Generation of Spatially Consistent Rainfall Data — Refinement and testing of simplified models.

Report to DEFRA, by:

Department of Civil and Environmental Engineering
Imperial College of Science, Technology and Medicine
London SW7 2BU
and
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT

Authors:

Professor H.S. Wheater **Professor V.S. Isham**
Dr. C. Onof **Dr. R.E. Chandler**
Mr R. Bird

Consultant:
Dr. D. Koutsoyiannis (Faculty of Civil Engineering, National Technical University of Athens)

January 2002

Contents

1 Introduction **4**

1.1 Background to the study 4

1.1.1 Spatial-temporal rainfall modelling in continuous space and time . . . 5

1.1.2 Alternatives for application in the short-term 5

1.2 Scope of the continuation study and structure of the report 7

2 Data used in the study **9**

2.1 The Brue 9

2.2 The Blackwater 9

2.3 North-East Lancashire 11

3 Multi-site disaggregation **12**

3.1 Problem formulation 12

3.2 Modelling approach 13

3.3 Estimation of the stochastic structure at the hourly level 14

3.4 Models involved 15

3.5 The simplified multivariate rainfall model 16

3.6 The transformation model 17

4 Testing of multi-site disaggregation procedure **20**

4.1 Brue catchment — performance with known hourly cross-correlations 20

4.2 Brue catchment — performance using daily–hourly correlation relationships . 21

4.3 Results for other regions 26

4.3.1 The Blackwater 26

4.3.2 North-East Lancashire 32

5 Generalized Linear Modelling of daily rainfall **35**
5.1 Overview of methodology 36
5.1.1 Multisite simulation 37
5.2 Theoretical developments 37
5.2.1 Notation 38
5.2.2 Choice of distribution for Z_t 38
5.2.3 Fitting 39
5.2.4 Interpretation of the model 39
5.2.5 The joint distribution of $\boldsymbol{Y}$ and Z 41
5.2.6 Practical implementation 44
5.2.7 Dealing with incompatibility 44
5.2.8 Imputation 45
5.3 Performance evaluation 46
5.3.1 The Blackwater 47
5.3.2 North-East Lancashire 51
5.4 Summary of chapter 54

6 Summary and conclusions **56**
6.1 Summary 56
6.2 Conclusions 57

Bibliography **58**

A Figures for GLM results in Section 5.3 **60**

APPENDIX **60**

B Results files for GLMs reported in Section 6.3 **91**
B.1 Blackwater Fitting Results 91
B.1.1 Blackwater logistic model fitting file 91
B.1.2 Blackwater gamma model fitting file 97
B.2 North East Lancashire Fitting results 103
B.2.1 North East Lancashire logistic model fitting file 103

B.2.2 North East Lancashire gamma model fitting file 108

Chapter 1

Introduction

1.1 Background to the study

The research described in this report is a continuation of a programme started in 1998, under the sponsorship of MAFF. It was recognised that the current guidance for the representation of rainfall for flood design was based on highly simplified concepts. Individual storm events were defined by an averaged storm temporal profile; an areal reduction factor was recommended to define areal rainfall from point rainfall estimates. However, the development of continuous rainfall-runoff simulation methods for flood design requires continuous rainfall inputs, hence new methods were required. In addition, new sources of rainfall data were becoming available (radar). The aim of the research programme is therefore to provide a new generation of rainfall simulation tools appropriate for hydrological application in the context of flood design.

Phase 1 of the research culminated in the production of a report *Generation of spatially consistent rainfall data* in February 2000 (Wheater *et al.* 2000). One of the main achievements of Phase 1 was the development of rainfall models that can be calibrated using radar rainfall data. Such data represent an important source of information on the fine-scale spatial-temporal structure of rainfall fields, and the resulting models are suitable for the simulation of rainfall sequences at any temporal and spatial scales of hydrological interest. However, given the current limitations of radar data with respect to accuracy and record length, a preliminary investigation of alternative methods of simulating rainfall sequences, suitable for application in the short-medium term, was also made. These include: a) a simulation model for daily rainfall, based on the family of Generalized Linear Models (GLMs), which can represent non-stationarity in rainfall in both space (allowing for topographic and other location effects) and time (allowing for the representation of climate variability), and b) a methodology for spatial-temporal disaggregation of observed or simulated daily rainfall. These methods are designed for use when radar data are scarce, unreliable or absent and can be used, for example, with daily raingauge network data, and limited availability of sub-daily data. The purpose of the continuation programme described in the current report is further

to investigate such methods.

In the rest of this Section we will give a brief overview of the research carried out under Phase 1. For more detail, see Wheater *et al.* (2000). In subsequent chapters of this report we describe the investigations and developments carried out during the continuation study. Given the dependence of this work on Phase 1, further description of these developments and of the structure of the report will be postponed to Section 1.2, following this overview.

1.1.1 Spatial-temporal rainfall modelling in continuous space and time

Much of the Phase 1 research was concerned with the development and validation of a family of models to represent rainfall in continuous space and time. This is the most general representation of rainfall fields, and model results can be aggregated as required to provide output at any specified space and time steps. This therefore gives complete flexibility of application, and the models, which are parsimonious and efficient, are computationally cheap to simulate. Full details of the final model developed, the Gaussian Displacements Spatial-Temporal Model (GDSTM), are given in Wheater *et al.* (2000).

The model was applied using radar rainfall data, with very promising results. It is able to reproduce well the detailed spatial patterns of rainfall and the main statistical properties with respect to both space and time. However, to represent the full levels of clustering in the model, the detailed observation of spatial structure available from rainfall radar data is needed. With a dense network of raingauges, one level of clustering is lost, and the velocity structure of the rainfall fields cannot be readily identified. Although radar data are routinely collected for the UK, historical record lengths are limited, and are restricted by problems of reliability. This situation is improving, but in the short term, at least, remains a constraint. A second issue is that, except for seasonally-varying parameters, the model, at least in its current form, is stationary in both space and time. It may therefore be inappropriate, without further development, for use in regions with strong topographic variability, or for the generation of long rainfall sequences when long term changes in climate are suspected. Hence there is a need for alternative approaches, which are consistent with data currently available, to support short-to-medium term application.

1.1.2 Alternatives for application in the short-term

Generalized Linear Models

Although radar and subdaily raingauge data are relatively scarce in the UK at present, long daily rainfall records are more abundant. These can be used to study nonstationarities in rainfall sequences, and to develop daily rainfall simulation models. In principle, the output from such models can be downscaled to any desired resolution using some appropriate method. Hourly data may be adequate for most flow simulation purposes.

Many techniques are available for generating daily rainfall sequences. Here, one of the primary concerns is the incorporation of nonstationarities. A powerful and flexible technique, which allows nonstationarities to be quantified and incorporated, is that of Generalized Linear Models (GLMs). These are standard in the statistical literature, and were introduced into the study of daily rainfall by Coe and Stern (1982). The basic idea, which is an extension of linear regression, is to use the values of various predictors to forecast a probability distribution for the amount of rainfall at a site on a given day. In fact, this distribution is specified in two parts: we model the probability of rainfall occurrence separately from the amount of rain if non-zero. Previous days' rainfall amounts can be used as predictors, to account for autocorrelation. Other predictors might include quantities representing regional variation (such as site altitude), seasonal variability, long-term trends and 'external' climatological factors such as the North Atlantic Oscillation (NAO). Models can be specified in such a way that the effect of a predictor depends on the values of others — for example, it is known that the NAO affects European climate predominantly in the winter months. Models can be fitted and compared, using Maximum Likelihood, and can be tested using a variety of simple but informative checks. By defining suitable dependence structures between sites, it is possible to build a multivariate GLM that allows simulation of nonstationary daily rainfall sequences over a network of sites.

The GLM methodology has been applied to many rainfall datasets from the UK and elsewhere. It has been found to be extremely useful for interpreting historical records, particularly when we want to study changes in the climate of an area. Moreover, providing models have been specified carefully, GLM simulations can reproduce a variety of features (including extremal behaviour) of observed rainfall sequences. Chapter 4 of Wheater *et al.* (2000) contains technical details, and a selection of results. Other applications include: a flooding study in Ireland, where GLM outputs were downscaled to hourly resolution for input to continuous flow simulations (OPW 1998); and an investigation into drought in the Yangtze River in China (Yang 2001).

Spatial-temporal rainfall disaggregation

The GDSTM can provide stationary sequences in continuous space and time, whereas the GLM can provide nonstationary sequences at a daily timescale. It is natural to try to obtain the advantages of both techniques. Various methods of doing this have been explored. Perhaps the most promising to date uses a method due to Koutsoyiannis (2001); this is described in Section 5.3 of Wheater *et al.* (2000). The situation we envisage is one in which there are no radar data covering the locations of interest and where only daily raingauge data are available. In many cases, raingauge data at a finer time scale (e.g. hourly) will be available at one or more nearby locations and the idea is to combine these fine scale rainfall data with the daily data to generate fine time-scale rainfall series at the locations of interest. This needs to be done in a way that preserves the daily rainfall totals at these locations, respects the important (hydrologically relevant) temporal properties of sub-daily rainfall, and is spatially consistent so that, in particular, the spatial correlation structure of

the rainfall field is maintained. In the simplest scenario, both daily and hourly raingauge data will be observational, but the procedure can be extended to use simulated data.

The basic idea is to start from a simple multivariate autoregressive model for the 'within-day' structure of a rainfall field at a set of locations. This model is able to reproduce what are regarded as the essential statistics of the hourly rainfall process. These are the mean, variance, coefficient of skewness and lag-one temporal autocorrelation coefficient for the series at each location, together with the lag-zero cross correlation coefficients between locations. This multivariate model is fitted using the available hourly series and then simulated, to generate hourly time series at the set of locations of interest. However, to do this we need to estimate the hourly lag-zero cross-correlations appropriate to these locations, since cross-correlations between sites are distance-dependent. A separate model is therefore developed to estimate these. Essentially, a scaling relation for cross-correlations is fitted using the available hourly data (we assume data for at least two hourly gauges are available). This relates sub-daily to daily cross-correlations and therefore enables the extrapolation of the cross-correlation structure of the daily series to an hourly time-scale. In the absence of any direct information on hourly cross-correlations (if only one hourly series is available), this has either to be assumed known, or to be estimated separately. One possibility for the latter is to fit the GDSTM to the hourly and daily data, see Wheater *et al.* (2000).

At this point, we can generate hourly series at the locations of interest with the desired temporal properties and the right spatial structure. The simulated model is, however, spatially and temporally stationary and the series do not as yet have the correct daily total rainfalls. The final step in our disaggregation procedure is therefore to apply a multivariate transformation to the simulated series to give the right daily totals. The transformation employed here does not affect the stochastic properties of the series, which are preserved. However, spatial and temporal nonstationarities of the rainfall field will be reflected in the daily totals at the locations of interest, and thus also in the final transformed series.

1.2 Scope of the continuation study and structure of the report

The first phase of research clearly demonstrated the viability of the GDSTM as a technique for simulation of rainfall in continuous space and time, but as noted above, with the requirement for radar data to support model parameterisation. It was therefore seen as the preferred option for medium term application. In the meantime, other methods (described in Section 1.1.2) using downscaled daily sequences may be more feasible. However, within the scope of the first phase of research, only very preliminary development and testing of the disaggregation methods was possible. The research described herein is the result of a short (9 man-months) extension that focuses on the models that form the basis of the disaggregation procedure. The principal aims were to develop further the disaggregation methods, to test them for different areas of the UK, and to address aspects of the spatial performance of the GLM which were identified as requiring improvement.

The structure of the rest of this report is as follows. In Chapter 2, we describe the observational data from three catchments in the UK (the Blackwater, NE Lancashire, the Brue) that are used to fit and validate the models. In Chapter 3, the full multi-site disaggregation procedure is described, while its implementation and the results of extensive testing of its performance are discussed in Chapter 4. The performance of the models when the cross-correlation structure is known, and when it has to be estimated, are both assessed. In Chapter 5, multivariate GLMs are developed that can be simulated to provide artificial daily rainfall data at a set of locations of interest. A particular challenge addressed here is the appropriate representation of the spatial dependence between locations. The use of GLMs to provide simulated multivariate daily data for input to the disaggregation procedure (to replace the observed daily series), provides the capability to generate hourly time series at locations where no daily observational data are available and, if desired, under a range of scenarios that could, for example, include climate change possibilities. Finally, in Chapter 6, we give a summary of the results of this continuation study and present our conclusions.

Chapter 2

Data used in the study

The procedures developed in this report have been tested using data from three contrasting study regions within the UK. We here give a brief overview of the data available from each of these regions.

2.1 The Brue

The Brue region is located in the catchment of the river Brue in southern England and is covered by the HYREX network of 49 0.2mm tipping bucket gauges. Data available are from September 1993 until August 1998. Gauge locations are shown in figure 2.1. The data have been quality controlled and some periods of data have been removed from each gauge's record. Two periods of hourly data with no missing values for 8 gauges for roughly 2 years have been derived and named set A and set B. Set A comprises data from gauges 2, 4, 9, 22, 23, 25, 29 and 33 and is for the period 8^{th} September 1994 to 20^{th} August 1996. Set B contains data from gauges 9, 12, 14, 15, 19, 25, 28 and 31 and is for the period 6^{th} October 1995 to 5^{th} October 1997. These data sets are used to test disagregation methods in Chapter 4.

2.2 The Blackwater

The Blackwater region is an area roughly $50 \times 40km^2$ around the catchment of the river Blackwater in Surrey. Daily data used in the Blackwater region are from a network of Meteorological Office rainfall gauges. There are 34 gauges in total and records run from 1908 until the present. There are 22 gauges whose records go back as far as 1975. In addition, in this study, 3 Environment Agency hourly records, derived from 0.2mm tipping bucket records, have been used for disaggregation. Locations of gauges are shown in Figure 2.2.

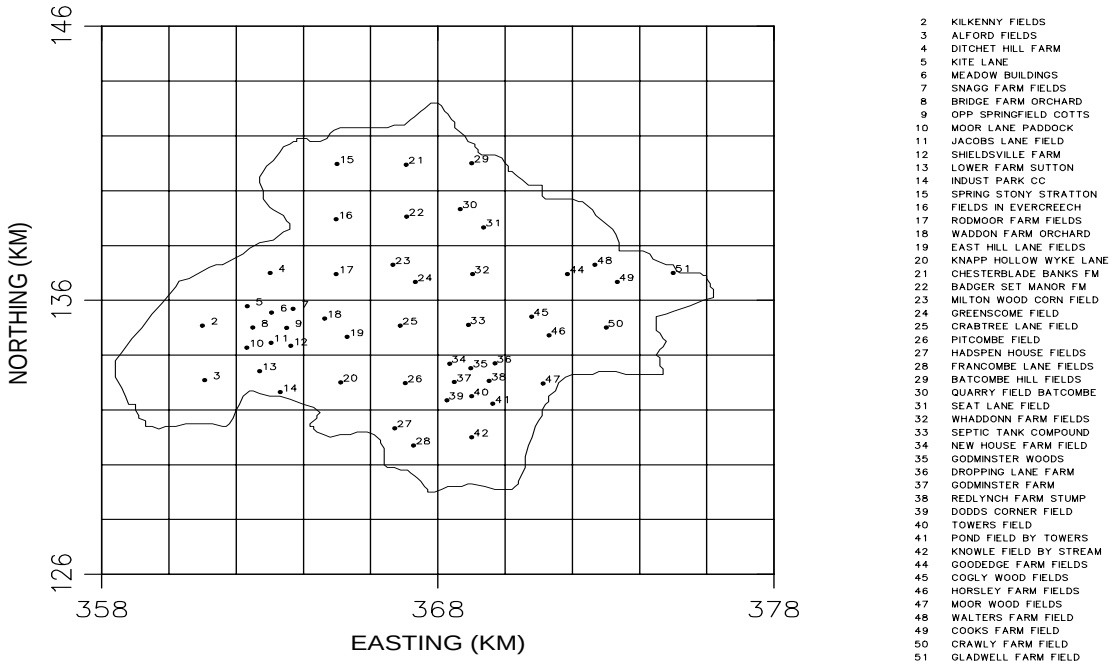

Figure 2.1: Locations of gauges in the Brue region.

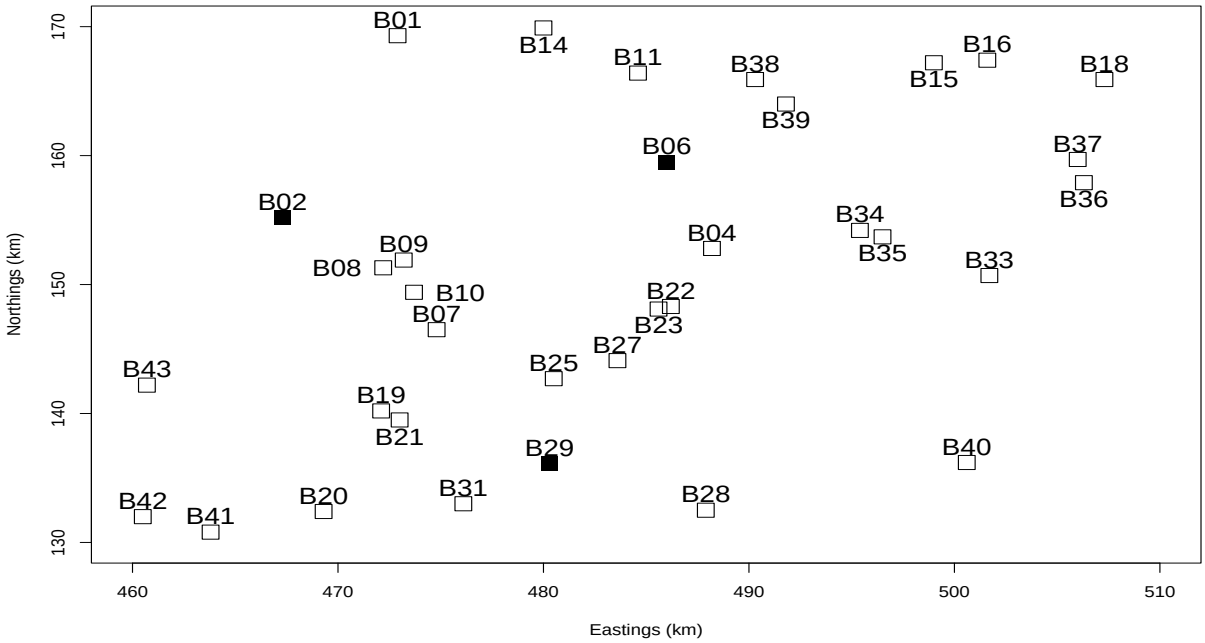

Figure 2.2: Location of gauges in the Blackwater region. Solid symbols represent gauges with hourly data.

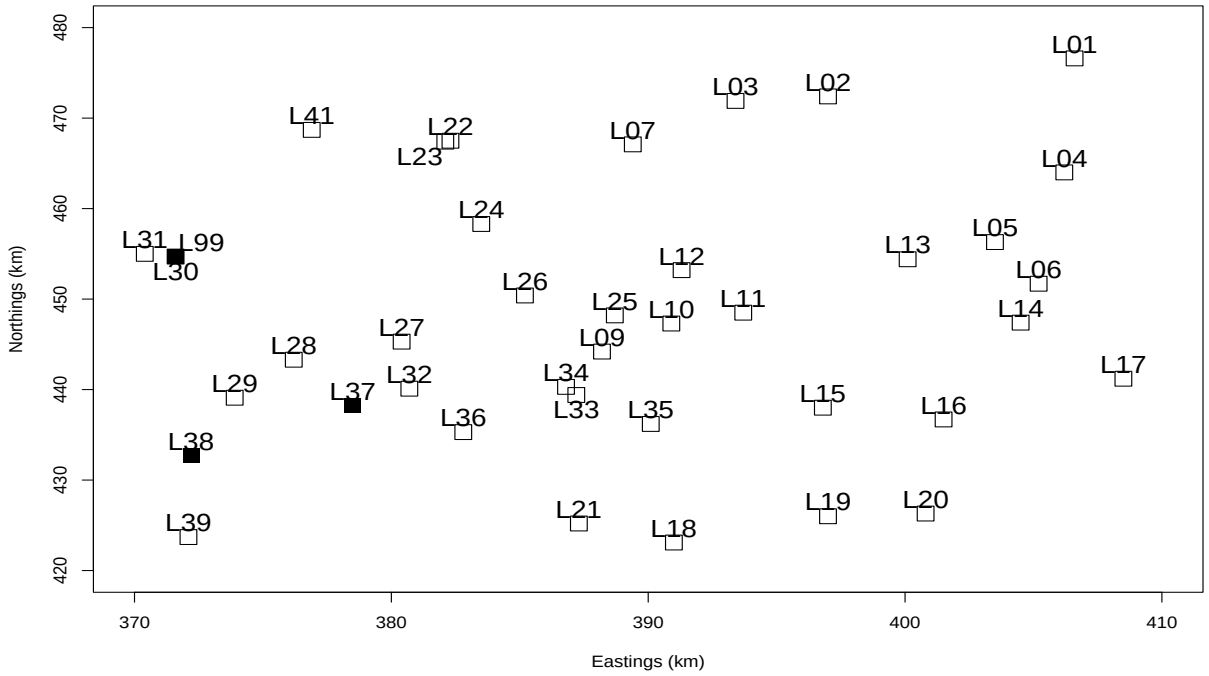

Figure 2.3: Location of gauges in the North-East Lancashire region. Solid symbols represent gauges with hourly data.

2.3 North-East Lancashire

The North-East Lancashire region covers about $40 \times 60 km^2$ at generally high elevation in North East Lancashire and Yorkshire. Daily data used runs from 1961 until the present and are from Meteorological Office daily gauges. There are 40 gauges in total and 23 of these have records going back to 1975. In addition, 3 Environment Agency hourly records, derived from 0.2mm tipping bucket records, have been used for disaggregation. Locations of gauges are shown in figure 2.3.

Chapter 3

Development of multi-site disaggregation procedure

Very frequently, situations with poor data availability arise. For example, radar data may not exist and raingauge data may be available only at a daily time scale at most locations of interest. However, there often exist raingauge data at a finer time scale (e.g. hourly) at a neighbouring site. The question then arises of whether we could utilise the available single-site fine scale rainfall information, in conjunction with the daily data, to generate spatially consistent rainfall series.

This can be considered as a particular case of a general multivariate spatial-temporal rainfall disaggregation problem. In another (commonly occurring) guise, this problem involves the use of observed fine-scale data from a single site to disaggregate historical multivariate daily series. Although there is substantial experience in multi-site disaggregation of rainfall from annual to monthly time scale (e.g. Wilby and Wigley (1997)), and in single-site disaggregation of rainfall to finer time scales (e.g. Wheater *et al.* (2000)), this multivariate fine-time-scale rainfall disaggregation problem has not been studied so far in the rainfall modelling literature. It presents significant differences from that of single-site disaggregation. In particular, the spatial correlation (cross-correlation between different sites) must be maintained. However, the spatial correlation can be turned to advantage since, in combination with the available single-site rainfall information, it enables realistic fine-scale rainfall series to be generated. Timings of rainfall events and maximum intensities at all sites can be guided by an observed or simulated fine-scale hyetograph.

