For example, we may have a time series model with a small *MSE*, but which shows no particular propensity to accurately anticipate turning points, a factor of particular interest in the present context.

Our early experiments in river level and flow prediction indicated, as one might expect, that an embedding model constructed from time series data recorded at a *single* site could not be used to produce a model that effectively predicted turning points. Although we were able to obtain models with a relatively low MSE, these models invariably *lagged* the actual data by one time step (in attempts to build predictive models for a random walk we see an identical behavior) and so were ineffective in anticipating future changes of level or flow. This is entirely reasonable, since what affects the behavior at a particular point on the river is predominately determined by the input from upstream, not the previous behavior at the same point. Obviously one should use upstream data, taken prior to the time of prediction, to model flow or level at downstream sites. Only in this way can we be sure of genuinely capturing the flow dynamics.

One way to determine the correct transfer times between successive measurement points is by direct on-site measurement, preferably under a variety of flow rate conditions. This would obviously be the recommended approach in a real system. It is relatively straightforward to accomplish and, once performed, leaves no room for doubt. Additionally such physical measurements act to validate algorithmic approaches to determining lags. We investigated two algorithmic techniques to determine the transfer times directly from the data measurements: the *Time-lag Gamma test*[4] and *Delta correlation*.

The Time-lag Gamma test compares the target time series $y(t)$ with an input time series $x(t)$ $(1 \leq t \leq M)$ by computing Gamma statistics for data sets $(x(t-d), y(t))$ for $d = 1, 2, 3, \ldots$ and then choosing the lag d which produces a Γ closest to zero. Although this worked well on simulated river flow data it proved less effective on real data.

Instead we used the second approach of Delta correlation and then validated our choice of inputs using the Gamma test. The delta correlation $\Delta_c(d)$ of $y(t)$ with $x(t)$ $(1 \leq t \leq M)$ at lag d is defined by

$$\Delta_c(d) = \frac{1}{AB} \sum_{i=1}^{M-d-1} (x(i+1) - x(i))\,(y(i+1+d) - y(i+d)) \tag{7}$$

where

$$A^2 = \textstyle\sum_{i=1}^{M-d-1} (x(i+1) - x(i))^2$$

$$B^2 = \textstyle\sum_{i=1}^{M-d-1} (y(i+1+d) - y(i+d))^2$$

The idea here is to correlate *changes* in the input time series with later *changes* in the output time series at some lagged time d. The time lag with the highest positive correlation[5] should indicate the flow time between sensor points.

[4]This had been called the sliding-ones Gamma test in previous work.

[5]In the context of river flows, the correlations will be positive since the expectation is that as an upstream river rises (or falls) then downstream the river will correspondingly rise (or fall) at a later point in time. In general, for other types of problem, strong negative correlations may be as significant as strong positive ones.

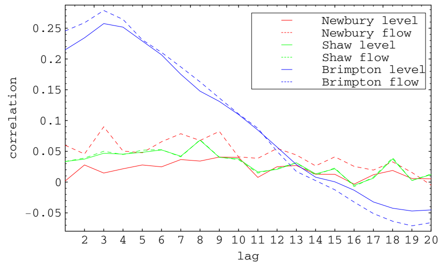
An important aspect of Delta Correlation is that it is very fast - so one can obtain an initial overview of what lags are likely to be important very quickly. Usually identification of lags for flow and level time series was not too difficult: we simply picked the lag time with the largest delta correlation, provided this was consistent with out understanding of the relative distances involved. Rainfall and aggregated rainfall lags were often harder to decide.

4.2.1 Determining the lags for the Theale area model

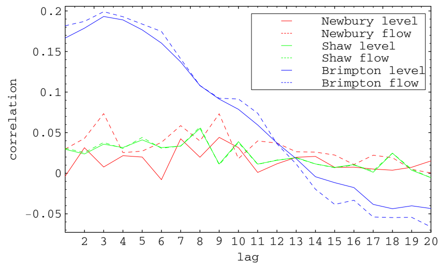
The Delta correlation on the Theale area data produced the graphs in Figure 6. The level and flow measurements at Newbury, Shaw and Brimpton were correlated to the level and flow measurements taken at Theale. After determining the lags by selecting the maximum correlation we arrive at the Delta correlation results shown in Table 1.