3.1 Problem formulation

We standardise the problem that we examine throughout this section in the following manner. We assume that we are given:

1. hourly rainfall series at one or two gauges at least (hereafter referred to as hourly sites).

If one series only is available, it may be the result of either:

- measurement by an autographic device or digital sensor;
- simulation with a fine time scale point rainfall model such as a point process model (e.g. a model of the Bartlett-Lewis rectangular Pulse type — see Onof *et al.* (2000) for example);
- simulation with a temporal point rainfall disaggregation model applied to a series of daily rainfalls (Wheater *et al.* (2000)).

If two series are available, they could be measurements or simulations from a spatial-temporal model such as the GDSTM (see section 1.1).

2. several daily rainfall series at neighbouring sites (hereafter referred to as daily sites), as a result of either:

   - measurement by conventional rain gauges (daily observations), or
   - simulation with a multivariate daily rainfall model (such as the GLM - see chapter 5).

We wish to produce series of hourly rainfall at these sites so that:

1. their daily totals equal the given daily values;
2. their stochastic structure resembles that implied by the available historical data.

3.2 Modelling approach

The proposed approach to multivariate fine-scale rainfall disaggregation problem involves the application of two separate models at the generation phase.

The first is a rather simplified multivariate model of hourly rainfall that can preserve the essential statistics of the multivariate rainfall process and, simultaneously, incorporate the available hourly information without any reference to the known daily totals at the other sites. The essential statistics considered here are the means, variances and coefficients of skewness, the lag one autocorrelation coefficients and the lag zero cross-correlation coefficients.

The second model is a transformation model that modifies the series generated by the first model, so that the daily totals are equal to the given ones. This uses a (multivariate) transformation, which does not affect the stochastic properties of the series. Both models are discussed further below.

Some questions may arise as to the adoption of a simplified multivariate model rather than a more sophisticated one which could preserve additional statistics of the rainfall process. While in the proposed modelling framework a simplified model is suggested, the adoption of a

more sophisticated approach is not excluded. For example, it might be possible to design the multivariate model so as to maintain a large number of autocovariance coefficients (for any lags). However, such complexity will generally be unnecessary since, for typical catchment sizes, spatial dependence between sites will be strong so that the primary control over the disaggregated output will come from the given fine-scale series.

Specifically, assuming that the daily sites are all close to the hourly ones and highly spatially correlated, the given hourly series at the latter sites can be used, with the simplified multivariate model, to:

- guide the generation of the hourly series at the sites with daily data, and act indirectly to preserve properties not modelled explicitly;
- properly locate the rainfall events in time;
- produce initial hourly rainfall series at the daily sites, whose departures from the actual hourly depths at those sites are not large (even though the known daily totals are not considered at all at this stage).

At a later stage, i.e. when the transformation model is applied, another source of information is additionally incorporated, that is the multi-site daily information. This results in preservation of additional properties, which are not captured by the statistics used. For example, as noted above, nonstationarities of the rainfall field (both in space and time) are reproduceable, even though the models used are both stationary.

The proportion of dry intervals, although considered as one of important properties to be preserved, is difficult to incorporate explicitly in either of the above described models, as it cannot be expressed in terms of statistical moments. However, it can be treated by an iterative procedure (Wheater *et al.* 2000).

3.3 Estimation of the stochastic structure at the hourly level

The essential statistics that we wish to preserve in the generated hourly series are:

1. the means, variances and coefficients of skewness;
2. the temporal correlation structure (autocorrelations);
3. the spatial correlation structure (lag zero cross-correlations);
4. the proportions of dry intervals.

All these statistics, apart from the cross-correlation coefficients, can be estimated at the hourly time scale using an hourly data set. At this parameter estimation stage, these are assumed to be spatially stationary. The generated hourly series, which is forced to respect the observed daily totals, will however reflect the non-stationarities of these data.

The most difficult statistics to estimate are the cross-correlations at the hourly level. The problem, as formulated here, assumes that at least two hourly series are available. On the basis of these and of the daily data at all the gauges, a scaling relationship which holds for all intergauge distances can be fitted to the cross- correlations. This scaling model relates the hourly to the daily cross-correlations.

Note it is possible to carry out the disaggregation with only one hourly gauge. There are two possible cases:

- either the hourly cross-correlation structure is known (see section 4.1);
- or if not, a spatial-temporal model can be fitted to the available hourly and daily data to estimate it, as explained in chapter 5 of Wheater *et al.* (2000); this is the most likely case.

3.4 Models involved

Several separate models are involved in the proposed disaggregation framework. These fall into three categories, as follows:

Category 1: The first category includes the models that are the core of this framework in the sense that they provide the required output (the hourly series). These are, as outlined in section 3.2 above,

- the simplified multivariate model for hourly rainfall, and
- the transformation model.

Category 2: This category contains models to provide the required input, if no observed series are available. These may include

- the GLM for providing daily rainfall depths;
- a single-site disaggregation model to disaggregate daily depths of one location into hourly depths (see Wheater *et al.* (2000) for a review of such procedures);
- a single-site Bartlett-Lewis model (Onof *et al.* 2000) for providing hourly depths at one location.

The single-site Bartlett-Lewis model and the GLM may be used in conjunction. Thus, under a stationary climate scenario, a single-site Bartlett-Lewis model can be simulated into the future, and a GLM can be simulated at neighbouring sites, conditional upon

the daily totals from the single-site Bartlett-Lewis model, so as to introduce spatial nonstationarity into the generated multivariate daily time series.

Category 3: The third category consists of a scaling model describing the relationship between hourly and daily cross-correlations in the standard case. This is described in section 4.2. However, in the case where only one hourly gauge is available, it rather includes models that are able to provide some of the required parameters of the spatial-temporal rainfall process given the statistical properties that can be estimated from the available data (see for example the GDSTM in Wheater *et al.* (2000)).

We now describe the simplified multivariate model and the transformation model which have been used in this work.

3.5 The simplified multivariate rainfall model

Let the n-vector $\boldsymbol{X}_s = (X_s^1, X_s^2, ..., X_s^n)^T$ represent the hourly rainfalls at time (hour) s at n locations. We assume that the simplified multivariate rainfall model is an AR(1) (autoregressive-moving process of order 1) model, expressed by

$$\boldsymbol{X}_s = \boldsymbol{a}\boldsymbol{X}_{s-1} + \boldsymbol{b}\boldsymbol{V}_s \tag{3.1}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are $(n \times n)$ matrices of parameters and $(\boldsymbol{V}_s)$ $(s = 0, 1, 2, \ldots)$ is an independent, identically distributed (i.i.d.) sequence of innovations (these are n-vectors of i.i.d. random variables, so that the innovations are both spatially and temporally independent). The time index s can take any integer value. The $(\boldsymbol{X}_s)$ are not necessarily standardised to have zero mean and unit standard deviation, nor are they normally distributed. On the contrary, their distributions are quite skewed. Consequently, the distributions of $(\boldsymbol{V}_s)$ are skewed too; a three-parameter gamma distribution is generally appropriate for the latter.

Alternatively, the model can be expressed in terms of some nonlinear transformation X_s^* of the hourly depths X_s, in which case (3.1) is replaced by

$$\boldsymbol{X}_s^* = \boldsymbol{a}\boldsymbol{X}_{s-1}^* + \boldsymbol{b}\boldsymbol{V}_s \tag{3.2}$$

It is natural to consider the power family of transformations here.

Equations to estimate model parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ and the moments of $\boldsymbol{V}_s$ are given by Koutsoyiannis (1999) for the most general case. As there is no need to preserve lagged cross-covariances in the problem examined, the parameter matrix $\boldsymbol{a}$ can be diagonal. The parameter matrix $\boldsymbol{b}$ should be defined here as lower triangular. This is necessary in order to incorporate the known hourly rainfalls at the hourly gauges.

3.6 The transformation model

Transformations that can modify a series generated by any stochastic process so as to satisfy some additive property (i.e. that the sum of the values of a number of consecutive variables be equal to a given amount), without affecting the first and second order properties of the process, have been studied by Koutsoyiannis (1994) and Koutsoyiannis and Manetas (1996). These transformations, more commonly known as adjusting procedures, are appropriate for univariate problems, although they can be applied to multivariate problems as well, but in an iterative framework. More recently, Koutsoyiannis (2001) has studied a true multivariate transformation of this type, which avoids any iteration, and also proposed a generalised framework for coupling stochastic models of different time scales.

This framework, tailoredto the problem examined here, is depicted in figure 3.1. Here $\boldsymbol{X}_s$ and $\boldsymbol{Z}_p$ represent the *actual* hourly- and daily-level processes, related by

$$\sum_{s=(p-1)k+1}^{pk} \boldsymbol{X}_s = \boldsymbol{Z}_p \ , \tag{3.3}$$

whereas $\tilde{\boldsymbol{X}}_s$ and $\tilde{\boldsymbol{Z}}_p$ denote auxiliary processes, represented by the simplified rainfall model in our case, which also satisfy 3.3. k in this equation is the number of fine-scale timesteps within each coarse-scale step (24 for the current application).

The problem is, given a time series $(\boldsymbol{z}_p)$ of the actual process $(\boldsymbol{Z}_p)$, to generate a series $(\boldsymbol{x}_s)$ of the actual process $(\boldsymbol{X}_s)$. To this aim, we first generate another (auxiliary) time series $(\tilde{\boldsymbol{x}}_s)$ using the simplified rainfall process $\left(\tilde{\boldsymbol{X}}_s\right)$. The latter time series is generated independently of $(\boldsymbol{z}_p)$ and, therefore, the $\tilde{\boldsymbol{x}}_s$s do not add up to the corresponding $\boldsymbol{z}_p$s, as required by the additive property (3.3), but to some other quantities, denoted as $(\tilde{\boldsymbol{z}}_p)$. Thus, in the next step, we modify the series $(\tilde{\boldsymbol{x}}_s)$ to produce a series $(\boldsymbol{x}_s)$ which is consistent with $(\boldsymbol{z}_p)$ (in the sense that the $\boldsymbol{x}_s$s and $\boldsymbol{z}_p$s obey (3.3)) without affecting the stochastic structure of the $\tilde{\boldsymbol{x}}_s$s. For this modification we use a linear transformation $f(\tilde{\boldsymbol{X}}_s, \tilde{\boldsymbol{Z}}_p, \boldsymbol{Z}_p)$ whose outcome is a process distributed identically to $(\boldsymbol{X}_s)$ (so that we can write $\boldsymbol{X}_s = f(\tilde{\boldsymbol{X}}_s, \tilde{\boldsymbol{Z}}_p, \boldsymbol{Z}_p)$), and which is also consistent with $(\boldsymbol{Z}_p)$ (it satisfies (3.3)).

Let $\boldsymbol{X}_{(p)}$ be the vector of hourly rainfall values for day p (for 5 locations, $\boldsymbol{X}_{(p)}$ contains $24 \times 5 = 120$ variables). Let also $\boldsymbol{Y}_p$ be a vector containing:

a. the daily values $\boldsymbol{Z}_p$;

b. the daily values $\boldsymbol{Z}_{p+1}$ of the next day and

c. the hourly values of the last hour of the previous day $p-1$ for all locations. This means that for 5 locations $\boldsymbol{Y}_p$ contains $3 \times 5 = 15$ variables in total.

Items (b) and (c) of vector $\boldsymbol{Y}_p$ were included to ensure that the transformation preserves not only the covariance properties among the hourly values of each day, but the covariances with

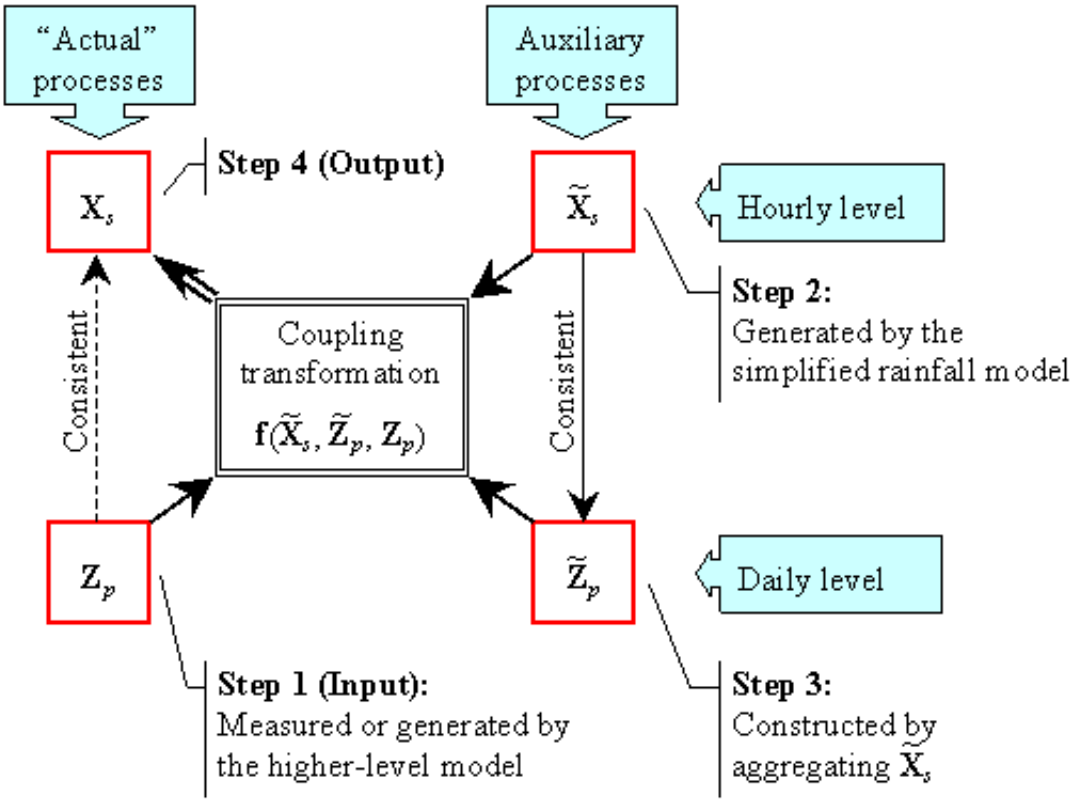

Figure 3.1: Schematic representation of actual and auxiliary processes, their links, and the steps followed to construct the actual hourly-level process from the actual daily-level process.

the previous and next days as well. Note that at day p the hourly values of day $p-1$ are known (therefore, in $\boldsymbol{Y}_p$ we enter hourly values of the previous day) but the hourly values of day $p+1$ are not known This is why, in $\boldsymbol{Y}_p$, we enter daily values of the next day, which are known. In an identical manner, we construct the time-series of variables $\tilde{\boldsymbol{X}}_{(p)}$ and $\tilde{\boldsymbol{Y}}_p$ from the time-series of $\tilde{\boldsymbol{X}}_s$ and $\tilde{\boldsymbol{Z}}_p$.

Koutsoyiannis (2001) showed that there exists a matrix of coefficients $\boldsymbol{h}$ such that if $\boldsymbol{X}$ is generated using

$$\boldsymbol{X}_{(p)} = \tilde{\boldsymbol{X}}_{(p)} + \boldsymbol{h}(\boldsymbol{Y}_p - \tilde{\boldsymbol{Y}}_p) \tag{3.4}$$

then:

1. $\boldsymbol{X}_{(p)}$ has mean and variance-covariance matrix identical to those of $\tilde{\boldsymbol{X}}_{(p)}$, and joint covariance matrix with $\boldsymbol{Y}_p$ identical to that of $\tilde{\boldsymbol{X}}_{(p)}$ and $\tilde{\boldsymbol{Y}}_p$;

2. any linear relationship which holds for $\tilde{\boldsymbol{X}}_{(p)}$ and $\tilde{\boldsymbol{Y}}_p$ and which can be written in the

form

$$\boldsymbol{g}_X^T \tilde{\boldsymbol{X}}_{(p)} = \boldsymbol{g}_Y^T \tilde{\boldsymbol{Y}}_p \tag{3.5}$$

where $\boldsymbol{g}_X$ and $\boldsymbol{g}_Y$ are matrices of coefficients, also holds for $\boldsymbol{X}_{(p)}$ and $\boldsymbol{Y}_p$, that is

$$\boldsymbol{g}_X^T \boldsymbol{X}_{(p)} = \boldsymbol{g}_Y^T \boldsymbol{Y}_p \tag{3.6}$$

Note that the additive property (3.3) can be written in the matrix form (3.6) (for appropriately selected $\boldsymbol{g}_X^T$ and $\boldsymbol{g}_Y^T$) and, therefore, its preservation by the transformation is ensured.

Details of how to determine $\boldsymbol{h}$ in terms of covariance properties of $\boldsymbol{X}_{(p)}$ and $\boldsymbol{Y}_p$ are given by Koutsoyiannis (2001).

Note that a number of peculiarities of the rainfall process at a fine time scale cause specific difficulties that are examined in detail in Wheater *et al.* (2000).

Chapter 4

Testing of multi-site disaggregation procedure

In this chapter we report the results of an exercise testing the performance of the multi-site dissagregation procedure. Data required by the method are hourly and daily time series and hourly and daily cross correlations for the gauges used. In general, daily time series will be available for all gauges and hourly time series will be available for a subset of the gauges used. In this case all daily cross correlations will be known and an estimate for unknown hourly cross correlations will have to be made, based on the known hourly and daily cross correlations. In the following sections, hourly data are used for 2 of the daily gauges to guide disaggregation of the other daily gauges as described in Chapter 3. Section 4.1 evaluates the performance of the disaggregation procedure using 8 hourly gauges in the Brue catchment for the ideal situation in which hourly cross correlations are known for all gauges. The same data are used in Section 4.2 but, more realistically, using hourly cross correlations at 2 gauges to estimate those at the others by means of a scaling method. Comparison of results with those of Section 4.1 gives an indication of how well the scaling method has worked. Section 4.3 uses the scaling method to estimate hourly cross correlations for the Blackwater and North East Lancashire regions, described in chapter 2, and so implement the disaggregation procedure there.

4.1 Brue catchment — performance with known hourly cross-correlations

Figure 2.1 shows the location of hourly gauges in the Brue region. Based on selection of periods with no missing data, two sets, A and B, of data from eight hourly gauges are used of approximately 2 years duration each. These are set A, 8$^{\text{th}}$ September 1994 to 20$^{\text{th}}$ August 1996 with gauges 2, 4, 9, 22, 23, 25, 29 and 33, and set B, 6th October 1995 to 5$^{\text{th}}$ October 1997 with gauges 9, 12, 14, 15, 19, 25, 28 and 31. Daily rainfall has been disaggregated for both

data sets, using hourly data at 2 gauges (9 and 48 for set A and 9 and 31 for set B) and known hourly cross correlations at all gauges. The software written to implement the disaggregation procedure has a number of options which may be set by the user. For consistency with the observed data, hourly rainfalls below a threshold of 0.2mm, corresponding to the bucket size of the tipping bucket gauges, are set to zero. Having imposed this threshold the software sums the total disaggregated hourly rainfall at each gauge during each day and ensures that it agrees with daily records (within user specified tolerances).

Figures 4.1 and 4.2 show examples of disaggregation results with and without the 0.2mm threshold for winter (months December, January and February) and summer (months June, July and August) respectively using data set B. Shown on the plots are standard deviation, proportion dry and autocorrelations lag one and two for actual hourly data and for disaggregated hourly data. Also shown for reference is the standard deviation for hourly rainfall obtained by uniformly allocating data at the daily scale over hourly intervals. This gives an indication of how much better the disaggregation procedure peforms than simply uniformly allocating the daily data. For both winter and summer the application of a 0.2mm threshold gives much better performance for proportion dry as well as yielding an improved performance for the other statistics. This threshold is hence applied in all subsequent analyses. Figures 4.3 and 4.4 show autocorrelations between gauges for winter and summer respectively for the same data set and Figures 4.5 and 4.6 show the frequency plots for the number of active (wet) gauges again for winter and summer respectively. Notice that the disaggregation performance is worse in terms of standard deviation and autocorrelations for summer than for winter due to the higher variability of summer rainfall. The disaggregation procedure reproduces well the frequency distributions for the number of active gauges for both winter and summer. Results for spring and autumn are better than for summer but not quite as good as for winter. In interpreting results, note that they have been obtained using known hourly correlations and as such represent the best achievable performance using this disaggregation procedure.

4.2 Brue catchment — performance using daily–hourly correlation relationships

In this section a more realistic assumption is made that hourly cross correlations are only known between 2 of the gauges for the data set used in the previous section. Hourly cross correlations are used from this pair of gauges together with daily cross correlations from all gauges to estimate the unknown hourly cross correlations. The disaggregation procedure is then implemented and results compared with those of the previous section.

A simple scheme used here to estimate the unknown hourly cross correlations makes use of the approximate scaling of rainfall time series. A rainfall intensity series, X_λ say, sampled at a time scale λ, is said to be *multiscaling* if

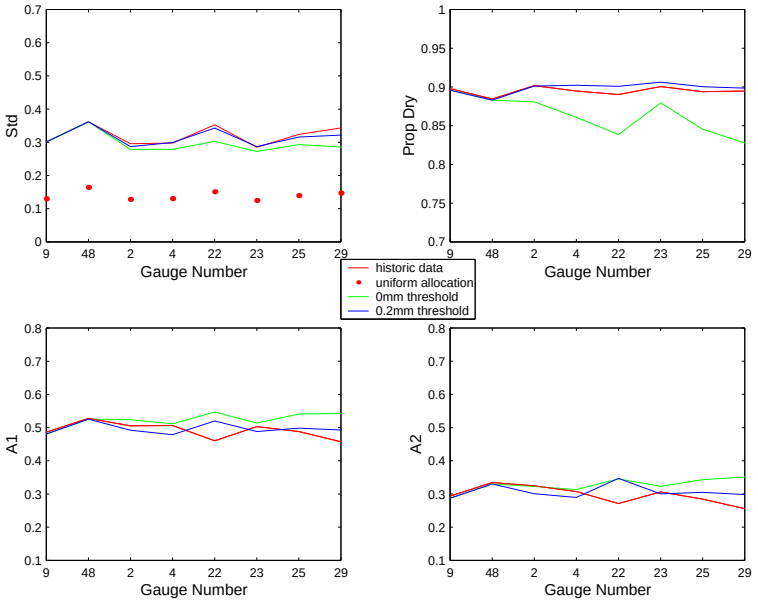

Figure 4.1: Statistics of disaggregated rainfall using thresholds of 0mm and 0.2mm together with those for actual hourly rainfall during winter. From left to right and top to bottom are hourly standard deviation, proportion of dry intervals, autocorrelations lag one and two.

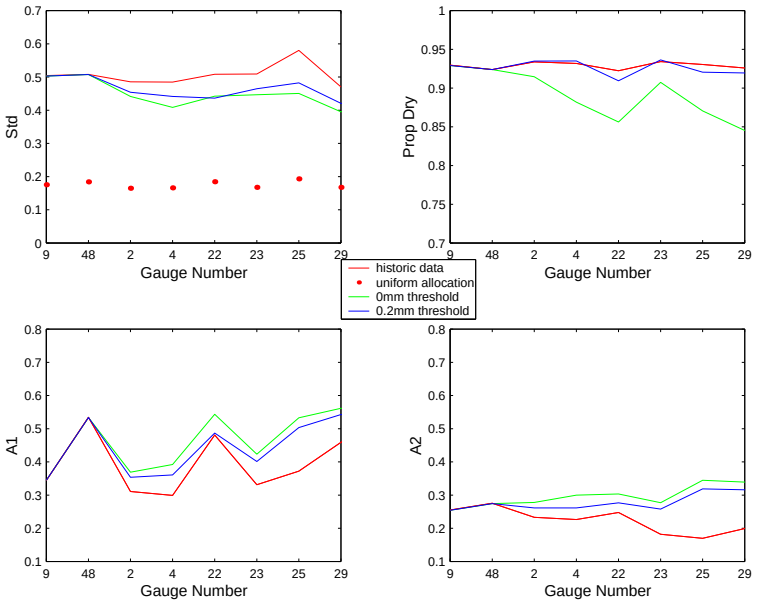

Figure 4.2: Statistics of disaggregated rainfall using a threshold of 0mm and 0.2mm together with those for actual hourly rainfall during summer. From left to right and top to bottom are hourly standard deviation, proportion of dry intervals, autocorrelations lag one and two.

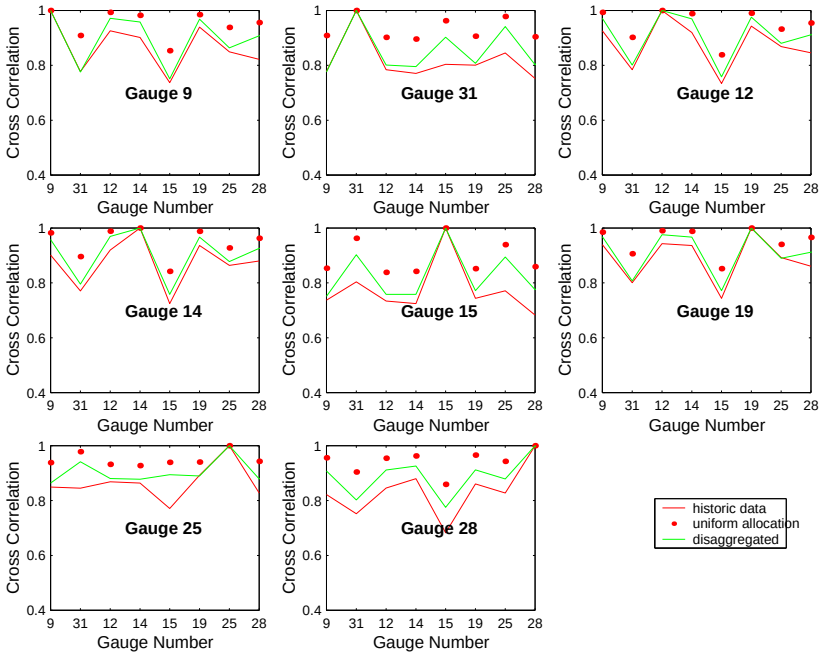

Figure 4.3: Hourly cross correlations for disaggregated and historic rainfall during winter for data set B.

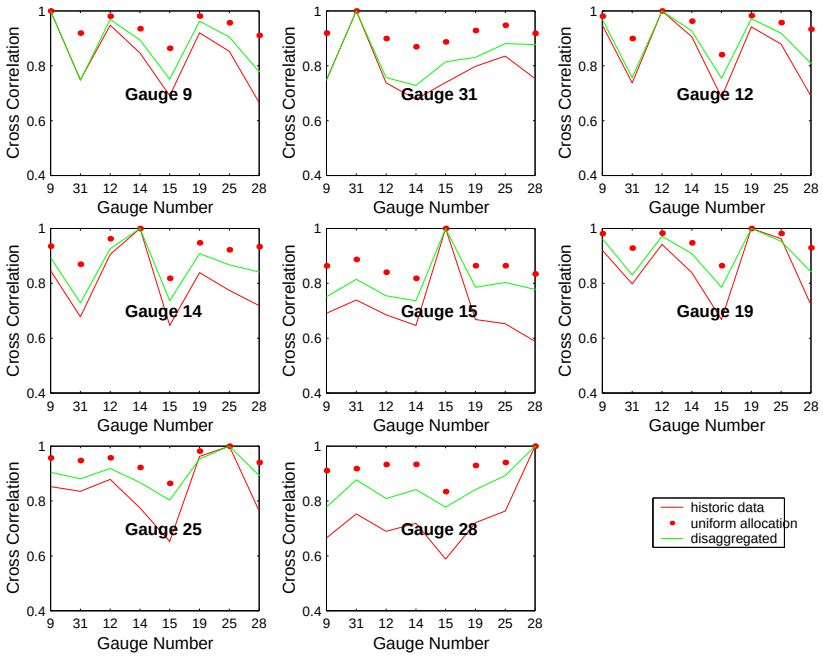

Figure 4.4: Hourly cross correlations for disaggregated and historic rainfall during summer for data set B.

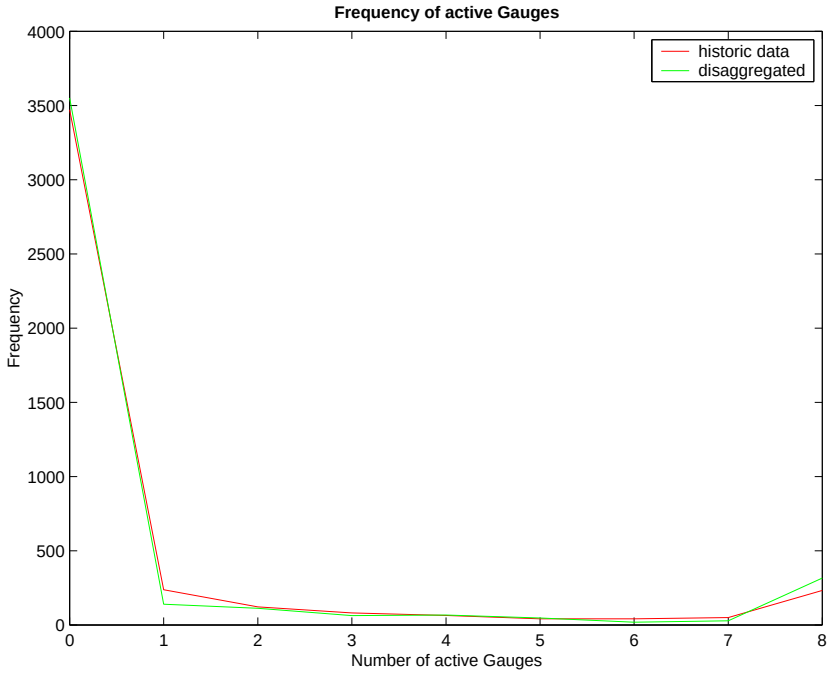

Figure 4.5: Frequency distribution of number of active gauges for disaggregated and historic hourly rainfall for data set B during winter.

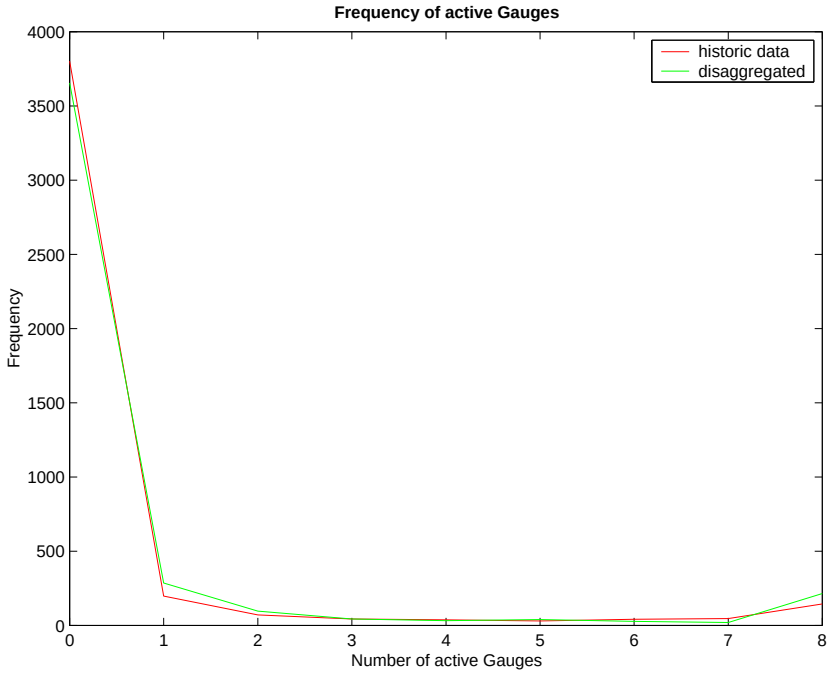

Figure 4.6: Frequency distribution of number of active gauges for disaggregated and historic hourly rainfall for data set B during summer.

$$E\left[X_{a\lambda}^{q}\right] = a^{k(q)} E\left[X_{\lambda}{}^{q}\right] \tag{4.1}$$

where $k(q)$ is the *scaling function* for the qth moment. For *simple* scaling $k(q) = kq$ where k is a constant.

The correlation between 2 gauges at a time scale λ, $C\left(X_\lambda, Y_\lambda\right)$ can be written

$$C\left(X_\lambda, Y_\lambda\right) = \frac{E\left[X_\lambda Y_\lambda\right] - E\left[X_\lambda\right] E\left[Y_\lambda\right]}{\left[\left(E\left[X_\lambda{}^2\right] - E\left[X_\lambda\right]^2\right)\left(E\left[Y_\lambda{}^2\right] - E\left[Y_\lambda\right]^2\right)\right]^{\frac{1}{2}}} \tag{4.2}$$

Since $E\left[X_\lambda\right] = E\left[X\right]$ is scale invariant then the only scale dependent components of the right hand side of equation 4.2 are $E\left[X_\lambda Y_\lambda\right]$, $E\left[X_\lambda{}^2\right]$ and $E\left[Y_\lambda{}^2\right]$. If these scale then equation 4.1 leads to

$$\begin{aligned} E\left[X_{a\lambda}{}^2\right] &= a^{\theta} E\left[X_\lambda^2\right] \\ i.e. \quad \ln\left(E\left[X_{a\lambda}{}^2\right]\right) &= \theta \ln a + \ln\left(E\left[X_\lambda^2\right]\right) \end{aligned} \tag{4.3}$$

and, taking λ to be 1 hour, $\ln\left(E\left[X_{a\lambda}^2\right]\right)$ is a linear function of $\ln a$ with gradient θ and intercept $\ln\left(E\left[X_1^2\right]\right)$.

Similarly

$$\ln\left(E\left[X_{a\lambda} Y_{a\lambda}\right]\right) = \phi \ln a + \ln\left(E\left[X_\lambda Y_\lambda\right]\right) \tag{4.4}$$

Estimates of θ and ϕ can be made by plotting log-moments against $\ln a$ for gauges at which hourly data (and hence data at other aggregated scales) are available and fitting straight lines to the points. If θ and ϕ do not vary over the region, similar lines can be fitted with these gradients to gauges at which only daily data are available. These can then be used to estimate the hourly statistics which are components of the hourly cross correlations and hence the cross correlations (equation 4.2). Hourly data are required from at least 2 gauges to use this method and, where there are multiple estimates of θ and ϕ, mean values can be taken. Due to the high cross correlation of the rainfall series at nearby gauges (typically between 0.7 and 0.9 at the hourly scale for the Brue region) any deviation from scaling behaviour will be qualitatively similar for $E\left[X_\lambda^2\right]$ and $E\left[X_\lambda Y_\lambda\right]$ so that errors in them will act to compensate each other in equation 4.2.

This method for obtaining hourly cross correlations has been implemented for the Brue region, for each season, using hourly data from only 2 gauges and daily data at all gauges for the data subsets A and B (Section 4.1) to obtain estimates of hourly cross correlations at all daily gauges. These estimates have been used together with data from the 2 hourly gauges to disaggregate the daily data using the described procedure in Chapter 3.

Figures 4.7 and 4.8 show the straight lines fitted to estimate θ and ϕ from gauges 9 and 31. For each season the mean values of θ and ϕ have been used to estimate the hourly cross correlations for the whole region by fitting lines with these gradients through points representing the 24, 48, 72 and 96 hour statistics as described above, and using equation 4.2.

Figures 4.9 and 4.10 show hourly cross correlations estimated using the scaling method for winter and summer respectively together with actual values for set B. These values for hourly cross correlations are then used to dissagregate daily rainfall using the disaggregation procedure (Chapter 3). Figures 4.11 and 4.12 show, for winter and summer respectively, plots of hourly statistics for rainfall disaggregated using known hourly cross correlations together with those using estimated hourly cross correlations. *For both winter and summer, disaggregation peformance using estimated and actual hourly correlations is similar* and is worse in summer, when rainfall is more variable. In summary, the methodology has been successful in this case, however for the Brue region distances between hourly gauges are relatively short. In many typical cases raingauges will be further apart and hourly cross correlations will be lower. Two regions where this is the case, the Blackwater and North East Lancashire regions, are studied in the next section.

4.3 Results for other regions

For both the Blackwater and North East Lancashire regions, hourly data have been used from three hourly gauges to disaggregate daily data at eight daily gauges (including the hourly gauges). By only using data at two of the three hourly gauges at a time this enables the third to be used to indicate performance of the disaggregation procedure using hourly cross correlations estimated with the scaling method.

4.3.1 The Blackwater

Hourly gauges used for the Blackwater region are B03, B30, and B06 (Figure 2.2). The scaling method has been used in the same way as in section 4.2 to estimate hourly correlations. Figure 4.13 shows estimated correlations for all seasons and using all combinations of pairs of the three hourly gauges available. In each plot, one gauge is kept fixed (that for which the cross correlation is 1) and is also the gauge whose hourly data have not been used during the estimation (i.e. the plots show performance at gauges for which hourly data were treated as unknown). As before performance is worse in summer and overall results are reasonable.

Figures 4.14 and 4.15 show plots of statistics for disaggregated and actual hourly data for winter and summer respectively (using hourly data from gauges B03 and B30 and treating gauge B06 as a daily gauge) together with statistics for actual hourly data. Results for other pairs of gauges are qualitatively similar. Results are not as good as for the Brue region as is to be expected due to the larger distances between gauges. But again, these results are plausible and encouraging given the simplicity of the scheme used to estimate hourly cross correlations.

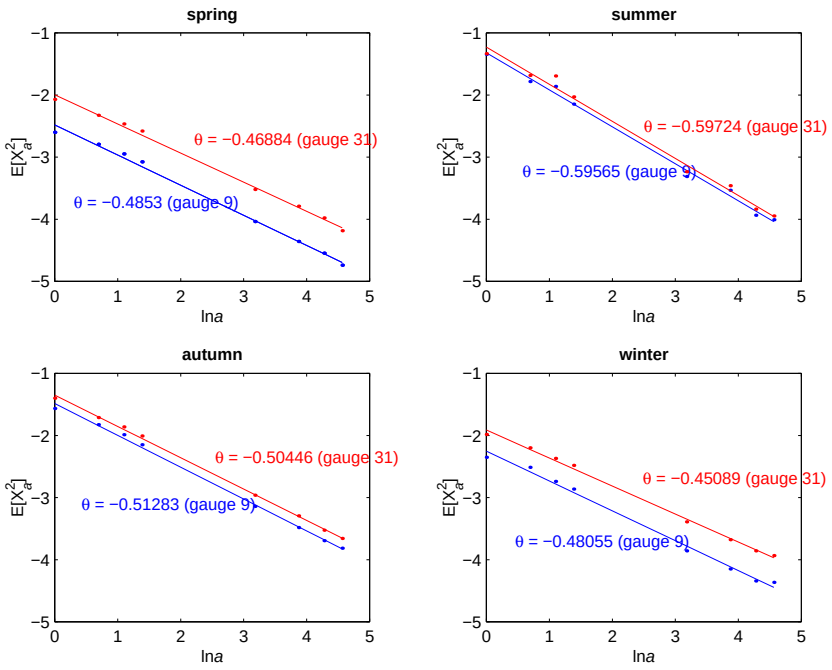

Figure 4.7: Approximate scaling of $E\left[X^2\right]$ for gauges 9 and 31.

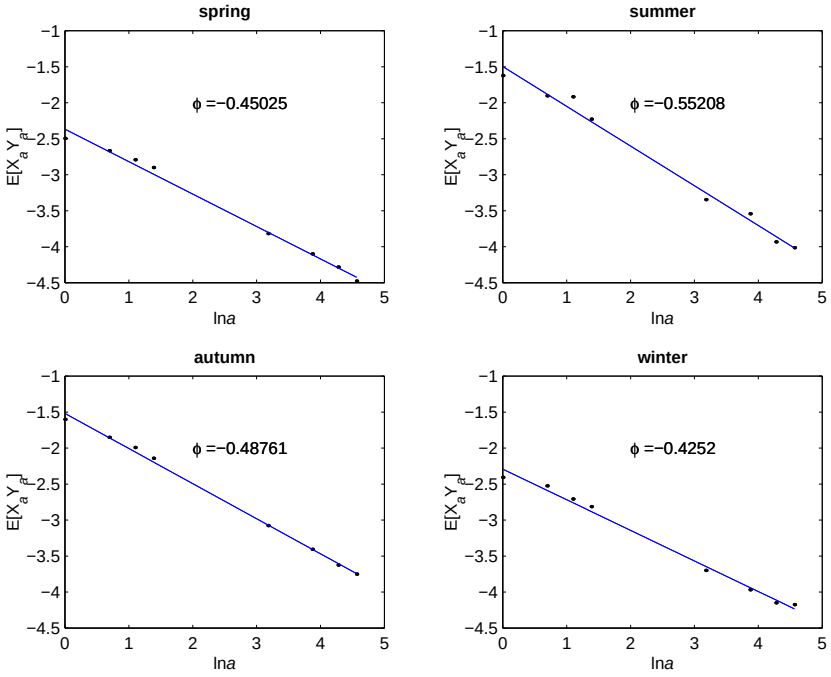

Figure 4.8: Approximate scaling of $E\left[XY\right]$ for gauges 9 and 31.

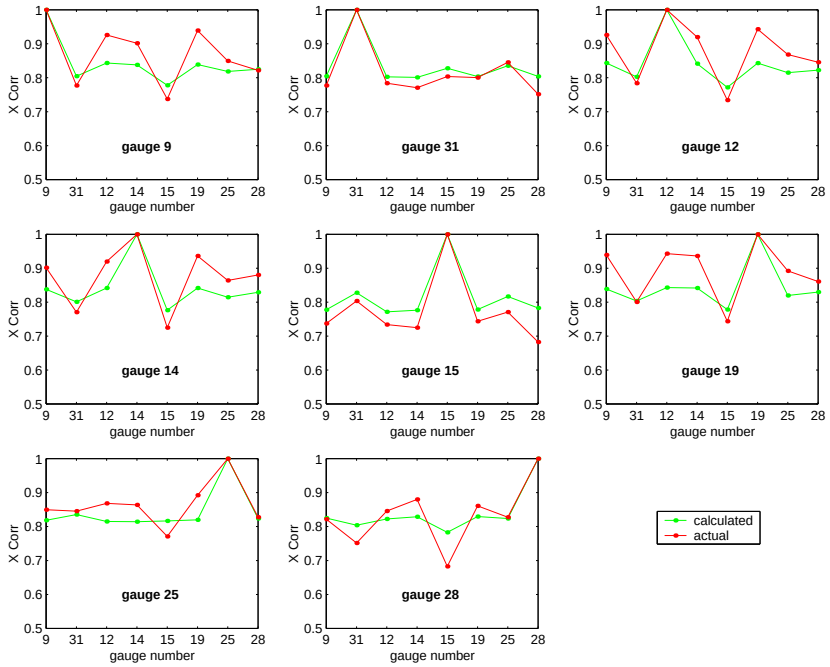

Figure 4.9: Estimated hourly cross correlations using the scaling method and actual hourly cross correlations for data set B during winter.

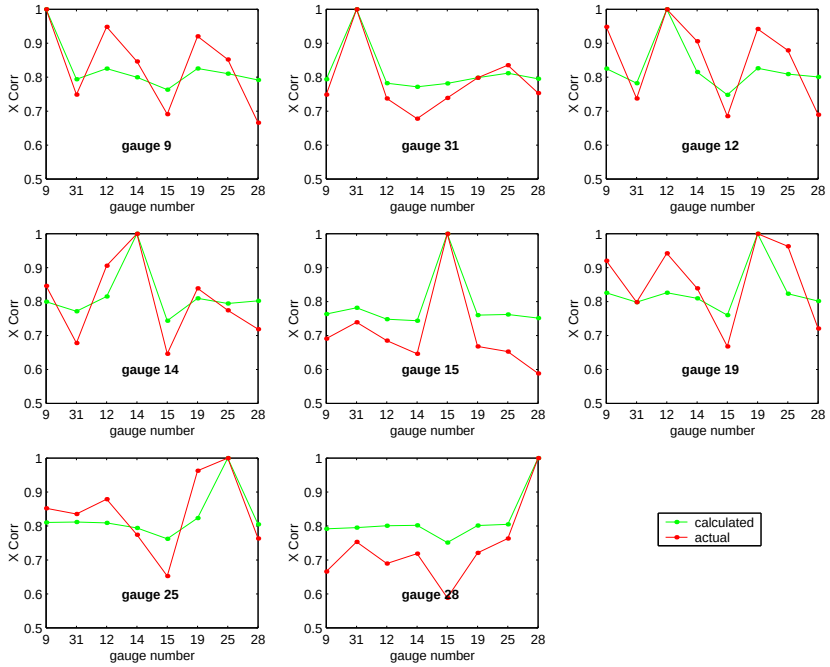

Figure 4.10: Estimated hourly cross correlations using the scaling method and actual hourly cross correlations for data set B during summer.