The Delta correlation analysis unambiguously identifies the lags from Shaw and Brimpton to Theale to be 8 hours and 3 hours respectively. The lag between Newbury and Theale is less clear cut. The analysis produces a 3 hour lag using the flow data and a 9 hour lag using the level data. The distance between Newbury and Shaw would suggest that the lag to Theale should indeed be around 9 hours. A closer examination of the data used to produce Figure 6 shows that the 3 hour lag had a correlation of 0.0735 and a correlation of 0.0732 for 9 hours. We can conclude that the likely lag is indeed 9 hours given all of the available evidence.

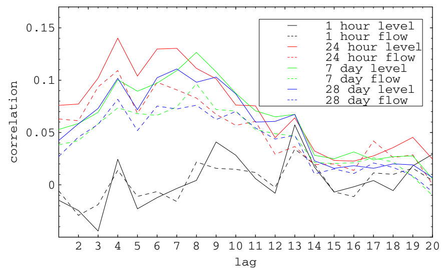
For the regional rainfall aggregated over 28 days we obtain a Theale level correlation of 0.111 corresponding to a lag of 7 hours, whereas for the flow we obtain a correlation of 0.102 corresponding to a lag of 4 hours. In this case the meaning of a lag against a 28 day aggregated rainfall is less clear cut, but examining the graphs we decide that a 7 hour lag may be more appropriate here. The Delta correlations between individual rainfall sensor sites and Theale



(a) The Delta flow and level correlations from Newbury, Shaw and Brimpton measured against the river level at Theale.



(b) The Delta flow and level correlations from Newbury, Shaw and Brimpton measured against the river flow at Theale.



(c) The rainfall measured against the level and flow at Theale.

Figure 6: The Delta correlation plots for the Theale model.

measurement	Delta level	Delta flow	used lag
Newbury level	9	9	9
Newbury flow	3	3	9
Shaw level	8	8	8
Shaw flow	8	8	8
Brimpton level	3	3	3
Brimpton flow	3	3	3
Regional rainfall 1-hour	13	13	13
Regional rainfall 1-day	4	4	4
Regional rainfall 7-days	8	8	8
Regional rainfall 28-days	7	4	7
Marlborough rainfall	13	9	13
Lambourn rainfall	13	9	9
Chieveley rainfall	13	13	13
Kingsclere rainfall	20	20	8

Table 1: Estimated lags for the Theale area measurements. The lags chosen for the analysis were derived from the Delta correlation analysis. The lag for Kingsclere rainfall was manually selected as 8 hours.

were also analyzed, as they could introduce additional local information that the aggregated regional rainfall cannot describe.

It is interesting to note that the results of the Delta correlation analysis are relatively consistent, regardless of whether the level or flow are used.

The lags calculated in Table 1 were used to construct a data set for the Theale area model. The choice of inputs was then validated using the full-embedding routine. This analysis determined that the rainfall at Lambourn and the flow at

	Including Lambourn rainfall and Newbury flow	Excluding Lambourn rainfall and Newbury flow
$\vert\Gamma\vert$	0.00077	2.0638×10^{-6}
Gradient A	0.01865	0.022833
Standard error	0.00066	0.00037164
$\vert V_{ratio}\vert$	0.00306	8.255×10^{-6}
Near neighbours	10	10
M	8076	8076
Mask	`11111111111111`	`11011111101111`

Table 2: The Gamma test analysis results on the scaled Theale level area data set. The two results compare the effect of including or excluding the Lambourn rainfall and the Newbury flow (indicated by a `1` or `0` in the mask respectively).

Newbury were irrelevant ($|\Gamma| = 0.00077$ with Lambourn rainfall and Newbury flow and $|\Gamma| = 2.1 \times 10^{-6}$ excluding Lambourn rainfall and Newbury flow). The results of the analysis are shown in Table 2.

The discovery that the rainfall measured at Lambourn and the flow at Newbury were not useful led us to try a number of manual revisions to the lag times, but at each stage the feature selection routines discarded the measurements. This observation may have arisen for a number of reasons. The Lambourn rainfall measurement site is relatively distant from Theale, making it difficult to find a correlation. Considering that when rain occurs it is actually distributed across a region, it may well be the case that the Lambourn rainfall measurements, insofar as they contribute at all, are contributing at around the noise level. The Newbury flow-level correlation, shown in Figure 6(c), indicates that flow and level are not highly correlated. The analysis has consequently selected the most useful of the Newbury measurements and discarded the less reliable flow information.