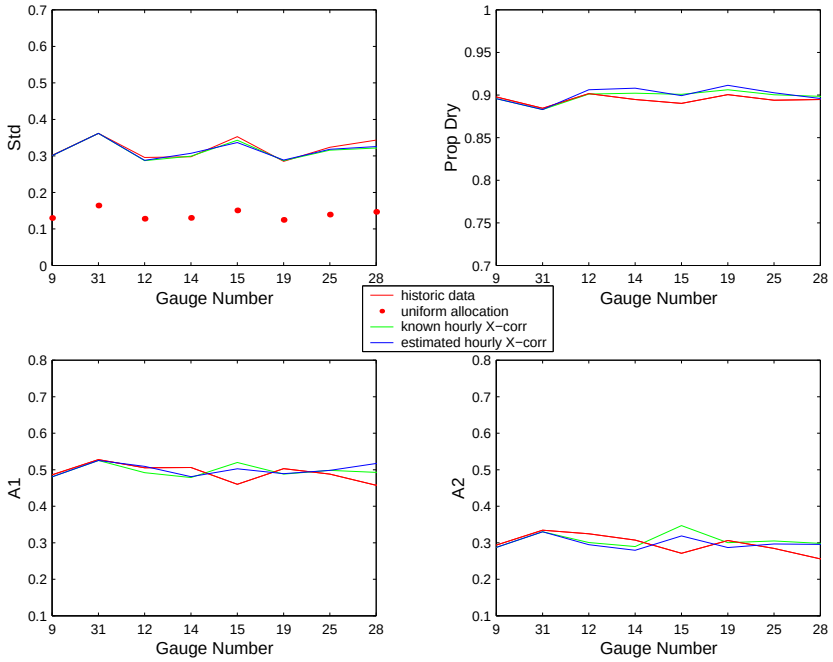

Figure 4.11: Statistics for disaggregated rainfall using estimated and actual hourly cross correlations for data set B during winter (standard deviation, proportion of dry hours and autocorrelations lag one and two).

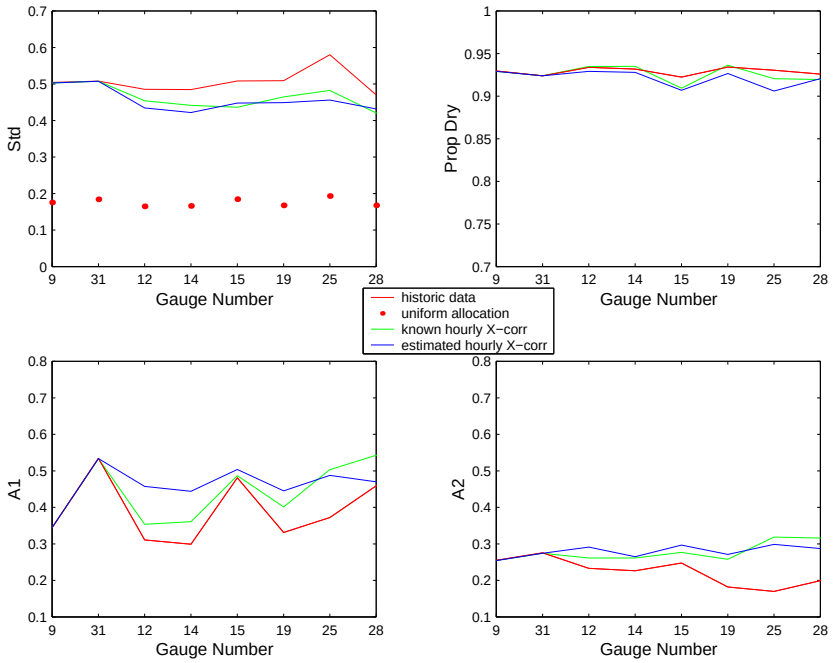

Figure 4.12: Statistics for disaggregated rainfall using estimated and actual hourly cross correlations for data set B during summer (standard deviation, proportion of dry hours and autocorrelations lag one and two).

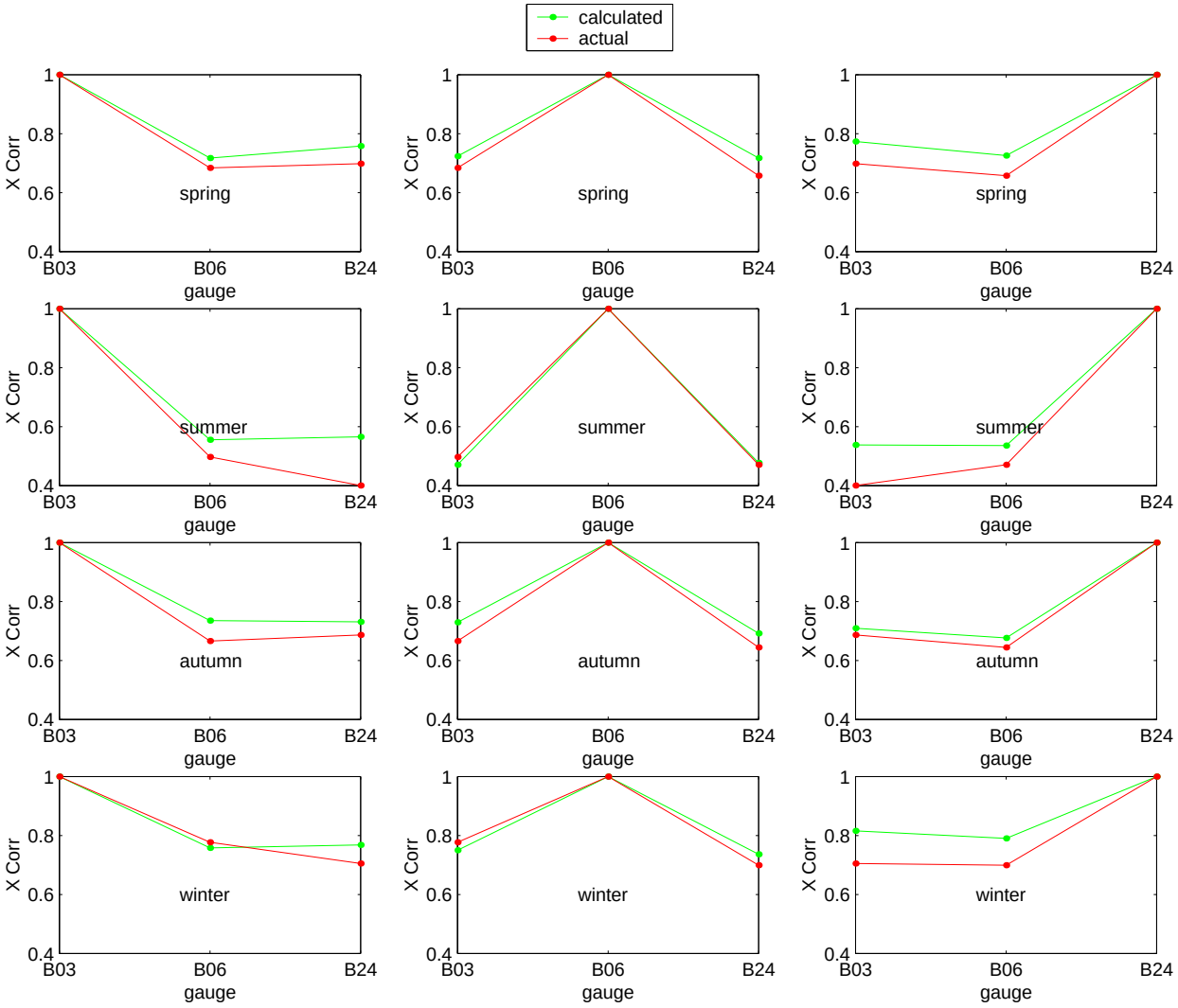

Figure 4.13: Estimated hourly cross correlations for the Blackwater region. Left column: cross-correlations involving site B03, estimated for sites B06 and B29 alone. Middle column: cross-correlations involving site B06, estimated for sites B03 and B29 alone. Right column: cross-correlations involving site B29, estimated for sites B03 and B06 alone.

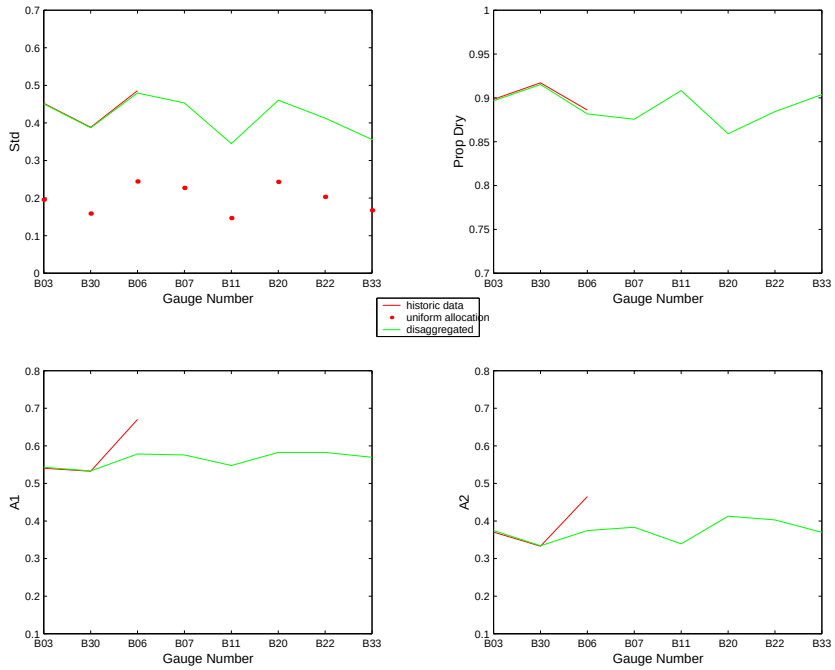

Figure 4.14: Statistics for disaggregated rainfall using estimated hourly cross correlations and actual values when known for the Blackwater region during winter (standard deviation, proportion of dry hours and autocorrelations lag one and two).

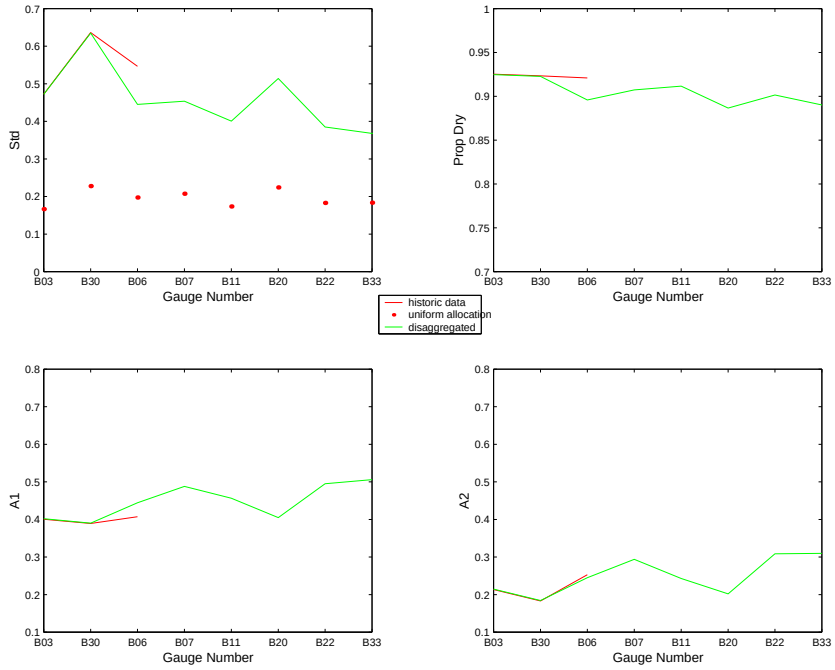

Figure 4.15: Statistics for disaggregated rainfall using estimated hourly cross correlations and actual values when known for the Blackwater region during summer (standard deviation, proportion of dry hours and autocorrelations lag one and two).

4.3.2 North-East Lancashire

For the North East Lancashire region hourly data hve been used from gauges L37, L38 and L99. The same methods were used as for the Blackwater region (section 4.3.1) and Figures 4.16, 4.17 and 4.18 show results in the same way. Again results are not as good as for the Brue region due here to larger distances between gauges and differences in altitude of gauges (not a significant feature of the Brue region).

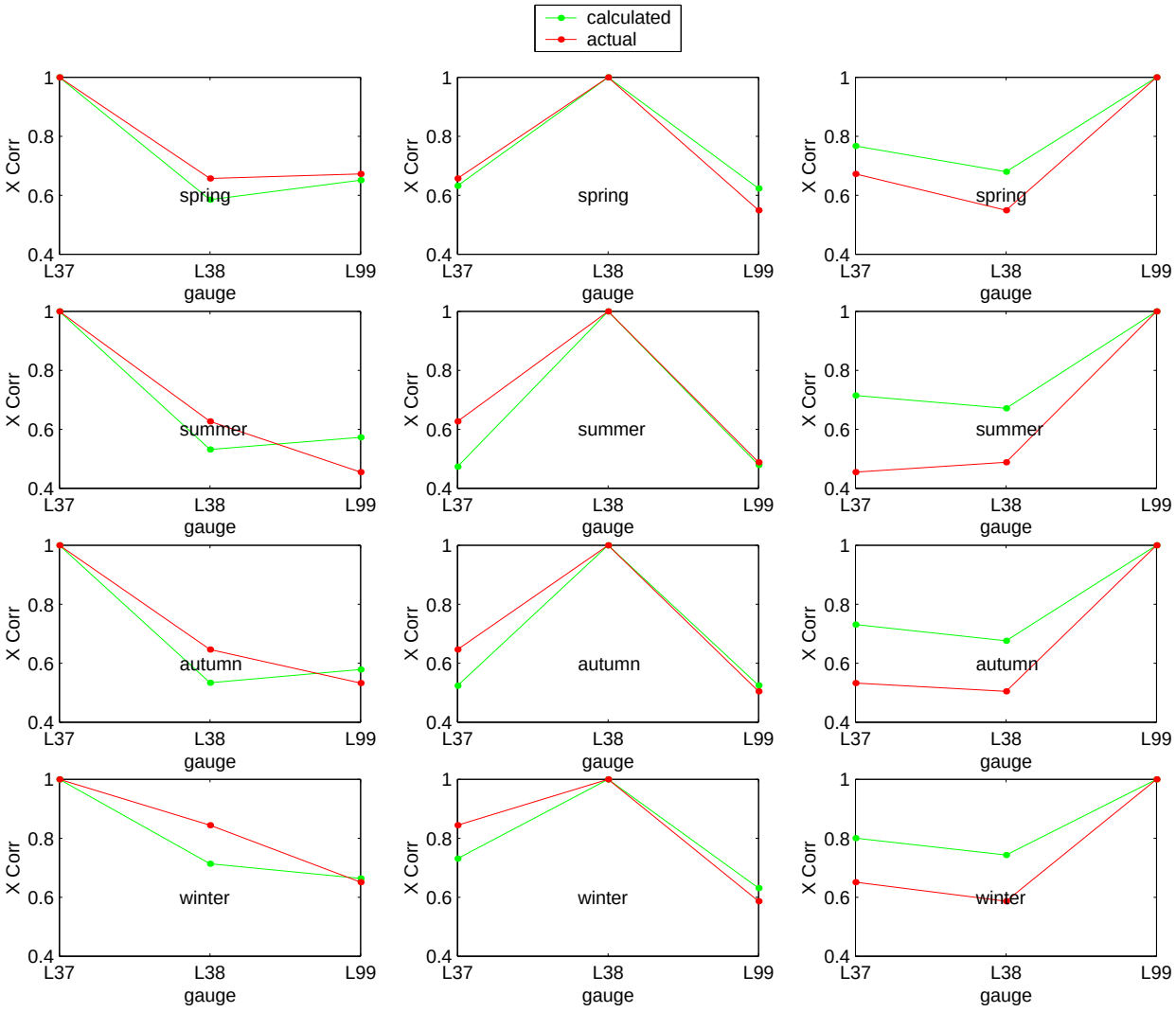

Figure 4.16: Estimated hourly cross correlations for the North-East Lancashire region. As for figure 4.13.

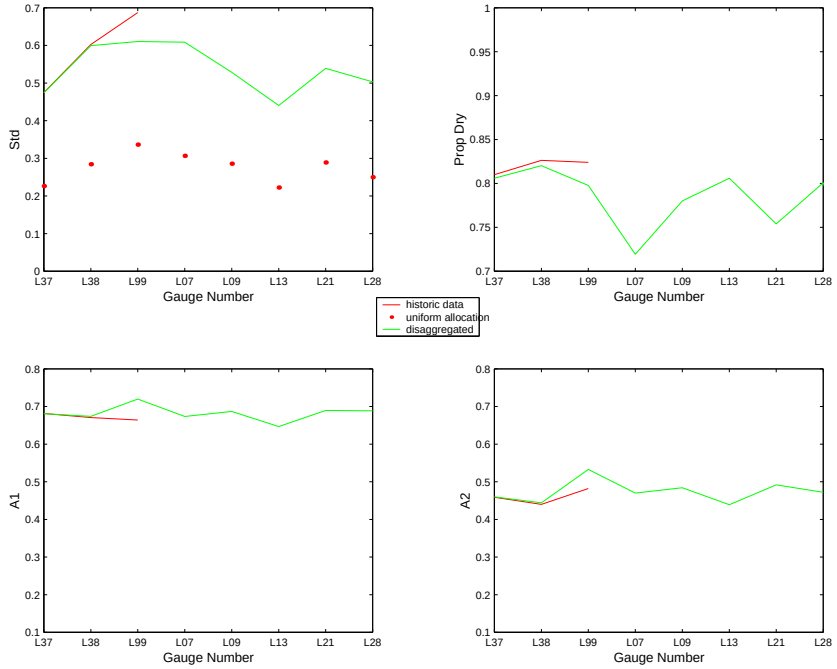

Figure 4.17: Statistics for disaggregated rainfall using estimated hourly cross correlations and actual values when known for the North-East Lancashire region during winter (standard deviation, proportion of dry hours and autocorrelations lag one and two).

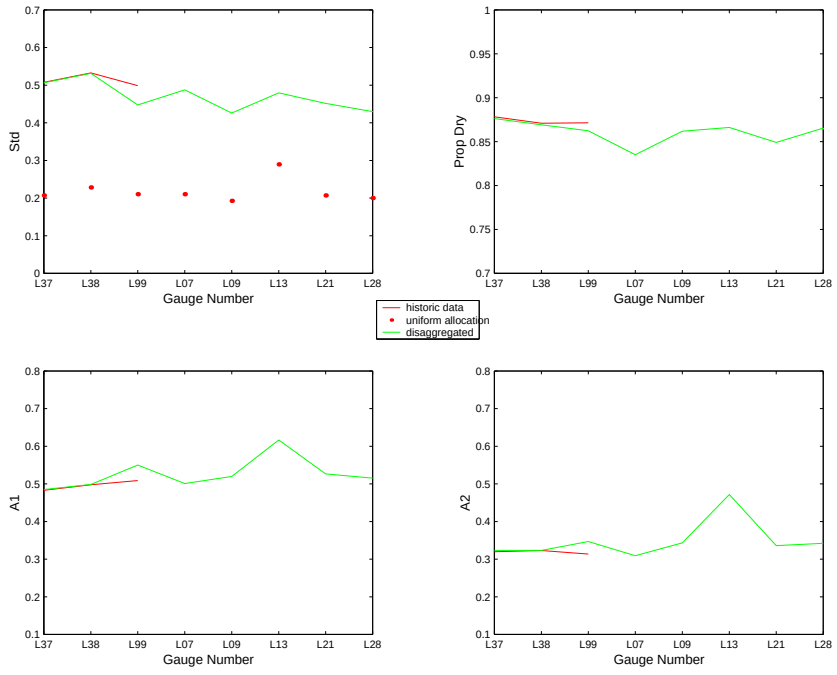

Figure 4.18: Statistics for disaggregated rainfall using estimated hourly cross correlations and actual values when known for the North-East Lancashire region during summer (standard deviation, proportion of dry hours and autocorrelations lag one and two).

Chapter 5

Generalized Linear Modelling of daily rainfall

In the previous chapter, the performance of the multi-site disaggregation procedure was tested by disaggregating observed daily rainfall sequences down to an hourly timescale. We now tackle the problem of modelling daily sequences at a network of sites. The ultimate aim of such an exercise is to produce simulated sequences of daily rainfall, which can be disaggregated to a subdaily timescale if necessary. There are two reasons for wanting to do this:

- Daily data are more widely available than subdaily data, so that structures in daily rainfall sequences (for example, trends) can be detected and parameterised more reliably than those in relatively scarce (and short) subdaily sequences.
- In the light of current concerns regarding changing climate and its potential impacts, scenarios based on disaggregation of historical daily records may not be valid into the future; therefore, the use of model-based scenarios to investigate possible future trends is potentially beneficial.

The possibility of using model-based daily sequences to derive hourly values was discussed in section 3.4.

As described in chapter 1, our daily rainfall modelling is based on Generalized Linear Models (GLMs). Our previous report (Wheater *et al.* 2000) gave a full description of the methodology, and presented three case studies by way of illustration. By and large, the results were extremely promising: the models were able to detect realistic structures in noisy rainfall records, and most properties of simulated sequences agreed closely with those observed in data. However, there was a tendency for the models to underestimate the probability that any two sites would both experience rain on the same day. In this chapter, we present an alternative mechanism for representing the spatial dependence in rainfall occurrence, and further explore the performance of the methodology using new datasets.

We start by giving a brief overview of the GLM methodology. For full details, see Wheater *et al.* (2000).

5.1 Overview of methodology

The basic idea underlying GLMs is to predict a probability distribution for each day's rainfall at every site of interest, by relating the mean of that distribution to the values of various other related quantities which we call *predictors*. Possible predictors include previous days' rainfall amounts, the month of the year and variables representing topographic effects. Our implementation broadly follows that of Coe and Stern (1982) and Stern and Coe (1984), who adopted a two-stage approach as follows:

Stage 1: model the pattern of wet and dry days at a site using logistic regression. If we denote by p_i the probability of rain for the ith case in the dataset, conditional on a vector of predictors $\boldsymbol{x}_i = (x_{1i}\ x_{2i}\ \ldots\ x_{pi})'$, then the logistic regression model is given by

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i\boldsymbol{\beta}\ = \sum_{j=1}^{p}\beta_j x_{ji}\ . \tag{5.1}$$

Stage 2: fit gamma distributions to the amount of rain on wet days. The rainfall amount for the ith wet day in the database is taken, conditional on a predictor vector $\boldsymbol{\xi}_i$, to have a gamma distribution with mean μ_i where

$$\ln\mu_i = \boldsymbol{\xi}_i\boldsymbol{\gamma}\ = \sum_{j=1}^{q}\gamma_j\xi_{ji}\ . \tag{5.2}$$

for some coefficient vector $\boldsymbol{\gamma}$. All gamma distributions are assumed to have a common shape parameter, ν say (if $\nu = 1$ the distributions are exponential).

These two models will be referred to as 'occurrence' and 'amounts' models respectively.

In the GLM framework, model fitting (estimation of the coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) and selection can be carried out using likelihood methods. Models can be checked using a variety of simple but informative residual plots; these will be illustrated in Section 5.3 below.

Further features that can be accommodated include interactions between predictors (two predictors are said to interact if the effect of one of them depends upon the value of the other), and the estimation of nonlinear transformations of predictors. Interactions are useful as they tell us about the mechanisms driving the rainfall process. For example, Chandler and Wheater (2001) found in a model for Irish rainfall that there was a significant interaction between the North Atlantic Oscillation (NAO) and predictors representing the seasonal cycle — increases in the NAO are associated with increases in winter rainfall, but have little effect in the summer months. This agrees with our understanding of the NAO as a phenomenon whose effects are mainly confined to the Northern Hemisphere winter (Hurrell 1995). One

of the potential advantages of the GLM methodology is that it allows us to incorporate such structures into simulated rainfall sequences.

5.1.1 Multisite simulation

When using a GLM to simulate rainfall sequences at several sites, it is necessary to allow for the fact that neighbouring sites are not independent. To account for spatial dependence in rainfall amounts, we study the correlation structure among the Anscombe residuals from the amounts model: these are residuals defined in such as way as to be normally distributed to a good degree of approximation. Once it is known which sites are wet on a particular day, it is therefore straightforward to allocate dependent rainfall amounts to these sites: simulate a correlated vector of Anscombe residuals (using standard algorithms for the generation of multivariate normal random vectors), and then back-transform the residuals at each site, to obtain values drawn from the modelled gamma distributions.

Dependence in rainfall occurrence is more difficult to deal with. In previous work, it was modelled using a binary 'weather state' variable representing 'wet' and 'dry' days in an area — on a wet day, the probability of rainfall at all sites is increased whereas on a dry day it is decreased. This was intended as a simple representation of the mechanism responsible for dependence in occurrence at spatial scales of interest — the dependence is mainly due to the fact that all sites tend to be influenced by the same weather systems on particular days. However, it can be shown theoretically that, under this mechanism, the probability of all sites being either wet or dry cannot be arbitrarily close to 1. The effect of this, in practice, is that multi-site simulations of a GLM do not reproduce the observed distribution of numbers of wet sites — the simulations do not give enough days when all sites are either wet or dry.

5.2 Theoretical developments

For hydrological purposes, it may be important to reproduce accurately the distribution of numbers of wet sites, since this is related to the proportion of an area which experiences rain. In view of the failure of the 'binary weather state' model in this respect, the current phase of research has investigated other possible mechanisms. There are a variety of methods in the literature for simulating dependent binary random variables: most are based either upon hidden variables (our 'weather state' model is a simple example of this), or upon correlation structure. Of all the correlation-based approaches, that of Oman and Zucker (2001) appears most promising, and it would be of interest to explore this in the future. However, there are a number of potential problems with the use of correlations to specify dependencies in binary data (see Section 4.1.5 of Wheater *et al.* (2000) for a discussion). In view of these potential difficulties, and of the limited timescale of this phase of research, we have not investigated correlation-based methods here. Rather, since one of the primary concerns is to reproduce the distribution of numbers of wet sites, we investigate a new method of incorporating spatial dependence, by modelling this distribution directly. This section gives details of the theory

involved.

5.2.1 Notation

We begin by establishing some notation. We wish to simulate, for any given day t, a vector of dependent binary random variables, $\boldsymbol{Y}_t = (Y_{1t} \ \ldots \ Y_{S_t t})'$ (S_t is the number of sites we are studying on day t). Our rainfall occurrence model allows us to calculate $E(Y_{st}) = p_{st}$, say. A (non-unique) dependence structure can be specified for $\boldsymbol{Y}_t$ through the distribution of $Z_t = \sum_{s=1}^{S_t} Y_{st}$. Since the p_{st}s vary from day to day, so does the distribution of Z_t — in particular, we have $\mathrm{E}(Z_t) = \sum_{s=1}^{S_t} p_{st}$.

5.2.2 Choice of distribution for Z_t

A flexible family of distributions for discrete random variables taking values in $\{0, 1, \ldots, S_t\}$ is the Beta-Binomial family:

$$P(Z_t = z) = \binom{S_t}{z} \frac{\Gamma(\alpha_t + z)\,\Gamma(S_t + \beta_t - z)\,\Gamma(\alpha_t + \beta_t)}{\Gamma(\alpha_t + \beta_t + S_t)\,\Gamma(\alpha_t)\,\Gamma(\beta_t)} \qquad (z = 0, 1, \ldots, S_t)\ , \tag{5.3}$$

for some parameters $\alpha_t, \beta_t \in \mathbb{R}^+$. For small values of S_t, these probabilities can be evaluated cheaply using a recurrence relation. The mean and variance of the distribution are

$$\frac{S_t \alpha_t}{\alpha_t + \beta_t} \qquad \text{and} \qquad \frac{S_t \alpha_t \beta_t (\alpha_t + \beta_t + S_t)}{(\alpha_t + \beta_t)^2 (\alpha_t + \beta_t + 1)}\ , \tag{5.4}$$

respectively. In standard applications, the distribution arises as that of a Binomial (S_t, θ) random variable, where θ is itself a random variable distributed according to a Beta distribution with parameters α_t and β_t. The uniform distribution corresponds to the special case $\alpha_t = \beta_t = 1$. If $\alpha_t = 0, \beta_t \neq 0$ then $P(Z_t = 0) = 1$, and if $\beta_t = 0, \alpha_t \neq 0$ then $P(Z_t = S_t) = 1$. The case when $\alpha_t = \beta_t = 0$ is discussed below.

It is convenient to reparametrise the Beta-Binomial distribution here: set

$$\theta_t = \frac{\alpha_t}{\alpha_t + \beta_t} \qquad \text{and} \qquad \phi_t = \alpha_t + \beta_t\ , \tag{5.5}$$

so that

$$\alpha_t = \theta_t \phi_t\ , \qquad \beta_t = \phi_t (1 - \theta_t)\ , \qquad \mathrm{E}(Z_t) = S_t \theta_t \text{ and } \mathrm{Var}(Z_t) = \frac{S_t \theta_t (1 - \theta_t)(\phi_t + S_t)}{\phi_t + 1}\ . \tag{5.6}$$

We can think of θ_t as a mean value parameter, and ϕ_t as a dispersion parameter (in fact, ϕ_t essentially controls the tendency of the distribution to be concentrated either at its extremities or around its mean — see Figure 5.1). As a first attempt at modelling in this way it will be convenient, and not implausible, to assume that $\phi_t = \phi$ is constant for all t, so

that θ_t (which is known, since it can be calculated from our rainfall occurrence model) is the only time-varying parameter of the distribution. As θ_t varies therefore, the effect is to move along one of the rows of Figure 5.1 — in this way we hope to reproduce typical 'summer' and 'winter' distributions of numbers of wet sites, for example.

The reparametrisation also allows us to investigate the shape of the distribution when $\alpha_t = \beta_t = 0$. In this case, we have $\phi = 0$. If we consider $\lim_{\phi \to 0} P(Z_t = 0)$ with θ_t fixed, we find that in the limit Z_t takes the values 0 and S_t with probabilities $1 - \theta_t$ and θ_t respectively. At the other extreme, it can be shown that the limiting distribution as $\phi \to \infty$ is Binomial with parameters S_t and θ_t. Since the Binomial arises if all the Y_{st}s are independent and identically distributed, we see that ϕ can be regarded as an overall summary of the dependence among the Y_{st}s — small values of ϕ correspond to strong dependence.

5.2.3 Fitting

Given data $\{(S_t, Z_t, \theta_t) : t = 1, \ldots, T\}$, a natural way to estimate the dispersion parameter ϕ is via a method of moments. Note that

$$\mathrm{E}\left(\frac{(Z_t - S_t\theta_t)^2}{S_t\theta_t(1-\theta_t)}\right) = \frac{\phi + S_t}{\phi + 1} = E\left(R_t^2\right) \quad \text{say.}$$

Hence

$$\mathrm{E}\left(\sum_{t=1}^{T} R_t^2\right) = T + \frac{1}{\phi+1}\sum_{t=1}^{T}(S_t - 1) \ ,$$

so that a natural estimator of ϕ is

$$\hat{\phi} \ = \ \frac{\sum_{t=1}^{T}(S_t - 1)}{\sum_{t=1}^{T}(R_t^2 - 1)} - 1 \ . \tag{5.7}$$

In the simple case where all the p_{st}s are equal and the Ys are independent, we have $Z_t \sim Bin(S_t, \theta_t)$, and $\mathrm{Var}(Z_t) = S_t\theta_t(1-\theta_t)$, whence $\mathrm{E}(R_t^2) = 1$. If all the R_t^2s were equal to 1, (5.7) would give $\hat{\phi} = \infty$ (corresponding to independence): overdispersion results in lower values of $\hat{\phi}$ as expected.

5.2.4 Interpretation of the model

In the special case when the probability of rain is the same at all sites, the beta-binomial model has an appealing heuristic interpretation. According to the discussion in Section 5.2.2, the beta-binomial distribution for Z_t can be generated in 2 steps:

1. Generate θ_t from a beta distribution with parameters (α_t, β_t).
2. Generate Z_t from a binomial distribution with parameters (S_t, θ_t).

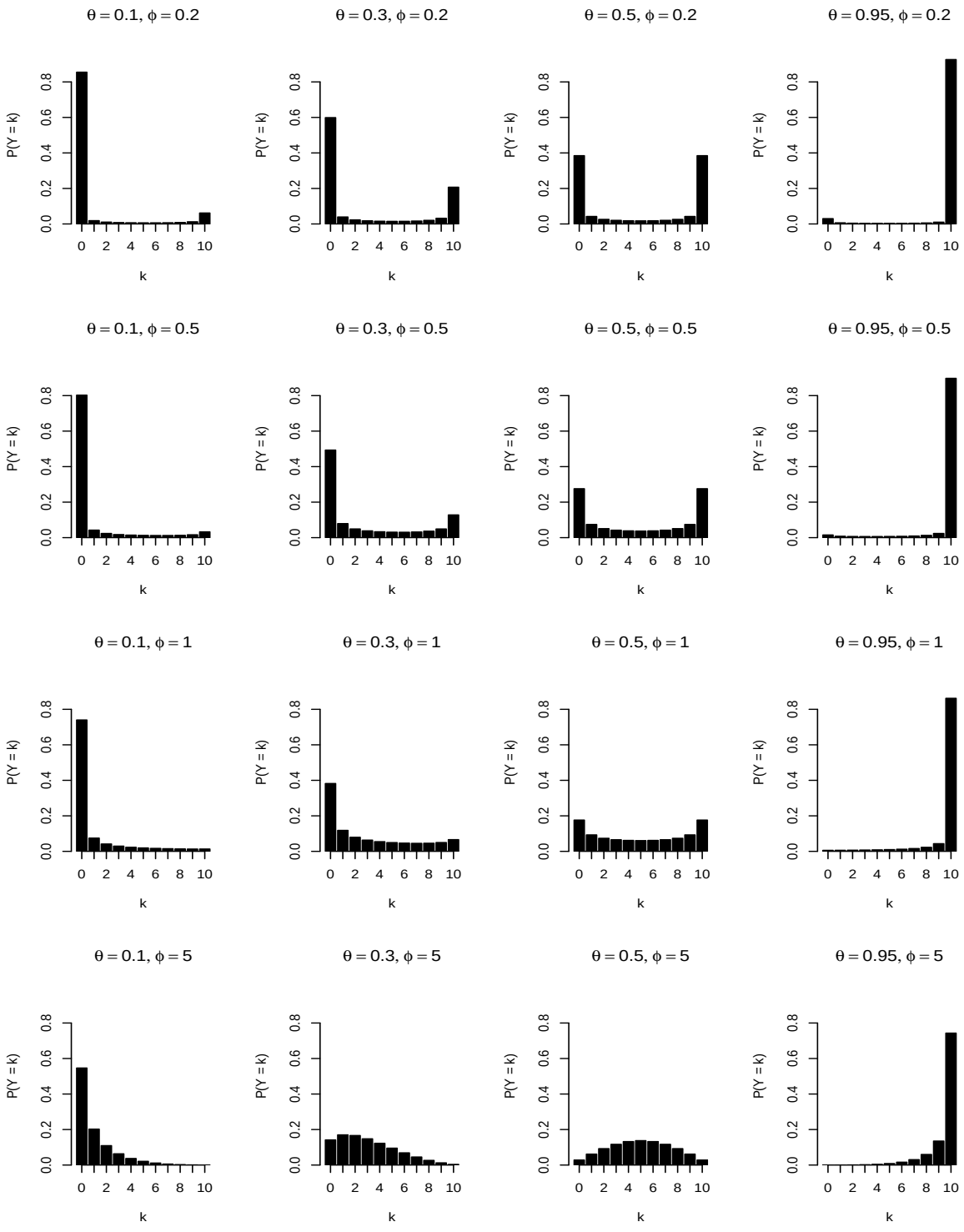

Figure 5.1: Examples of Beta-Binomial distributions when $S_t = 10$. Each row corresponds to a fixed value of the dispersion parameter ϕ; each column corresponds to a fixed value of the mean parameter θ.

Thus the following simple mechanism would give rise to a beta-binomial distribution for the number of wet sites in a fixed region sampled at S_t locations: a proportion θ_t of the region is wet, where θ_t is a beta-distributed random variable. Given θ_t, individual locations are wet or dry independently of each other.

Although this mechanism is clearly idealised (it is valid only when all probabilities are equal and when all points in space are independent given θ_t), it provides a useful insight into the model. In particular, it suggests that we may expect problems if sites are too close together (since in this case the assumption of conditional independence given θ_t is unlikely to hold even approximately).

This mechanism also enables us to explore the effect of increasing the number of sites. The proportion of the study area experiencing rain on day t can be approximated by Z_t/S_t. The approximation improves as S_t, the number of sites, tends to infinity. As $S_t \to \infty$ in step 2 above, Z_t/S_t converges to θ_t in probability; hence the distribution of Z_t/S_t tends to that of θ_t. Thus if we have a large number of sites, under our beta-binomial model we expect the proportion of wet sites to be distributed approximately as Beta(α_t, β_t).

5.2.5 The joint distribution of $\boldsymbol{Y}$ and Z

We have now specified a plausible model for the distribution of Z_t, the number of wet sites on day t. A natural strategy for simulation is to sample the number of wet sites from the distribution of Z_t, and then to allocate the positions of these wet sites. However, this will only yield sequences with the correct properties if the conditional probabilities of rain at each site, given the overall number of wet sites, are correctly specified. This can be achieved providing a valid joint distribution can be found for $\boldsymbol{Y}_t$ and Z_t. In this section, we present an algorithm for finding such a joint distribution (which may not be unique). The subscript t is now unnecessary, so we drop it and write $\boldsymbol{Y}, Z$. We assume in this section that the given distribution of Z is compatible with the individual probabilities of the Ys (this is not guaranteed, as we will see below).

From our earlier notation, we have $P(Y_s = 1) = p_s$. We also define $\pi_z = P(Z = z)$, and $w_{s,z} = P(Y_s = 1 \text{ and } Z = z)$. A first step in determining a joint distribution of $\boldsymbol{Y}$ and Z is to find $\{w_{s,z} : s = 1, \ldots, S; z = 0, \ldots, S\}$, from which we can calculate the conditional probabilities at each site:

$$P(Y_s = 1 | Z = z) = \frac{w_{s,z}}{\pi_z} \; . \tag{5.8}$$

The following relationships must hold:

$$0 \leq w_{s,z} \leq \pi_z \; ; \tag{5.9}$$

$$\sum_{z=0}^{S} w_{s,z} = p_s \; ; \quad \text{and} \tag{5.10}$$

$$\sum_{s=1}^{S} w_{s,z} = z\pi_z \; . \tag{5.11}$$

		s				
		1	2	$\cdots$	S	**TOTAL**
z	0	0	0	$\cdots$	0	0
	1	$w_{1,1}$	$w_{2,1}$	$\cdots$	$w_{S,1}$	π_1
	2	$w_{1,2}$	$w_{1,2}$	$\cdots$	$w_{S,2}$	$2\pi_2$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	S	π_S	π_S	$\cdots$	π_S	$S\pi_S$
TOTAL		p_1	p_2	$\cdots$	p_S	$\mathrm{E}(Z)$

Figure 5.2: Contingency table illustrating restrictions on the weights $\{w_{s,z} : s = 1, \ldots, S; z = 0, \ldots, S\}$.

The second of these is the Law of Total Probability; the third can be seen by noting that $\mathrm{E}\left[(\sum_s Y_s) | \, Z = z\right] = z$. But

$$\mathrm{E}\left[\left(\sum_s Y_s\right) \middle| \, Z = z\right] = \sum_s P\left(Y_s = 1 | Z = z\right) = \sum_s \frac{w_{s,z}}{\pi_z} \, ,$$

and the condition follows.

To visualise the problem, it is helpful to lay the ws out in the form of a contingency table, as in Figure 5.2. The only constraint which is not apparent from this is (5.9).

In principle, ws satisfying constraints (5.9)–(5.11) can be found using linear programming methods, since the first step in any linear programming problem is to find a feasible configuration i.e. one in which all the constraints are satisfied (see Press *et al.* (1992), Section 10.8 for example). However, with $S(S+1)$ values to allocate, a direct application of such techniques may be slow and hence unsuitable for use in simulation applications. We therefore develop our own algorithm, which is based on ideas of linear programming insofar as each step takes us closer to a feasible solution, but which makes use of the rather tight constraints to speed up the search. Additionally, our algorithm is insensitive to the order in which sites are considered (which would not be the case for a standard implementation of linear programming methods).

The basic procedure is to allocate a row of the table at a time, starting with $w_{s,0} = 0$ ($s = 1, \ldots, S$) and noting that $w_{s,S} = \pi_S$ ($s = 1, \ldots, S$) — both of which follow from constraints (5.9) and (5.11) above. As each row is allocated, the constraints on the remaining entities

will change. Specifically, suppose we have allocated rows $0, 1, \ldots, z-1$ and are currently considering row $z \leq S-1$. Constraints (5.9) and (5.11) above are unchanged. If we define $p_{s|z} = P(Y_s = 1 \text{ and } Z \geq z)$, then constraint (5.10) becomes

$$\sum_{j=z}^{S} w_{s,z} = p_{s|z} = p_s - \sum_{j=0}^{z-1} w_{s,z} \ . \tag{5.12}$$

Since we know that $w_{s,S} = \pi_S$, we must have $w_{s,z} \leq p_{s|z} - \pi_S$ for $z \leq S-1$. A further inequality can be deduced from constraint (5.9) applied to the subsequent rows of the table: $p_{s|z+1} = \sum_{j=z+1}^{S} w_{s,j} \leq \sum_{j=z+1}^{S} \pi_j \Rightarrow p_{s,z} - w_{s,z} \leq \sum_{j=z+1}^{S} \pi_j$, so that $w_{s,z} \geq p_{s|z} - \sum_{j=z+1}^{S} \pi_j$. Putting these inequalities together we must have, for $z \leq S-1$,

$$\max\left(0, p_{s|z} - \sum_{j=z+1}^{S} \pi_j\right) \leq w_{s,z} \leq \min\left(\pi_z, p_{s|z} - \pi_S\right) \ . \tag{5.13}$$

Writing the lower and upper bounds as $LB_{s,z}$ and $UB_{s,z}$ respectively, and summing, gives

$$\sum_{s=1}^{S} LB_{s,z} \leq \sum_{s=1}^{S} w_{s,z} \leq \sum_{s=1}^{S} UB_{s,z} \ . \tag{5.14}$$

Now from (5.11) above, we require $\sum_{s=1}^{S} w_{s,z} = z\pi_z$. Hence, providing $\sum_{s=1}^{S} LB_{s,z} \leq z\pi_z \leq \sum_{s=1}^{S} UB_{s,z}$, we can set

$$w_{s,z} = LB_{s,z} + \frac{z\pi_z - \sum_{s=1}^{S} LB_{s,z}}{\sum_{s=1}^{S} UB_{s,z} - \sum_{s=1}^{S} LB_{s,z}} \tag{5.15}$$

for each s, and proceed to the next row of the table. If the desired row total $z\pi_z$ is outside the interval $\left[\sum_{s=1}^{S} LB_{s,z}, \sum_{s=1}^{S} UB_{s,z}\right]$ then we must return to a previous row and re-allocate some of the joint probabilities. The following result is useful:

Result: providing the entries in rows 0 to $z-1$ of the table all satisfy inequalities of the form (5.13), the inequality $\sum_{s=1}^{S} UB_{s,z} \geq z\pi_z$ is automatically satisfied.

The proof is omitted, since it is not particularly instructive and the margin is too small to contain it. ■

The result tells us that in our algorithm, the only problem that can arise is when $\sum_{s=1}^{S} LB_{s,z} > z\pi_z$. Since $LB_{s,z} = \max\left(0, p_{s|z} - \sum_{j=z+1}^{S} \pi_j\right)$, the only way around this is to re-allocate some probabilities so as to reduce $p_{s|z}$ at sites where $p_{s|z} - \sum_{j=z+1}^{S} \pi_j > 0$ (because the πs are fixed), with a corresponding increase at sites where $p_{s|z} - \sum_{j=z+1}^{S} \pi_j < 0$. If this cannot be achieved, the given distribution of Z is incompatible with the marginal probabilities of the Ys.

We are now in a position to summarise the algorithm for calculating the ws. For each z:

1. Compute $p_{s|z} = p_s - \sum_{j=1}^{z-1} w_{s,j}$.

2. Compute $LB_{s,z} = \max\left(0, p_{s|z} - \sum_{j=z+1}^{S} \pi_j\right)$ and $UB_{s,z} = \min\left(\pi_z, p_{s|z} - \pi_S\right)$.

3. Compute $\sum_{s=1}^{S} LB_{s,z}$ and $\sum_{s=1}^{S} UB_{s,z}$.

4. If $\sum_{s=1}^{S} LB_{s,z} \leq z\pi_z$, calculate the ws for the current row according to (5.15). Otherwise, re-allocate some probabilities in previous rows of the table, if possible. For practical purposes, when re-allocating probabilities we try to avoid setting any value of $w_{s,z}$ to either 0 or π_z (except when $z = 0$ or S), since this can lead to problems in imputation (see Section 5.2.8 below) and is unrealistic.

The joint distribution of $\boldsymbol{Y}$ and Z is only partially specified by the ws in Figure 5.2, since the dependencies among the Ys are not represented. However, for simulation purposes it is not necessary to specify the distribution completely, as we illustrate in the next section.

5.2.6 Practical implementation

We are now in a position to simulate a dependent vector of binary random variables $\boldsymbol{Y}$, with a known distribution for their sum Z. The first step is to calculate the joint probabilities $\{w_{s,z}\}$ according to the algorithm in the previous section. It is necessary to compute the entire table regardless of the value of Z, since the initial allocation to any row of the table may need to be changed when subsequent rows are allocated.

The next step is to generate a value, z, from the distribution of Z, and to calculate the corresponding conditional probabilities at each site according to equation (5.8).