The result of the analysis is to use the inputs and lags in the Theale model that correspond to those shown in Table 1 without the rainfall measurements at Lambourn and the flow measurements at Newbury.

4.2.2 Determining the lags for the Windsor area model

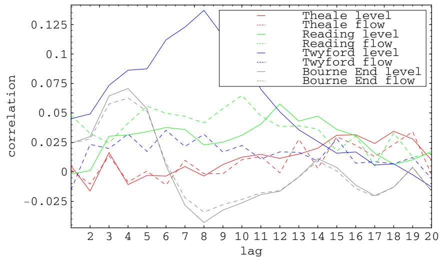
The Delta correlation for the Windsor area data produced the graphs shown in Figure 7. The level and flow measurements at Theale, Reading, Twyford and Bourne End were correlated to the level and flow measurements taken at Windsor. The lags shown in Table 3 were then determined from the Delta correlations.

The Delta correlation lag analysis identified that the lag between Theale and Windsor should be approximately 18 hours. The lag between Reading and Windsor was chosen to be 11 hours. Using Figure 7 we deduced that the remaining lags were likely to be 8 and 4 hours between Twyford and Windsor, and Bourne End and Windsor, respectively.

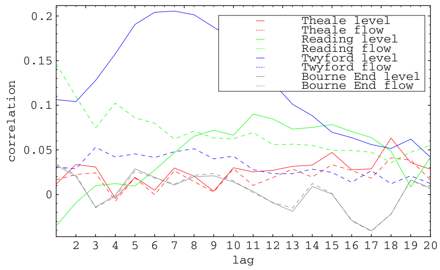
The regional rainfall was much harder to correlate. The evidence seems to suggest a lag of either 1 or 8 hours for hourly and weekly aggregated rainfall. Since we need a lag equal to or longer than the lag for Bourne End we chose 8 hours. A longer lag of 8 hours seems appropriate for the daily, 28 day, and Caversham rainfall.

The results of the Delta correlation analysis for level and flow were not as consistent as those calculated for the Theale area model. This could be due to the low correlation between level and flow measurements at the Reading, Twyford and Windsor sensors (Figure 4).

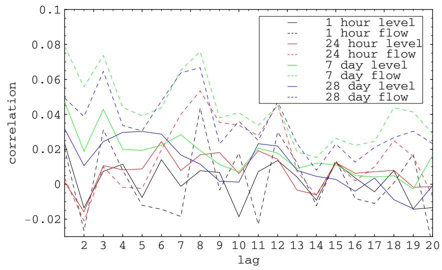
Gamma analysis on the scaled data set in Table 4 suggests that the inputs corresponding to Reading Flow and Caversham rainfall should be discarded.



(a) The Delta flow and level correlations from Theale, Reading, Twyford and Bourne End measured against the river level at Windsor.



(b) The Delta flow and level correlations from Theale, Reading, Twyford and Bourne End measured against the river flow at Windsor.



(c) The rainfall measured against the level and flow at Windsor.

Figure 7: The Delta correlation plots for the Windsor model.

measurement	Delta level	Delta flow	used lag
Theale level	18	18	18
Theale flow	19	18	18
Reading level	12	11	11
Reading flow	10	1	11
Twyford level	8	7	8
Twyford flow	6	3	8
Bourne End level	4	1	4
Bourne End flow	4	1	4
Regional rainfall 1-hour	1	8	8
Regional rainfall 1-day	6	8	8
Regional rainfall 7-days	1	1	8
Regional rainfall 28-days	1	8	8
Caversham rainfall	11	8	8

Table 3: Estimated lags for the Windsor area measurements. The lags chosen from the analysis were derived primarily from the Delta correlation analysis.

However, the model complexity leaving out these inputs (as judged by the gradient A) is approximately twice as large. It emerged that slightly better results could be obtained by retaining these inputs and building the simpler model than by discarding them.