Finally, we take each site in turn and sample its value, then update the probabilities at the remaining sites to condition upon the sampled value. This updating can be done using the algorithm of the previous section. To see this, consider site 1. Define $\boldsymbol{Y}^* = (Y_2 \ \ldots \ Y_S)'$, and $Z^* = Z - Y_1$. Then Z^* is the sum of the elements of $\boldsymbol{Y}^*$. Given $Z = z$, Z^* takes values $z - 1$ and z with probabilities $P(Y_1 = 1|Z = z)$ and $1 - P(Y_1 = 1|Z = z)$ respectively. We therefore apply the algorithm of the previous section to $\boldsymbol{Y}^*$ and Z^*; then set $Y_1 = 1$ with probability $w_{1,z}/\pi_z$ to determine the value of Z^*, and move on to the next site. As before, the resulting conditional probabilities for $\boldsymbol{Y}^*$ are not unique.

It may appear from the above discussion that the computation time for this approach is prohibitive. In fact, this is not the case in our experience: simulation using this method is certainly slower than that using the 'hidden variables' approach, but it is still feasible to generate many long simulated sequences at several sites in a couple of hours, on a reasonable PC.

5.2.7 Dealing with incompatibility

The discussion so far has assumed that the distribution of Z is compatible with the marginal probabilities of the Ys. This is not guaranteed — obvious examples of incompatibility arise

when $P(Z = S) > \min_s P(Y_s = 1)$ and when $P(Z = 0) > \min_s P(Y_s = 0)$. When simulating long sequences of rainfall, the probability of rain at each site changes every day, and the beta-binomial distribution for Z is specified independently of these probabilities apart from ensuring the correct mean. It is therefore almost inevitable that at some stage we will encounter a situation where the specified probabilities are incompatible with the distribution of Z. If this occurs, to continue simulation we must either adjust the probabilities or the distribution of Z.

We find that in practice, incompatibility occurs when one or two of the individual ps are very different from the majority (e.g. at most sites, the probability of rain may be in the region of 0.8–0.9, but at a couple of sites it is 0.3). The reason for this seems to be the incorporation of previous days' rainfalls into the rainfall occurrence model — there can be substantial variability between sites here, resulting in different forecast probabilities of rain at each site. However, such differences in the probabilities of rain over a small area are probably unrealistic. Moreover, the objective of the procedure developed here is to reproduce a plausible distribution for the number of wet sites. We therefore propose to deal with incompatibility by modifying the ps rather than the distribution of Z. Specifically, we shrink the ps towards their mean i.e. for each s replace p_s with

$$p_s - \lambda (p_s - \theta) \ , \tag{5.16}$$

where $\lambda \in (0, 1)$, and θ is the mean of the ps (and the mean-value parameter of the beta-binomial distribution — see equation (5.6)). The expected number of wet sites is unchanged by this adjustment, which can be regarded as an attempt to robustify the forecasts made by the occurrence model. For small values of λ, the adjustment is a small one, in which case it may need to be repeated until a compatible set of probabilities is found. Repeated adjustments are guaranteed to find a compatible set of probabilities, since in the limit all of the probabilities are equal. In Section 5.2.4 we demonstrated that in this case, there is a mechanism which results in the beta-binomial distribution for Z; hence at least one joint distribution exists.

5.2.8 Imputation

Daily rainfall records often contain missing values. If we can determine the distribution of these missing values conditional upon the observed values at all sites, then we can simulate from this distribution many times to construct uncertainty envelopes for historical rainfall statistics. We refer to this process as *imputation* (rather than 'interpolation' as in our previous report).

Imputation of the wet-dry field on any particular day, using the beta-binomial model, is straightforward. Without loss of generality, we assume that the values of $Y_1, \ldots, Y_k$ are observed, and that $Y_{k+1}, \ldots, Y_S$ are missing. The starting point is the table of joint probabilities $\{w_{s,z}\}$. We work with the 'observed' sites one at a time, at each stage updating both the distribution of Z, and the probabilities of rain at the remaining sites, to condition on

the observations. For example, defining $Z^* = Z - Y_1$ as previously, we have for each z

$$\begin{aligned} P(Z^* = z|Y_1 = y_1) &= \frac{P(Z^* = z \text{ and } Y_1 = y_1)}{P(Y_1 = y_1)} = \frac{P(Z = z + y_1 \text{ and } Y_1 = y_1)}{P(Y_1 = y_1)} \\ &= \frac{P(Y_1 = y_1|Z = z + y_1)\, P(Z = z + y_1)}{P(Y_1 = y_1)} \ . \end{aligned} \tag{5.17}$$

All of the required probabilities here are either specified in advance, or can be obtained from the table of ws. Note that this step will fail if $P(Y_1 = y_1) = 0$. It is therefore important, when allocating the ws, to avoid situations where any conditional probabilities evaluate to 0 or 1 if possible (see comment on page 44).

To update the probabilities at sites $2, \ldots, S$ we use

$$P(Y_s = 1|Y_1 = y_1) = \sum_z P(Y_s = 1|Y_1 = y_1 \text{ and } Z^* = z)\, P(Z^* = z|Y_1 = y_1) \ . \tag{5.18}$$

The second term in the sum here is given by (5.17); the first can be calculated, for each value of z, using the method described in Section 5.2.6 above.

Having worked through all of the 'observed' sites in this way, we end up with a conditional distribution for the number of wet 'missing' sites. We sample from this conditional distribution, and then continue exactly as before.

Software, implementing all of the theory described in this section, has been written and is freely available from `http://www.ucl.ac.uk/~ucakarc/work/rain_glm.html`.

5.3 Performance evaluation

Generalised Linear Models for daily rainfall have been fitted to the Blackwater and North-East Lancashire regions using likelihood methods to decide which predictors to include. Daily rainfall data were available from a network of rain gauges for each region, as described in chapter 2 (gauge locations are shown in Figures 2.2 and 2.3). For both regions, a few gauges were excluded from model fitting so that these could be used to assess model performance. The fitted GLMs were used to simulate daily rainfall during the period 1st April 1975 to 31st March 1997, after a warm up period of one month using actual data, for the Blackwater region and during the period 1st October 1974 to 30th September 1998 for the North-East Lancashire region. For both regions, rainfall was simulated at all gauges operative for the entire simulation period (including those excluded from model fitting).

For both regions, 100 realisations of GLM simulated daily rainfall have been compared with historic data from the gauges operating during the period of simulation (with any missing values filled by the GLM, as described in Section 5.2.8) to assess the performance of the models. From the 100 simulated realisations the 5th, 25th, 50th, 75th and 95th percentiles have been calculated for statistics of the model generated rainfall, and these have been compared with historic values. Statistics are for daily rainfall and include mean,

standard deviation, proportion of wet days, mean and standard deviation given that it is a wet day, autocorrelations at lags 1–3 and annual maxima. For each region statistics have been calculated for 6 individual gauges (including 2 which were excluded during model fitting), the averaged rainfall from gauges within each 10km and 20km square, and the averaged rainfall for all gauges in the region operative throughout the simulation period. For each region the frequency distribution of the number of active gauges (those recording rain) is also constructed along with the percentiles for the simulated rainfall.

The following subsections document the fitting and simulation results for each region.

5.3.1 The Blackwater

Fitting

The occurrence (logistic) model has been fitted to the Blackwater region using 29 predictors. These include sine and cosine terms at the first two Fourier frequencies over the region to represent spatial site effects, whole and half year cycles to represent seasonality, temporal dependence, temporal persistence and interactions between seasonality and temporal dependence and persistence. An indicator for site B38 is also used as without it residuals for site B38 have non zero mean significantly larger than at other sites.

The amounts (gamma) model was fitted using 34 predictors again including sine and cosine terms at the first two Fourier frequencies over the region to represent spatial site effects, whole and half year cycles to represent seasonality, temporal dependence, temporal persistence and interactions between seasonality and temporal dependence and persistence. Again an indicator was used for site B38.

Extracts from files generated by the fitting programs containing residual information are provided in Appendix B.

Figures 5.3 and 5.4 show contours for fitted site effects (i.e. the contributions to equations (5.1) and (5.2)) for the Blackwater region for the occurrence and amounts models respectively. Gauges marked by red squares (numbers B7, B14, B28, B35 and B42) were not used to fit the models.

For both models the contours show an oscillatory nature of the fitted site effects which is due to the Fourier basis used to represent them. There are peaks and troughs in both surfaces, not supported by gauge data, which are artefacts of the Fourier basis. The Blackwater region is relatively flat and altitude is not a significant predictor for either model. With no other predictors the site effects have been represented purely as a function of location. For the logistic model, mean residuals at 22 out of 34 gauges (including 5 out of the 6 not used to fit the model) are non zero with a significance greater than 95 % (Appendix B1.1). It is possible that this measure of performance could be improved by using higher frequency terms in the Fourier representation. This was done however, using the first three Fourier frequencies and, although mean residuals improved at gauges used in fitting, the troughs and peaks of the surfaces between those gauges became more pronounced, indicating that the model had been

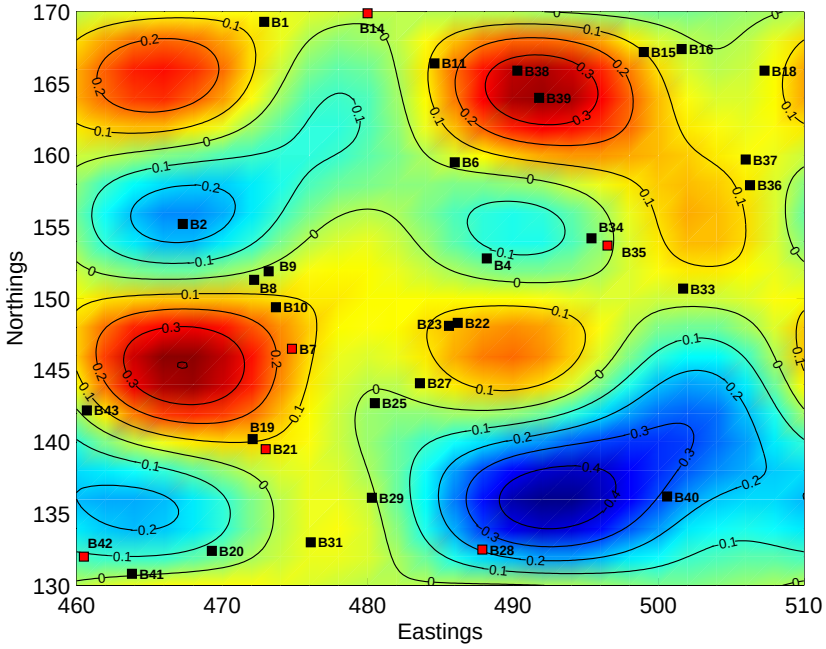

Figure 5.3: Contours of contributions made to log odds for rainfall occurrence by site effects for the logistic model fitted to the Blackwater region.

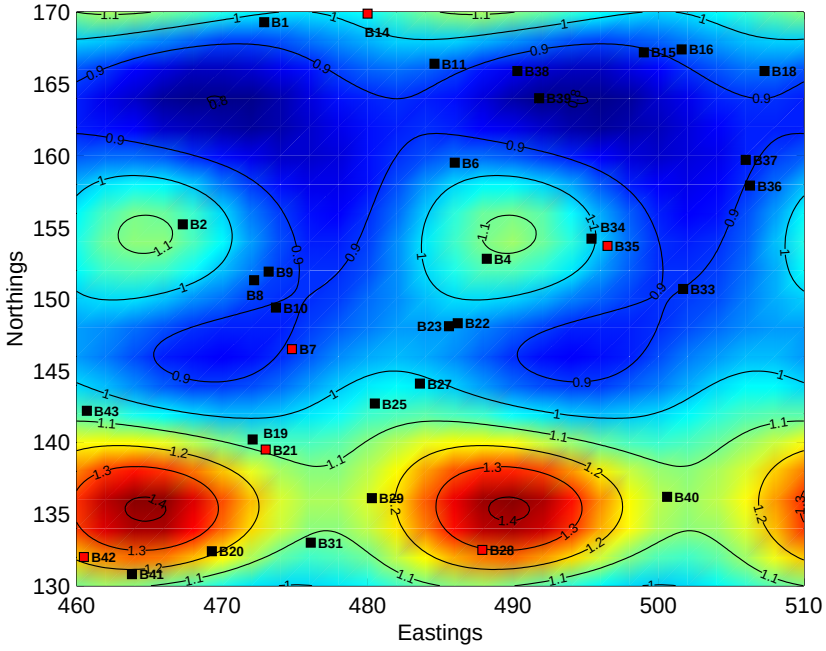

Figure 5.4: Contours showing contributions to multiplicative scaling factors applied to mean rainfalls made by site effects for the gamma model fitted to the Blackwater region.

overfitted — this would result in performance deterioration at locations away from gauges used to fit the models. The same applies to the gamma model.

Other measures of goodness of fit have been calculated, such as model performance by month and by year for the logistic and gamma models, and the forecasting performance of the logistic model. These are documented in Appendix B in files Blackwater logistic.res and Blackwater gammamdl.res. Most of the residuals show little structure; notice, however, that the mean annual residuals for the occurrence model (Appendix B1.1.1) are predominantly negative at the beginning of the record and predominantly positive thereafter. Moreover, the annual residuals for the amounts model (Appendix B1.2) are predominantly negative towards the end of the record. This indicates that there are trends in the area's rainfall patterns, which have not been accounted for by the models. Within the timescale of this phase of research, it has not been possible to investigate these in detail; however, we note that the presence of such trends may cause problems for the subsequent simulation exercise.

Simulation

The logistic and gamma models fitted to the Blackwater region have been used to simulate rainfall; properties of the simulated sequences have been compared with those observed, as described at the top of section 5.3. Statistics have been calculated for 6 gauges, including B07 and B35 which were excluded during model fitting, the averaged rainfall from gauges within each 10km and 20km square and the averaged rainfall for all gauges in the region operative throughout the simulation period. Plots showing historic statistics and percentiles for simulated statistics are in Appendix A1. Figure 5.5 shows gauges used for simulation, which are the ones recording during the simulation period. Again gauges marked by red squares are those excluded during fitting. Also shown are squares of size 10km x 10km and 20km x 20km, each enclosing gauges whose rainfall can be averaged to represent rainfall at that scale.

Plots of statistics for single sites (B02, B07, B27, B29, B34 and B35 — see Figures A.8–A.13 in Appendix B) show that for all sites there is a tendency for the simulated proportion of wet days to be lower than historically whilst the overall mean simulated rainfall for single sites is close to that historically. This is consistent with the simulated 'wet-day' mean daily rainfalls being higher than the historical ones. None of these results are surprising, given the residual analyses reported above — it was noted that residuals for the occurence model were mostly positive towards the end of the record (which is the time period used for simulation), indicating underprediction, whereas those for the amounts model were mostly negative indicating overprediction. It is expected that the inclusion of predictors representing these trends would resolve these problems.

For single sites, the standard deviation of daily rainfall for all days and for wet days and the first three autocorrelations are close to those of historic data. Simulation results for gauge B07, not used for model fitting, do not appear worse than typical results for other gauges. At gauge B35 (also excluded from model fitting), the simulated number of wet days is lower than the observed number by a larger amount than is typical here; however, results

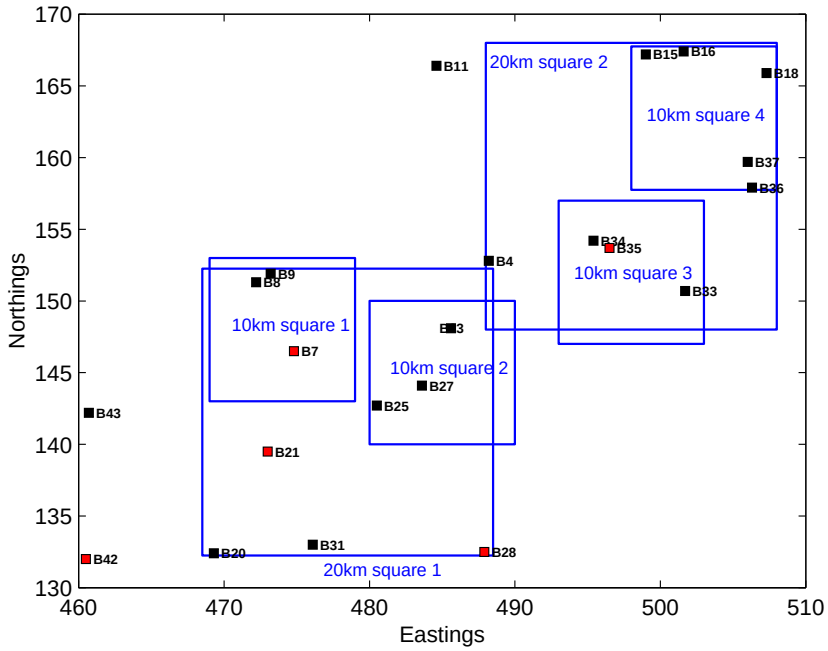

Figure 5.5: Blackwater region: gauges used for simulation. Red gauges are those not used in model fitting. 10km and 20km squares are also shown.

for nearby gauge B34 are only slightly better. A difference in mean residuals can be seen for gauges B22 and B23 which again are close (Appendix B1.1). This illustrates that data from neighbouring gauges may conflict in their statistical properties which may be influenced, for example, by gauge type, location and recording method. This data uncertainty will contribute to fitting difficulties; Chandler and Wheater (2001) report similar experiences in the context of another study. At larger scales (10km, 20km and whole region plots) the simulations' tendency to underestimate the number of wet days is reduced.

Appendix A also includes plots (Figures A.14–A.15) of historic annual maxima for single sites and 10km, 20km and whole region scales together with percentiles for simulated rainfall. At least visually, there appears to be reasonable agreement between historic and simulated annual maxima at all scales shown.

All of these simulation results have been obtained using the new method for treating spatial dependence (described in section 5.2 above). To investigate the performance of this new method, figure 5.6 shows the historic distribution of the number of wet sites during the simulation period together with percentiles for the simulated distribution. This indicates excellent agreement during all seasons, so that the new method enables results to be improved considerably over those reported in Wheater *et al.* (2000). It may be worth noting that where observed rainfall values are missing, they have been imputed using the model so that the picture presented here is perhaps over-optimistic; however, there are sufficiently few missing values that the effect of this is expected to be slight.

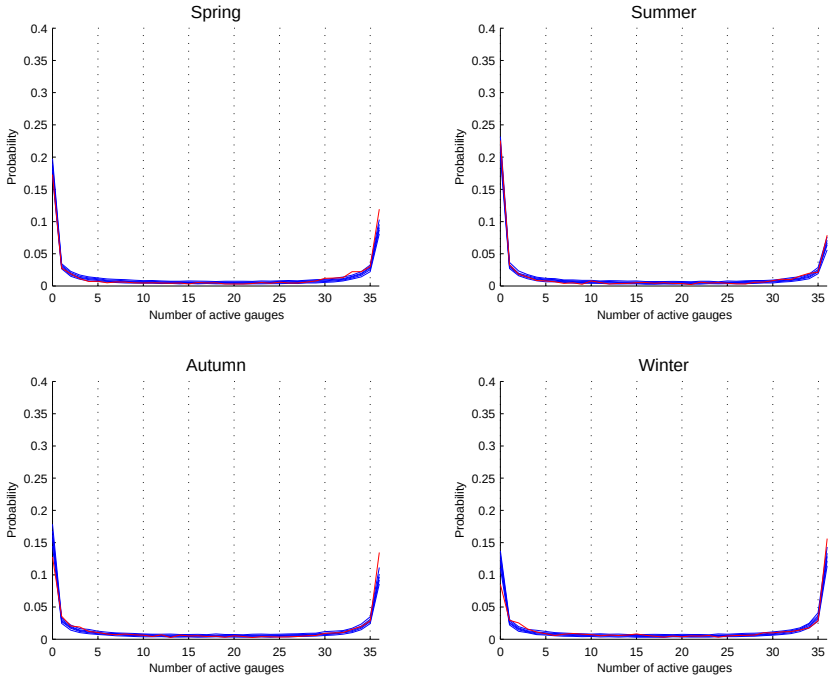

Figure 5.6: Blackwater region: Observed (red) and simulated (blue) distributions of numbers of wet sites.

5.3.2 North-East Lancashire

Fitting

The GLM model has been fitted to the NE Lancashire region and daily rainfall has been simulated for the period 1st October 1974 to 30th September 1998.

The fitted occurrence (logistic) model incorporates 37 predictors. These include site altitude and three Legendre polynomials of Eastings and Northings to represent spatial site effects, whole and half year cycles to represent seasonality, temporal dependence, temporal persistence and interactions between seasonality and temporal dependence and persistence.

The amounts (gamma) model contains 31 predictors, again including site altitude and three Legendre polynomials to represent spatial site effects, whole and half year cycles to represent seasonality, temporal dependence, temporal persistence and interactions between seasonality and temporal dependence and persistence.

Figures 5.7 and 5.8 show contours for fitted site effects, other than altitude, for the North East Lancashire region for the logistic and gamma models respectively. Gauges marked by red squares (numbers L10, L17 and L35) were not used to fit the models.

For both models the surfaces representing site effects have some steep gradients near the edges of the region away from gauges. These steep gradient are an artefact of the Legendre

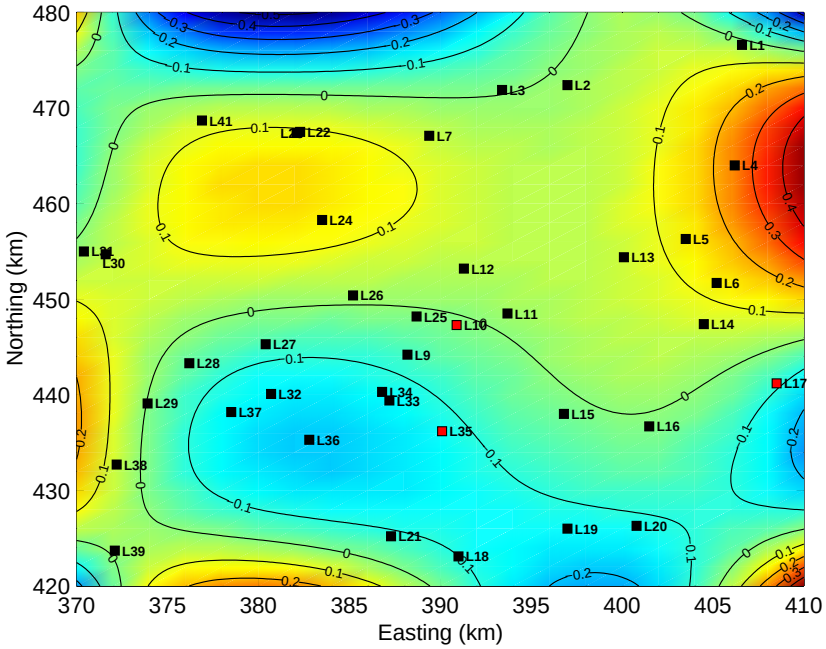

Figure 5.7: Contours of contributions made to log odds for rainfall occurrence by site effects for the logistic model fitted to the North-East Lancashire region.

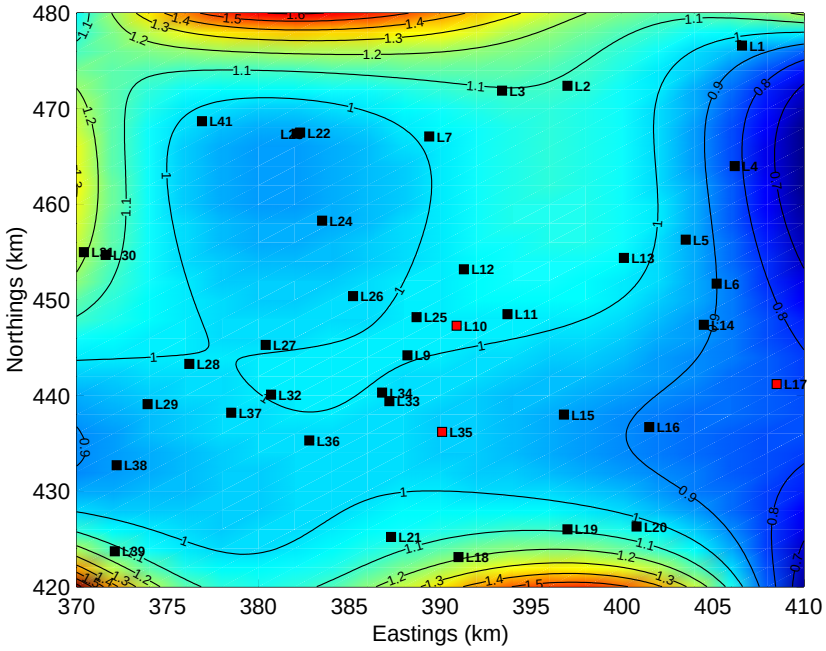

Figure 5.8: Contours showing contributions to multiplicative scaling factors applied to mean rainfalls made by site effects for the gamma model fitted to the North-East Lancashire region.

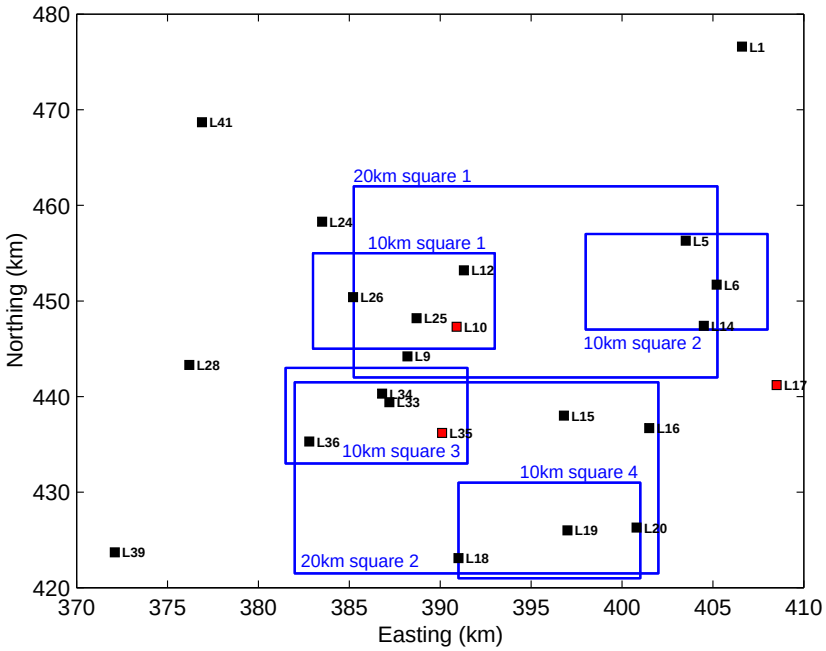

Figure 5.9: North-East Lancashire region: gauges used for simulation. Red gauges are those not used in model fitting. 10km and 20km squares are also shown.

polynomial representation of site effects and are not supported by gauge data. The surface gradient and mean residuals (Appendix B2) are high at Gauge L04 for both models, for example.

The model fitting programs generate residual analysis files (Appendix B). These show that model performance by site for the logistic and gamma models is worse than by month or year. As demonstrated for the Blackwater region, users should avoid overfitting of spatial dependence for both models as improvements in model peformance at the sites used in fitting may result in implausible surfaces for site effects. Another similarity with the Blackwater study is the presence of apparent trends in the annual residual series, which have not been captured by the models — again, these trends may be expected to cause some difficulties in simulation.

Simulation

As before, the logistic and gamma models fitted to the North East Lancashire region have been used to simulate rainfall as described above.

Figure 5.9 shows gauges used for simulation, which are the ones recording during the simulation period. Gauges L10, L17 and L35, marked by red squares, are those excluded during fitting. The 10km and 20km squares are also shown.

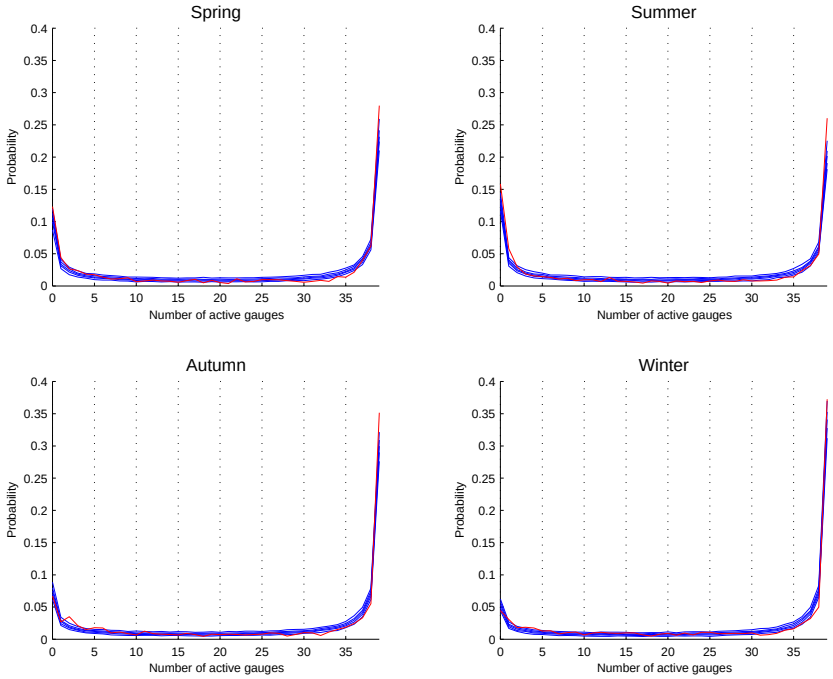

Figure 5.10: North-East Lancashire region: observed (red) and simulated (blue) distributions of numbers of wet sites.

Plots of simulated and observed rainfall statistics (Appendix A2, Figures A.16–A.30) show that at all scales (single sites, 10km square, 20 km square and whole region) there is a tendency, slightly enhanced at larger scales, for the simulations to overestimate historical mean daily rainfalls. The proportion of simulated wet days tends to be high at larger scales in summer. At all scales simulated annual maxima are slightly higher than historically. The performance at sites not used to fit the models does not appear worse than at those used for fitting.

Figure 5.10 shows the historic distribution of the number of wet sites during the simulation period together with percentiles for the simulated distribution. The historic distribution of the number of active (wet) sites is well reproduced by the logistic model fitted here.

5.4 Summary of chapter

The potential of GLMs for modelling and simulating daily rainfall sequences has been demonstrated elsewhere, prior to this report (Wheater *et al.* 2000; Chandler and Wheater 2001). The main aim of the work reported here has been to improve the representation of spatial dependence in rainfall occurrence, and to demonstrate the applicability of the method in two contrasting areas. The results indicate that the new dependence scheme for rainfall occur-

rence performs extremely well, in terms of reproducing the distribution of numbers of wet sites at different times of year and in different locations. For the models reported here, some discrepancies were observed between properties of observed and simulated rainfall sequences. However, the important point to note here is that these discrepancies can be anticipated by a careful analysis of residuals obtained while fitting models — in this case, they are mainly due to trends that are not represented in the models. The inclusion of trends is straightforward within the GLM framework (see Chandler and Wheater (2001), for example); the only reason for not including them here is the limited timescale of the project. On the basis of previous experience with a variety of datasets, we are confident that the discrepancies between observed and simulated properties can be resolved almost entirely by accounting for these trends. However, it is worth noting that the GLM approach does need to be applied with care to ensure that results will be realistic.

Chapter 6

Summary and conclusions

6.1 Summary

In previous work, a range of modelling approaches for continuous simulation of spatial rainfall was developed and examined. A model for simulation of rainfall in continuous space and time was developed and tested using radar rainfall data, with promising performance, and is recommended for medium-term application. However, further development is required to represent spatial and temporal non-stationarity, and the full model can only be identified from radar data. Even relatively dense raingauge networks do not adequately define the full clustered structure of rain cells or storm velocity.

Also in this previous work, alternative approaches were developed, to be compatible with currently-available raingauge data. Generalized Linear Models (GLMs) were shown to be a powerful tool for the representation of daily rainfall in space and time, and have the capability to represent both spatial and temporal non-stationarity. They can thus be used for analysis and simulation of, for example, topographic and other location effects on rainfall distribution and climate variability. However, for flood application, sub-daily spatial fields are needed. Hence the possibility of a combined methodology was considered, and a modelling framework for downscaling, or spatial-temporal disaggregation, was defined.

In this report, further work has been done to develop the combined methodology, and the multi-site disaggregation procedure is defined fully in Chapter 3. It is envisaged that model application would typically be based on data from a network of daily raingauges, perhaps with one or two sub-daily gauges (however, the procedure can, in principle, be based on simulation if necessary). The GLM can be fitted using the daily data, and used to interpolate missing data, define daily rainfall at additional locations, or extend the data series by simulation. Generation of long time-series is readily achieved. A simple multivariate model of hourly rainfall is used as the basis of the disaggregation procedure. This preserves the properties of the hourly spatial structure and is transformed to assimilate the daily values. A particular development has been the development of a scaling relationship, which holds for all intergauge distances, and relates the hourly to the daily cross-correlations. This

can be used when the hourly cross-correlation structure is not known.

Performance of the disaggregation scheme was tested for the Brue catchment, in south-west England, using the dense raingauge network established by the NERC HYREX progamme, and for two contrasting areas of the Blackwater (south-east England) and north-east Lancashire (Chapter 4). For the Brue, it was possible to test the scheme for the ideal situation in which the hourly cross-correlation structure is known. A particular issue investigated was the effect of the threshold of 0.2mm associated with the finite size of raingauge tipping buckets. For comparability, this was also applied to the simulation results. Results were generally good across a wide range of performance measures, although hourly cross-correlations for summer were less well represented than for winter. The results were then compared with the more realistic situation in which it was assumed that only two sub-daily gauges were available. A scaling analysis was carried out, using the two gauges only, to support the disaggregation. Results were most encouraging, with very similar performance to the situation in which the full hourly cross-correlation structure was known.

Results for the Blackwater and north-east Lancashire were also based on analysis of just two sub-daily gauges, and tested using a third sub-daily gauge. These comparisons were therefore much more limited due to the lack of sub-daily observation sites. Results were less good than for the Brue, as could be expected considering the greater separation of the gauges, but were plausible and encouraging.

Previous work with the GLM had shown some theoretical limitations in the representation of spatial dependence of rainfall (in particular difficulties in representing the spatial occurrence of wide-spread rainfall), and hence new theoretical developments were investigated and implemented, to overcome this (Chapter 5). A new dependence scheme was developed, based on use of the Beta-Binomial family of distributions to represent rainfall occurrence. Results from modelling the Blackwater and north-east Lancashire are reported in detail. Comparison of the observed and simulated distributions of numbers of wet sites were excellent for both, and it is concluded that the previously-identified limitations have been satisfactorily resolved.

6.2 Conclusions

The results above present two important methodological developments to resolve issues associated with the use of generally-available raingauge data to support continuous simulation modelling of spatial rainfall at hourly resolution. Within the constraints of a 9 month study, it was not possible to go further, but next steps obviously should include a) study of the integrated performance of the GLM and disaggregation procedure, b)the use of single-site rainfall models in conjunction with the GLM for long-term simulation, and c) joint performance of rainfall and rainfall-runoff models, including, in particular, analysis of extreme value performance.

Bibliography

Chandler, R.E. and Wheater, H.S. (2001). Climate change detection using Generalized Linear Models for rainfall — a case study from the West of Ireland. *Water Resources Research*. Submitted.

Coe, R. and Stern, R.D. (1982). Fitting models to daily rainfall. *J. Appl. Meteorol.*, **21**, 1024–31.

Hurrell, J.W. (1995). Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, **269**, 676–9.

Koutsoyiannis, D. (1994). A stochastic disaggregation method for design storm and flood synthesis. *Journal of Hydrology*, **156**, 193–225.

Koutsoyiannis, D. (1999). Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology. *Water Resources Research*, **35, no.4**, 1219–29.

Koutsoyiannis, D. (2001). Coupling stochastic models of different time scales. *Water Resources Research*, **37**, 379–92.

Koutsoyiannis, D. and Manetas, A. (1996). Simple disaggregation by accurate adjusting procedures. *Water Resources Research*, **32, no. 7**, 2105–17.

Northrop, P.J. (1998). A clustered spatial-temporal model of rainfall. *Proc. Roy. Soc. London.*, **A454**, 1875–88.

Oman, S.D. and Zucker, D.M. (2001). Modelling and generating correlated binary variables. *Biometrika*, **88, No.1**, 287–90.

Onof, C., Chandler, R.E., Kakou, A., Northrop, P., Wheater, H.S., and Isham, V. (2000). Rainfall modelling using Poisson-cluster processes: a review of developments. *Stoch. Env. Res. & Risk Ass.*, **14**, 384–411.

OPW (1998). *An investigation of the flooding problems in the Gort-Ardrahan area of South Galway, by Southern Water Global and Jennings O'Donovan and partners.* Office of Public Works, Dublin.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical recipes in FORTRAN (second edition).* Cambridge University Press.

Rodriguez-Iturbe, I., Cox, D.R., and Isham, V. (1987). Some models for rainfall based on stochastic point processes. *Proc. R. Soc. Lond.*, **A410**, 269–88.

Rodriguez-Iturbe, I., Cox, D.R., and Isham, V. (1988). A point process model for rainfall: further developments. *Proc. R. Soc. Lond.*, **A417**, 283–98.

Stern, R.D. and Coe, R. (1984). A model fitting analysis of rainfall data (with discussion). *J. Roy. Stat. Soc.*, **A147**, 1–34.

Wheater, H.S., Isham, V.S., Onof, C., Chandler, R.E., Northrop, P.J., Guiblin, P., Bate, S.M., Cox, D.R., and Koutsoyiannis, D. (2000). Generation of spatially consistent rainfall data. Report to the Ministry of Agriculture, Fisheries and Food (2 volumes). Also available as Research Report no. 204, Department of Statistical Science, University College London (http://www.ucl.ac.uk/Stats/research/abstracts.html).

Wilby, R.L. and Wigley, T.M.L. (1997). Downscaling general circulation output: a review of methods and limitations. *Prog. Phys. Geog.*, **21**, 530–48.

Yang, C. (2001). *Observed changes and simulative predictions of climate extremes in China.* PhD thesis, Institute of Atmospheric Physics, Beijing. In Chinese.

Appendix A

Figures for GLM results in Section 5.3

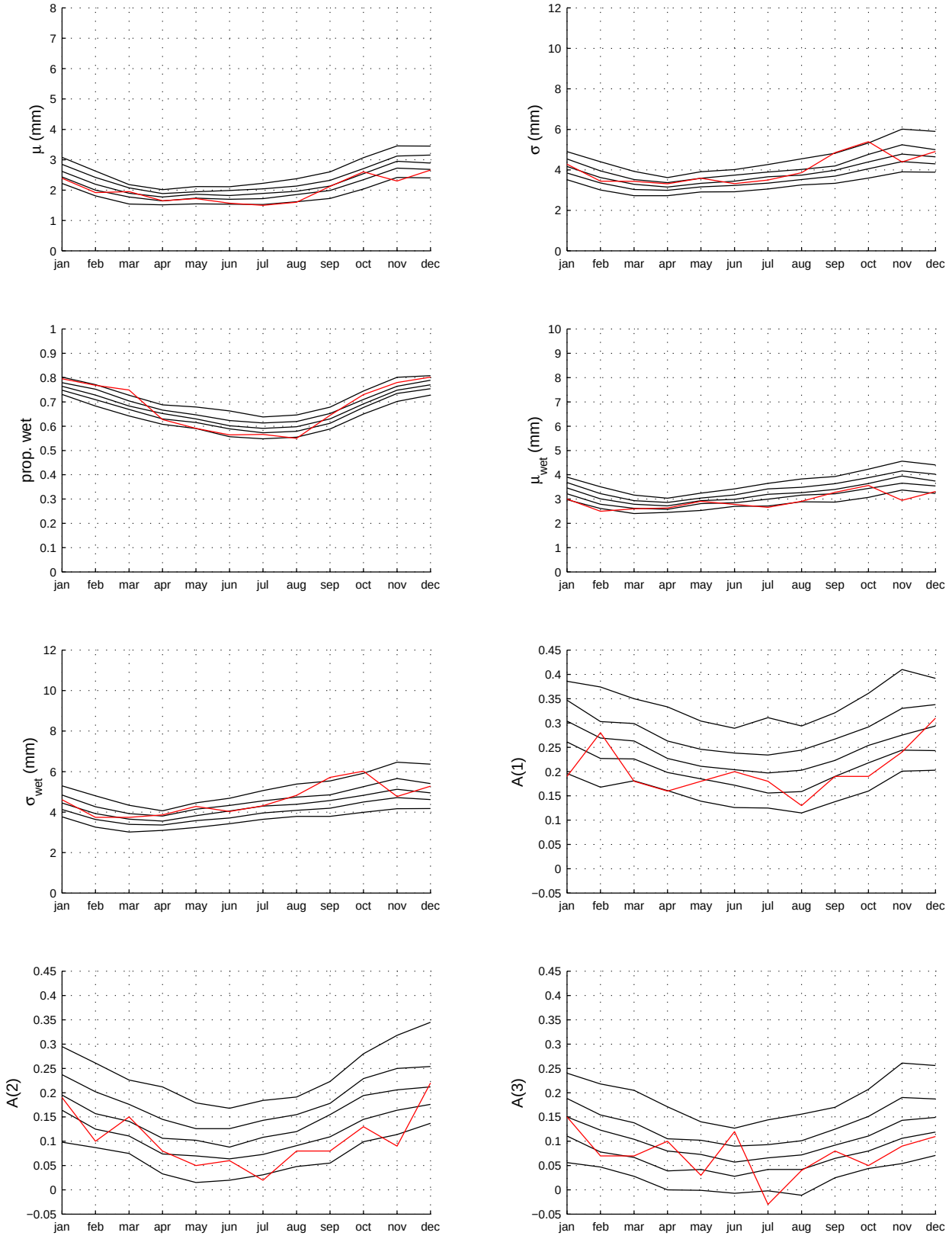

Figure A.1: *Blackwater whole region:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

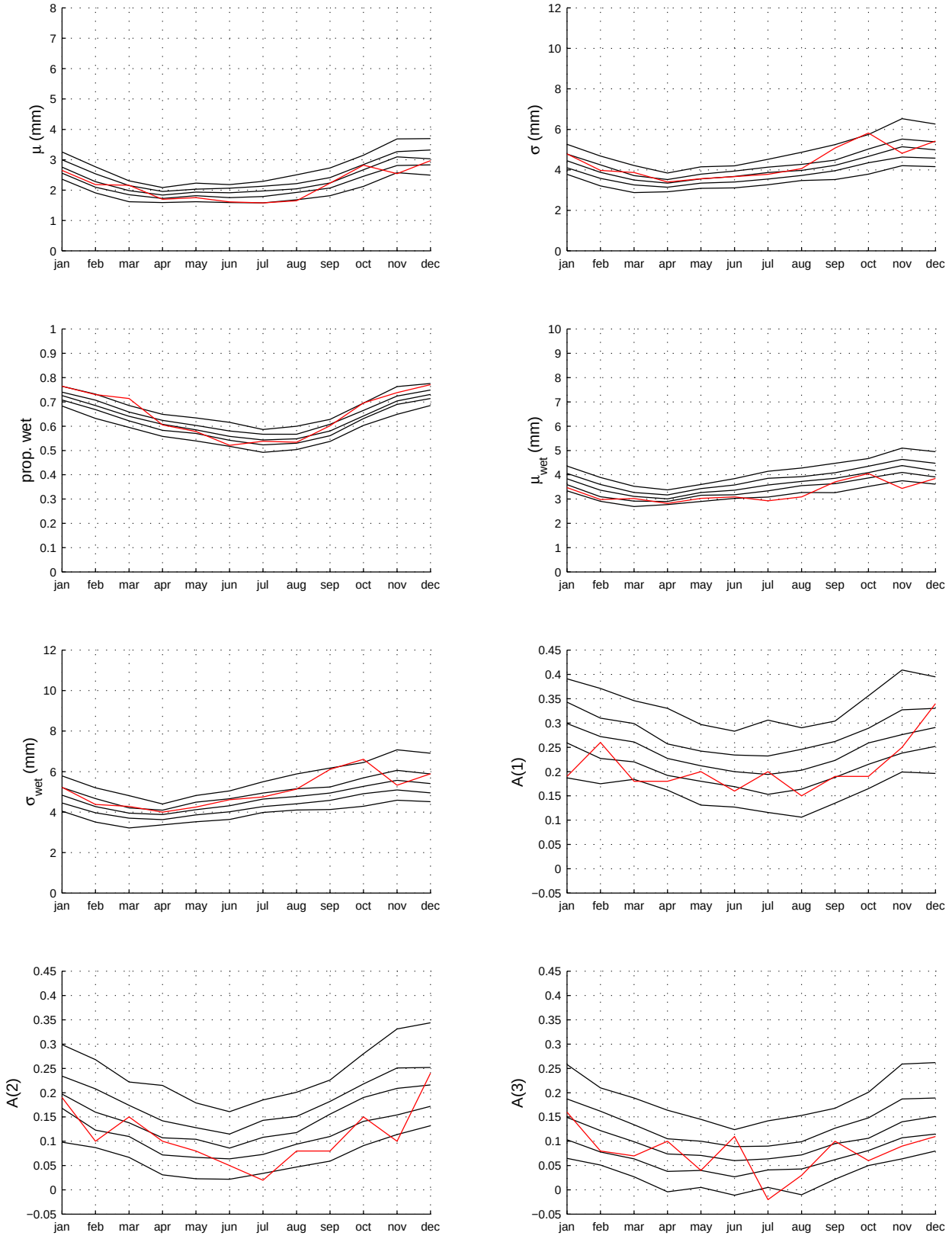

Figure A.2: *Blackwater 20km square 1:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