5 Modelling Results

Data files were constructed using the appropriate lags and imported into *winGamma*. These data files were used to build models for the Theale and

	Including Caversham rainfall and Reading flow	Excluding Caversham rainfall and Reading flow
$\|\Gamma\|$	0.00015385	1.254×10^{-5}
Gradient A	0.032345	0.062577
Standard error	0.000211	0.00027477
$\|V_{ratio}\|$	0.0006154	4.5017×10^{-5}
Near neighbours	10	10
M	8071	8071
Mask	`1111111111111`	`1110111101111`

Table 4: The Gamma test analysis results on the scaled Windsor level area data set. The two results compare the effect of including or excluding the Caversham rainfall and Reading flow (indicated by a `1` or `0` in the mask respectively).

Windsor areas.

5.1 Theale area model results

The best predictive inputs for the level at Theale were selected in section 4.2.1. We next use the Gamma test with this input mask to analyze the *quantity* of data, using the M-test to determine whether there was sufficient data to provide an asymptotic Gamma estimate and subsequently a reliable model. The results of this analysis are shown in Figure 8. To average the seasonal effects implicit in the data, an M-test was performed on order-randomized data and the results plotted. As the M-test proceeded, the Gamma test algorithm was exposed to points randomly sampled throughout the year. This produced an asymptotic convergence of the Gamma statistic, $\Gamma = 0.0007$, and indicated that there was sufficient data at around $M = 6500$ data points.

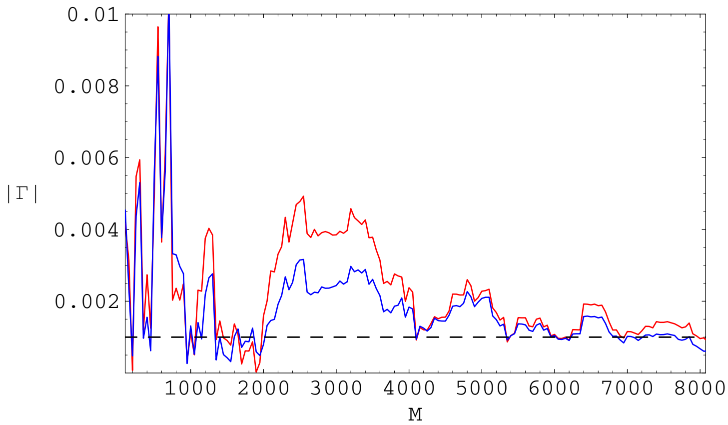


Figure 8: M-test performed on the Theale area model randomized, scaled data. The red lines correspond to the Gamma statistic calculated for the river level at Theale and the blue lines correspond to the flow. The dashed line shows that the data asymptotes at around 6500 points.

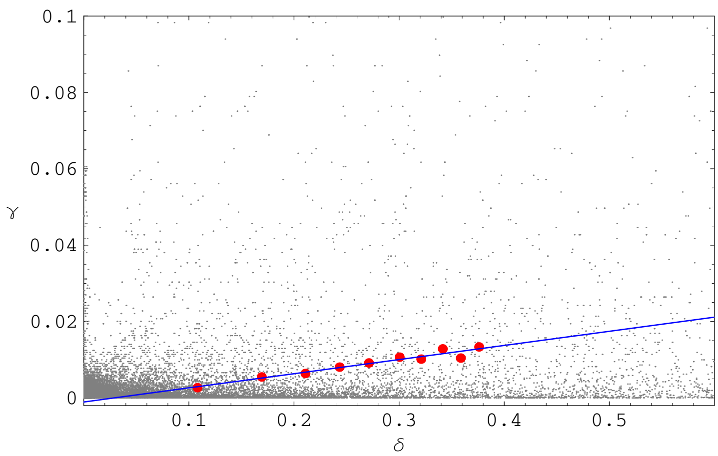


Figure 9: A Gamma scatter plot generated from the data for the Theale area model using level as the output. The scatter plot shows a moderate level of noise for the optimal choice of lags and inputs.

The form of the charts in Figure 8 indicates that there is very little difference between modelling the level or the flow (the measurements are reasonably well correlated at Theale, shown in Figure 4(f)).

An M-test analysis (and common sense) shows that in order to capture the seasonal dynamics of the river system, the river environment must be sampled for at least a full year. Since in this case we only have one year of available data we cannot build a successful model using the data in chronological order, by selecting one continuous time period for model training and a second disjoint period for testing. Instead we chose to order-randomize the data and use a proportion of the data for training and a separate proportion for testing. Using the M-test in Figure 8 we know that at least 6000 randomized data points would be required to build a reliable model.

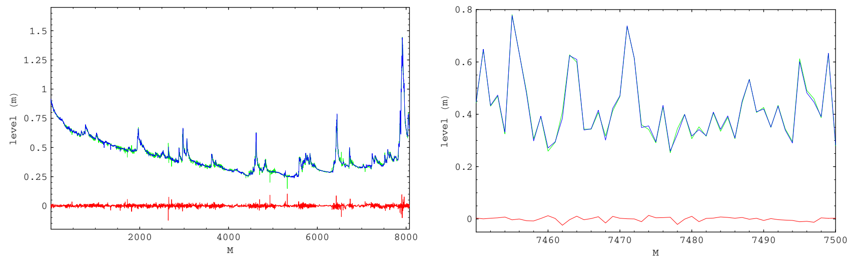
The Gamma scatter plot for this data set shows a moderate level of noise, Figure 9, even though the lags have been optimized and the best combination of inputs selected.

	Selected Theale area inputs
$\|\Gamma\|$	0.0008407
Gradient A	0.025359
Standard error	0.0004489
$\|V_{ratio}\|$	0.0034078
Near neighbours	10
M	6500

Table 5: The Gamma test analysis result on the randomized, scaled Theale level area data set to determine the target MSE.

The estimated model performance was then determined using the Gamma test, the results of which are shown in Table 5. The target MSE for the model was 0.0008407 (scaled) for $M = 6500$.

We constructed a feedforward 12-10-10-1 neural network trained using the BFGS algorithm [Press et al., 1992] which produced the results shown in Figure 10. Since the minimum lag used is the 3 hour lag from Brimpton to Theale, these models give a three hour ahead prediction.



(a) The model response in chronological order (20% of points are unseen).

(b) A closer inspection of the BFGS model performance on the (order-randomized) unseen data shows an acceptable error level.

Figure 10: The performance of the 3 hour look-ahead BFGS Theale area model. The green line shows the actual river level at Theale, the blue line shows the model prediction for the river level, and the red line shows the error between the actual and predicted level.

On the unscaled data the neural network reached a MSE of 8.002×10^{-5} on the training set and 0.000115 on the unseen test set (Table 6). The similarity of these two values would seem to indicate that the model generalized well and had not been overtrained.

Overall the results are rather encouraging. In Table 6 the unscaled MSE of 0.000115 translates to an error standard deviation deviation[6] of 0.0107m, for the 3 hour neural network prediction, i.e. around 1cm.

[6]Calculated by taking the square root of the MSE.

	MSE (scaled)	MSE (unscaled)
Training data	0.000863	8.002×10^{-5}
Test data	0.00124	0.000115

Table 6: The MSE values of the Theale area neural network model. The scaled MSE is measured using the same scaling as used for the Γ value. The unscaled MSE is measured in the same units as the river level.

5.2 Windsor area model results

The best predictive inputs for the level at Theale were selected in section 4.2.2. We next use the Gamma test with this input mask to analyze the *quantity* of data, using the M-test to determine whether there was sufficient data to provide an asymptotic Gamma estimate and subsequently a reliable model.

Figure 11 shows that the Windsor level and flow Gamma statistics on scaled data do not converge to approximately the same value, as did the corresponding values for scaled Theale data, see Figure 8. There is no reason in principle why these values should be the same. For example, differing precision in measurement of flow and level could cause different degrees of noise. In the Theale data, flow and level were highly correlated, whereas in the Windsor data the correlation between level and flow was not as significant.

The results of the M-tests in Figure 11 show that the seasonal effects can indeed be eliminated this way. The M-test performed on the randomized data shows that the Gamma statistic asymptotes on scaled data to $\Gamma \approx 0.00015$, at approximately 6500 points.

The estimated model performance was then determined using the Gamma test, the results of which are shown in Table 7.

A training data set was created from 6500 randomly selected data points using all of the available inputs (see Table 4 for the analysis). The target MSE

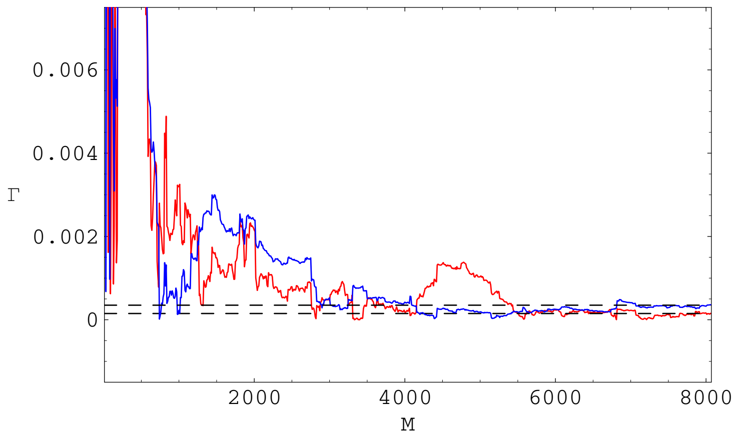


Figure 11: M-test performed on the Windsor area model randomized, scaled data. The red lines correspond to the Gamma statistic calculated for the river level at Windsor and the blue lines correspond to the flow (scaled data) . The dashed lines show the asymptotic Γ values for level and flow (approximately 0.00015 and 0.00035 respectively.)

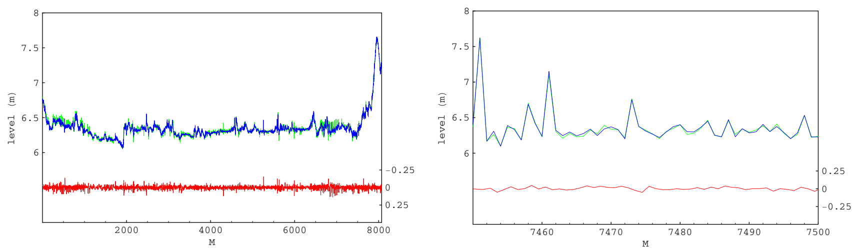
	All Windsor area inputs
$\|\Gamma\|$	0.00029824
Gradient A	0.031805
Standard error	0.0002325
$\|V_{ratio}\|$	0.0012537
Near neighbours	10
M	6500

Table 7: The Gamma test analysis result on the randomized, scaled Windsor level area data set to determine the target MSE.

for the model was 0.00029824 (scaled) for $M = 6500$.

We constructed a 13-20-15-1 feedforward neural network trained using the

BFGS method. The model is shown in Figure 12. Since the minimum lag used is the 4 hour lag from Bourne End to Windsor, these models gave us a four hour ahead prediction.



(a) The model response in chronological order (20% of points are unseen).

(b) A closer inspection of the BFGS model performance on the (order-randomized) unseen data shows an acceptable error level.

Figure 12: The performance of the 4 hour look-ahead BFGS Windsor area model. The green line shows the actual river level at Windsor, the blue line shows the model prediction for the river level, and the red line shows the error between the actual and predicted level.

	MSE (scaled)	MSE (unscaled)
Training data	0.0031426	0.0004565
Test data	0.0049463	0.0007185

Table 8: Results for the Windsor area BFGS model showing the performance on the test and training sets. The scaled MSE is measured using the same scaling as used for the Γ value. The unscaled MSE is measured in the same units as the river level.

The neural network reached a MSE of 0.0004565 on the training set which evaluated to an MSE 0.00071853 on the unseen test test on the unscaled data. These MSE figures are shown in Table 8. The neural network predicts with an error standard deviation of 0.026m, i.e. of the order of 3cm for an average level of around 6.3603m, which gives a percentage error of well under 0.5%.

6 Discussion of Modelling Results

There was evidence in the Windsor area model that perhaps other factors, beyond the rainfall and flow considered during the modelling, were affecting the river levels. In comparing the model predictions to the actual data we discovered there seemed to be occasional periodic (in fact daily) activity that affected river levels. Figure 13, where the vertical grid-lines indicate 24 hour intervals starting at midnight, illustrates an interval displaying this periodic behaviour of the river levels. The phenomenon is a rapid rise and then fall in level. Preceding the rise the model is often over-estimating the level.

Although only occasional, when it occurs the effect is too regular to be explained through coincidence or data measurement errors, but instead is either an environmental effect or has arisen through human intervention[7]. The approximate time for the peaks in river level are daily between 5pm and 10pm, with most being measured between 7pm and 8pm. This could be attributed to human domestic activity, regular engineering or agricultural work, or some other regular activity.

If we could attribute these fluctuations to periodic extractions and replacements, such as those created by industrial or human use of the river water, then we could use our techniques to develop an automatic monitoring program for detecting the (unlicensed) use of river water.

[7] Sometimes termed an ethnocentric signal

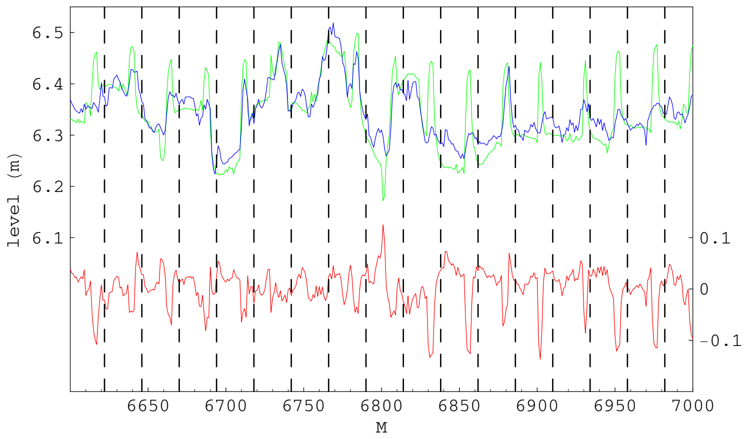


Figure 13: A comparison between the Windsor area model (blue) and the actual observations (green) shows that the model does not predict the daily fluctuations. The periodic fluctuations show up in the error (red), which is plotted on the right-hand scale.

7 Modular System

Finally we constructed a modular prediction system of the River Thames at Windsor using the predicted behaviour of the River Kennet at Theale as an input to the Windsor area model, replacing the measured data at Theale. This approach demonstrates the feasibility of a modular flood prediction system. Table 9 describes the performance of the Windsor area model using the modified data set and should be compared with Table 8.

This work demonstrates a modular system consisting of many sub-regional models can be built. Unfortunately because of the location of sensors in this case-study we were unable to increase the prediction time. However, given an appropriate tributary structure and sensor placement we could enhance the look-ahead prediction time.

The data sets described as hybrid data 1 and 2 in Table 9 are data sets derived from the original training and tests sets, respectively. The original sets

	MSE
Hybrid data 1	0.0006696
Hybrid data 2	0.0008865

Table 9: A comparison of the MSE values of the two Windsor area models. The actual river measurements at Theale have been substituted by predicted values from the Theale area model. In all other respects, the data sets were identical to those used in the Windsor area model described in Section 5 and which produced the results in Table 8.

were used to train and test the Windsor area model in Section 5. The only difference between these hybrid data sets and the original data sets is that the hybrid sets include the *predictions* for Theale rather the original Theale measurements. In addition, hybrid data set 1 was not used for training, because the model had already been constructed during our earlier experiments, but was used to compare the two models.

The results here compare favourably to the results of the original Windsor area model (Table 8). The MSE on the original test set was 0.0007185 awhich becomes 0.0008865 in the hybrid model. This represents a 11% greater error standard deviation in the modular system and translates to an average error of 2.97cm in the prediction, which remains under 0.5% of the average level $6.3603m$ at Windsor.

8 Conclusions

It would appear that there is no barrier in principle to producing very accurate post-precipitation river level and flow predictions apart from the data quality issue.

In addition to providing predictions using observable data, the overall utility of the system could be enhanced by providing facilities to run what-if scenarios to predict what effect certain activities would have on the river. For example, long-term weather forecast could be used to provide an advanced warning of the *possibility* of flooding.[8]

Provided these issues are adequately addressed then it would seem entirely feasible to develop an adaptive modular system for modelling and predicting river flow and levels over a time scale basically determined by the delay between the precipitation occurring and the water arriving at the prediction point.

One significant advantage of such an integrated monitoring, telemetry, and modelling system is that, subject to careful choice of sensor location, the entire system is independent of the actual location or river basin being modelled, i.e. having designed and produced the hardware and software, the entire package could be sold as a complete system capable of being installed at any location and producing effective river data modelling within a period of two to three years following installation.[9]

Of course, there are other issues that become necessary to resolve. One notable failure of conventional level prediction techniques was the Columbia River (Oregon-Washington, USA) flood of February 1996, in which accumulated snow in the mountains melted rapidly when the temperature rose sharply accompanied by high rainfall. Upstream flooding was widespread and only extreme measures prevented the west coast city of Portland from serious flooding. Thus, apart from the obvious input variables considered here, it would be wise to include the current depth of surface snow, ambient temperature, relative humidity etc. and to augment the modelling techniques to accommodate these

[8] Of course, predicting the weather is a complicated undertaking, but if done with a suitable level of accuracy could enable the flood prediction system to provide accurate early warning alerts.

[9] *MAPFLOWS* is such a conceptual software system but has yet to receive funding.

variables. This would require further research which was outside of the scope of this feasibility study.

References

[Aðalbjörn Stefánsson et al., 1997] Aðalbjörn Stefánsson, Končar, N., and Jones, A. J. (1997). A note on the gamma test. *Neural Computing & Applications*, 5(3):131–133. ISSN 0-941-0643.

[Chuzhanova et al., 1998] Chuzhanova, N. A., Jones, A. J., and Margetts, S. (1998). Feature selection for genetic sequence classification. *Bioinformatics*, 14(2):139–143.

[Durrant, 2001] Durrant, P. J. (2001). *winGamma: A non-linear data analysis and modelling tool with applications to flood prediction*. PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.

[Evans, 2001] Evans, D. (2001). *Data Derived Estimates of Noise using Near Neighbour Asymptotics*. PhD thesis, Cardiff University, Wales, UK.

[Evans and Jones, 2002a] Evans, D. and Jones, A. J. (2002a). Asymptotic moments of near neighbour distance distributions. *To appear in Proceedings of the Royal Society Series A.*

[Evans and Jones, 2002b] Evans, D. and Jones, A. J. (2002b). A proof of the gamma test. *To appear in Proceedings of the Royal Society Series A.*

[Georgakakos, 1986a] Georgakakos, K. P. (1986a). A generalized stochastic hydrometeorological model for flood and flash-flood forecasting. 1. Formulation. *Water Resources*, 22(13):2083–2095.

[Georgakakos, 1986b] Georgakakos, K. P. (1986b). A generalized stochastic hydrometeorological model for flood and flash-flood forecasting. 2. Case studies. *Water Resources*, 22(13):2096–2106.

[Guedes de Oliveira, 1999] Guedes de Oliveira, A. (1999). *Synchronisation of chaos and applications to secure communications*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.

[Jones et al., 2002] Jones, A. J., Tsui, A., and Oliveira, A. G. (2002). Neural models of arbitrary chaotic systems: construction and the role of time delayed feedback in control and synchronization. *Complexity International Vol 09. With html electronic supplement. Paper ID: tsui01, URL: http://www.csu.edu.au/ci/vol09/tsui01/*.

[Kitadinis and Bras, 1980a] Kitadinis, P. K. and Bras, R. L. (1980a). Real-time forecasting with a conceptual hydrological model 1 - analysis of uncertainty. *Water Resources*, 16(6):1025–1033.

[Kitadinis and Bras, 1980b] Kitadinis, P. K. and Bras, R. L. (1980b). Real-time forecasting with a conceptual hydrological model 2 - applications and results. *Water Resources*, 16(6):1034–1044.

[Končar, 1997] Končar, N. (1997). *Optimisation methodologies for direct inverse neurocontrol*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.

[Mutreja et al., 1987] Mutreja, K. N., Yin, A., and Martino, I. (1987). Flood forecasting model for Citandy river. In Singh, V. P., editor, *Flood Hydrology*, pages 211–220. Reidel, Dordrecht, The Netherlands.

[Press et al., 1992] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, second edition. ISBN 0-521-43108-5.

[Thirumalaiah and Deo, 1988] Thirumalaiah, K. and Deo, M. C. (1988). River stage forecasting using artificial neural networks. *Journal of Hydrologic Engineering*, 3(1):26–31. ISSN 1084-0699.

[Tsui, 1999] Tsui, A. (1999). *Smooth Data Modelling and Stimulus-Response via Stabilisation of Neural Chaos*. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London.

[Tsui et al., 2002] Tsui, A. P. M., Jones, A. J., and de Oliveira, A. G. (2002). The construction of smooth models using irregular embeddings determined by a gamma test analysis. *Neural Computing & Applications*, 10(4):318–329.

[Wright and Dastorani, 2001] Wright, N. G. and Dastorani, M. T. (2001). Effects of river basin classification on artificial neural networks based ungauged catchment flood prediction. Proceedings of the International Symposium on Environmental Hydraulics, Pheonix.