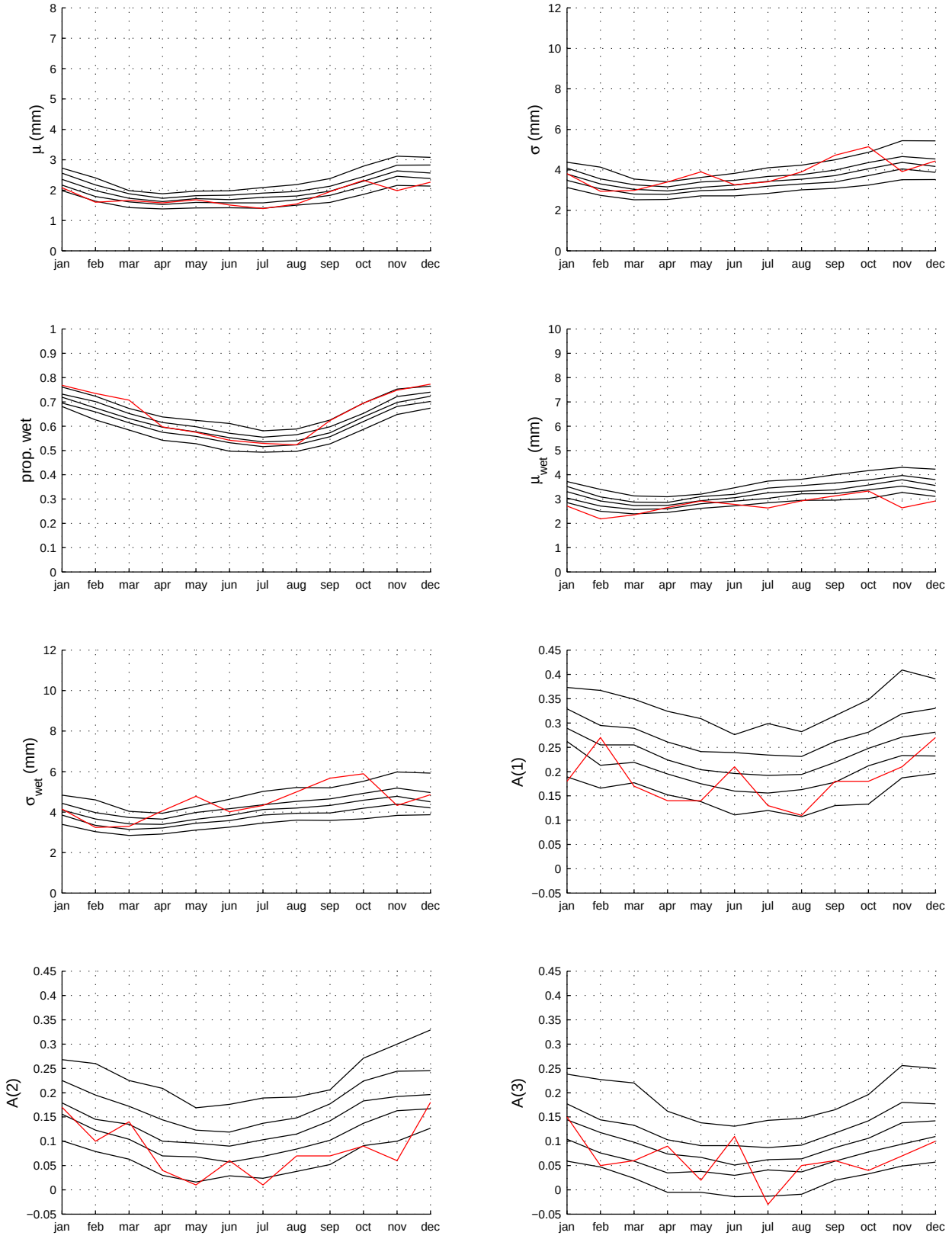

Figure A.3: *Blackwater 20km square 2:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

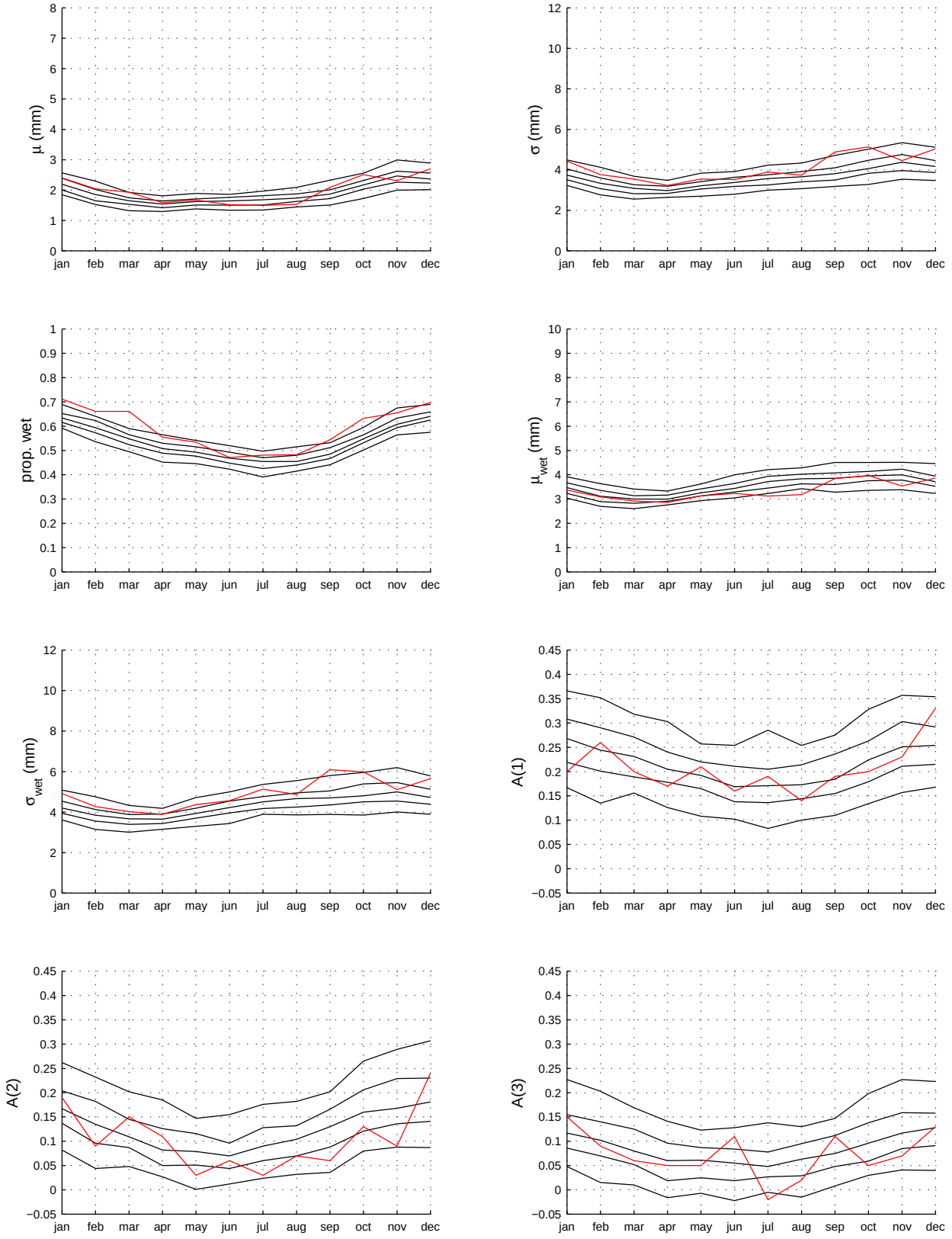

Figure A.4: *Blackwater 10km square 1:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

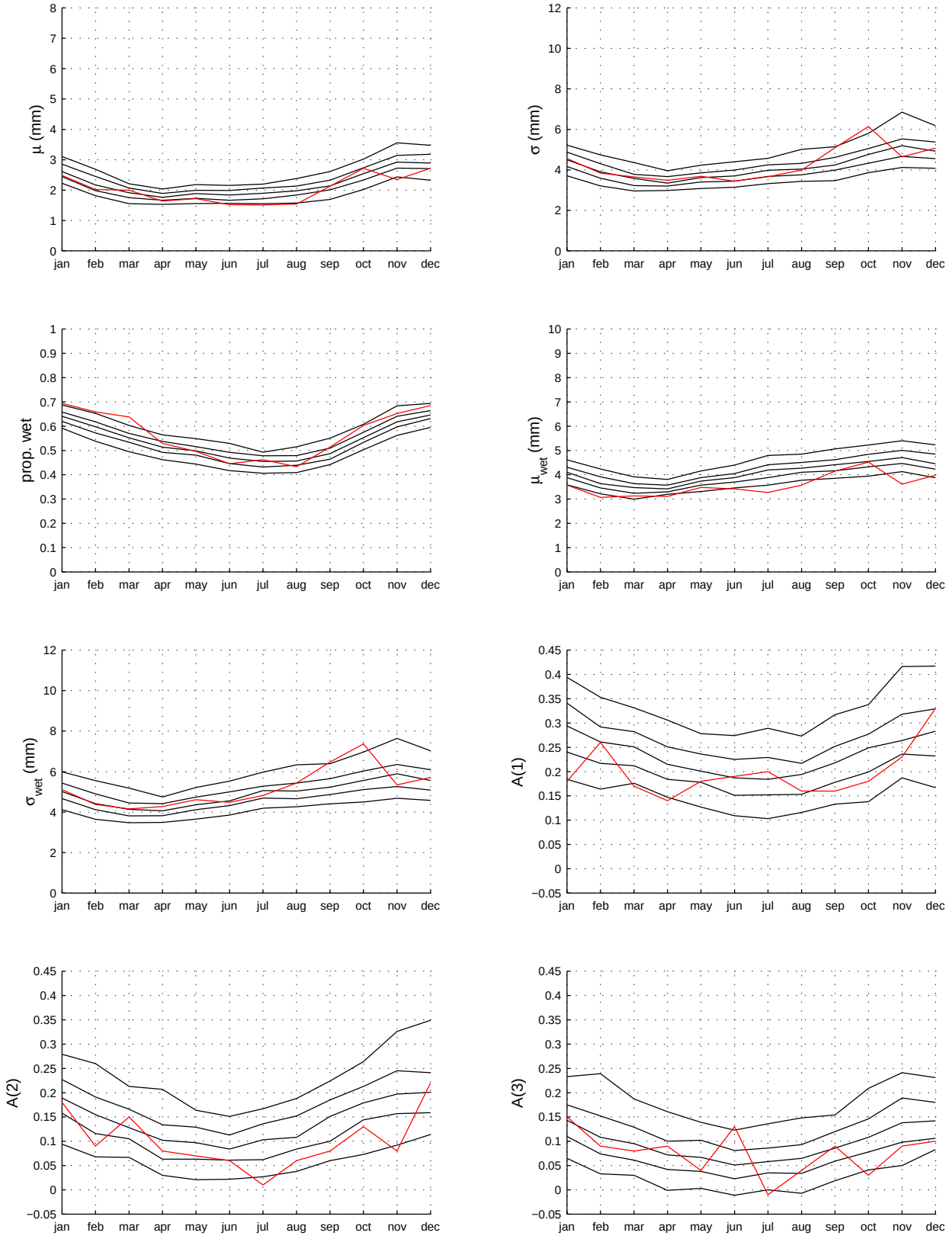

Figure A.5: *Blackwater 10km square 2:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

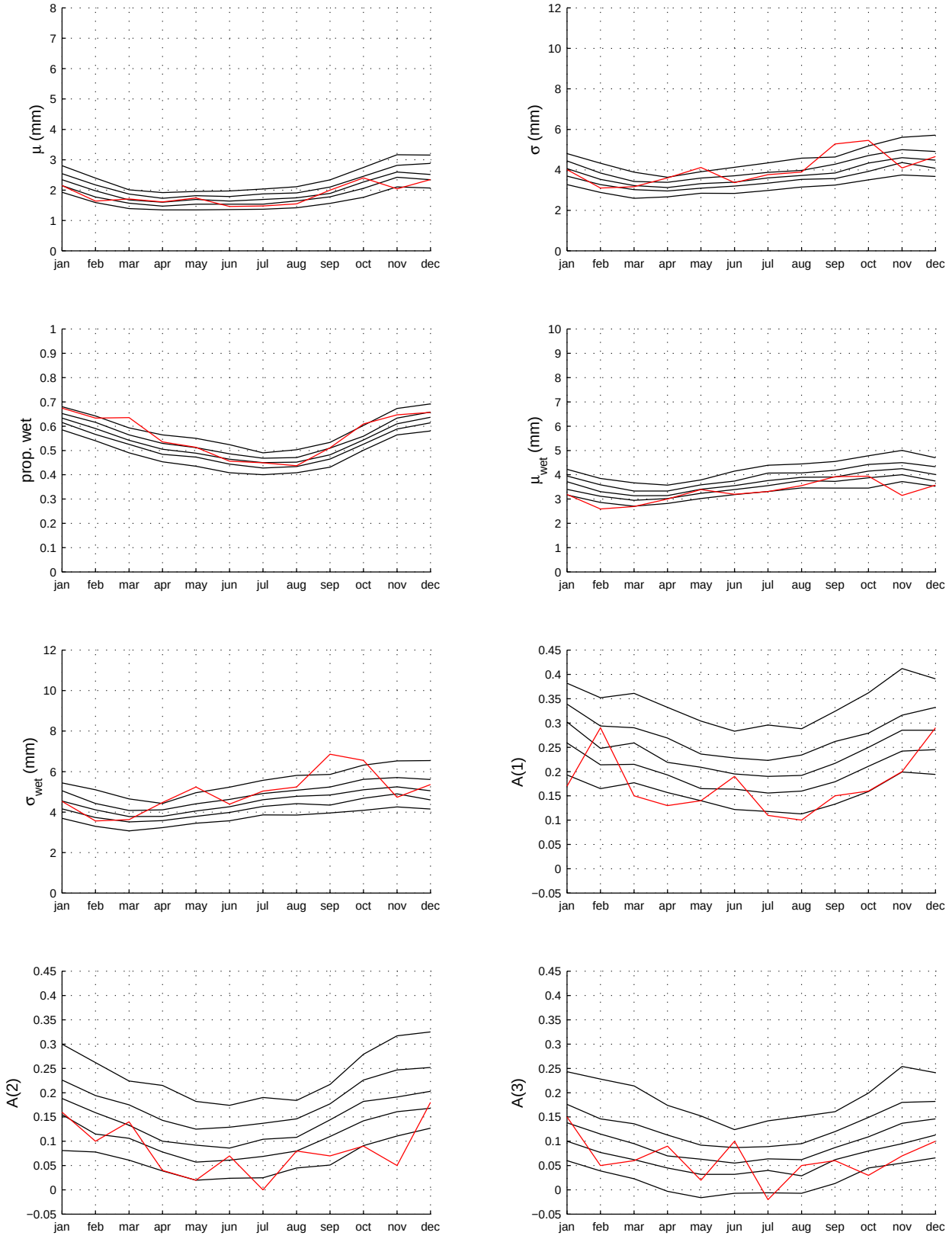

Figure A.6: *Blackwater 10km square 3:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

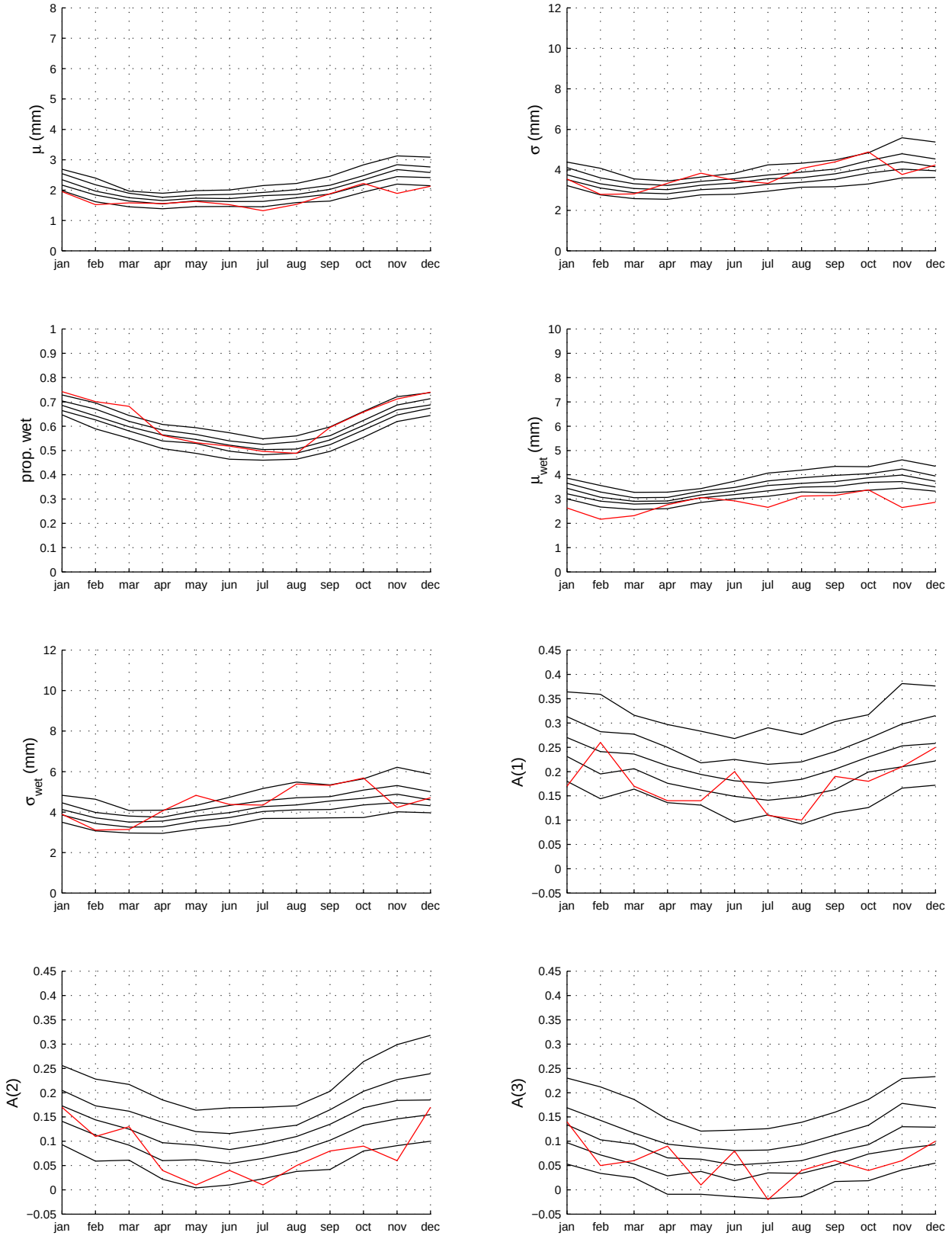

Figure A.7: *Blackwater 10km square 4:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

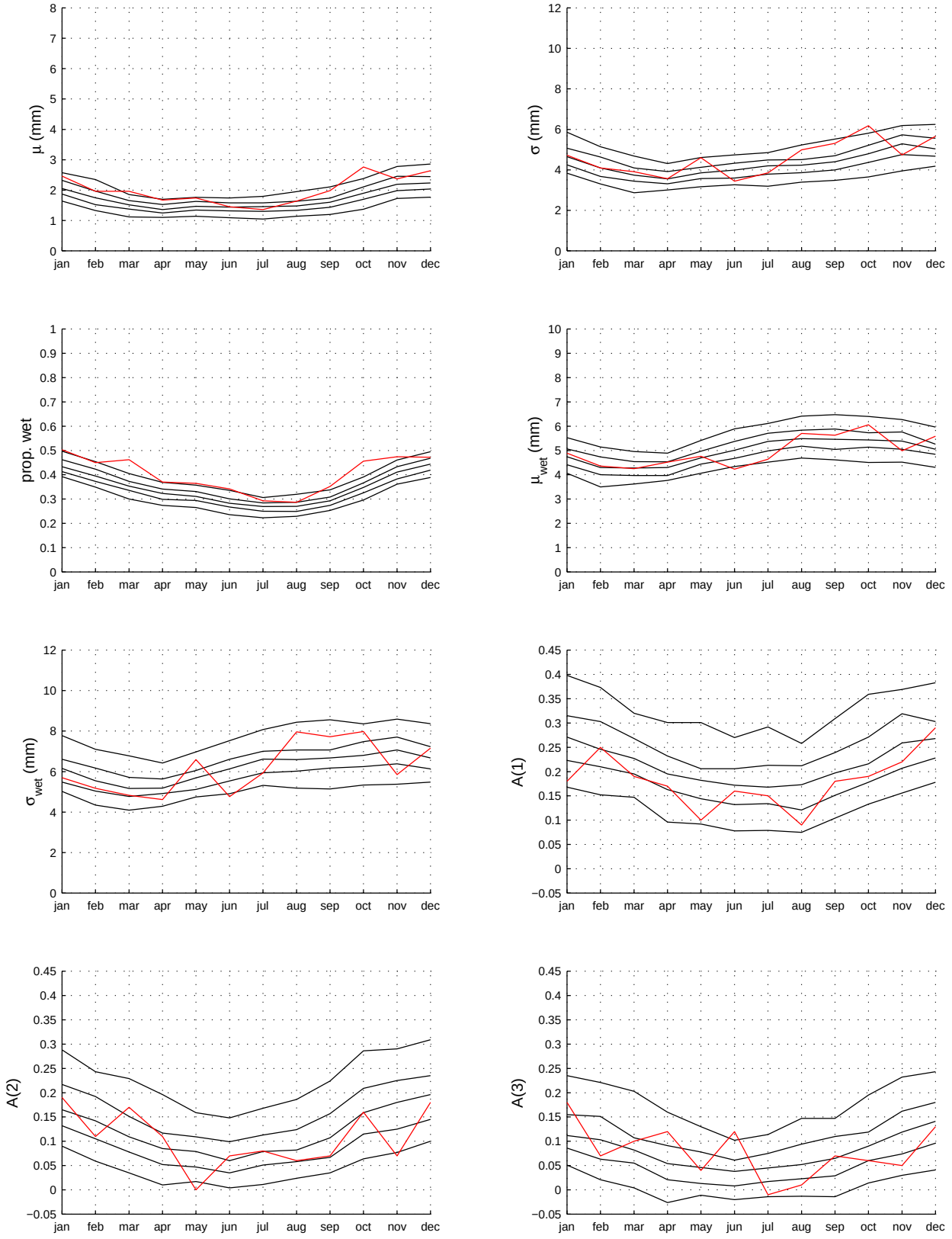

Figure A.8: *Blackwater gauge B02:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

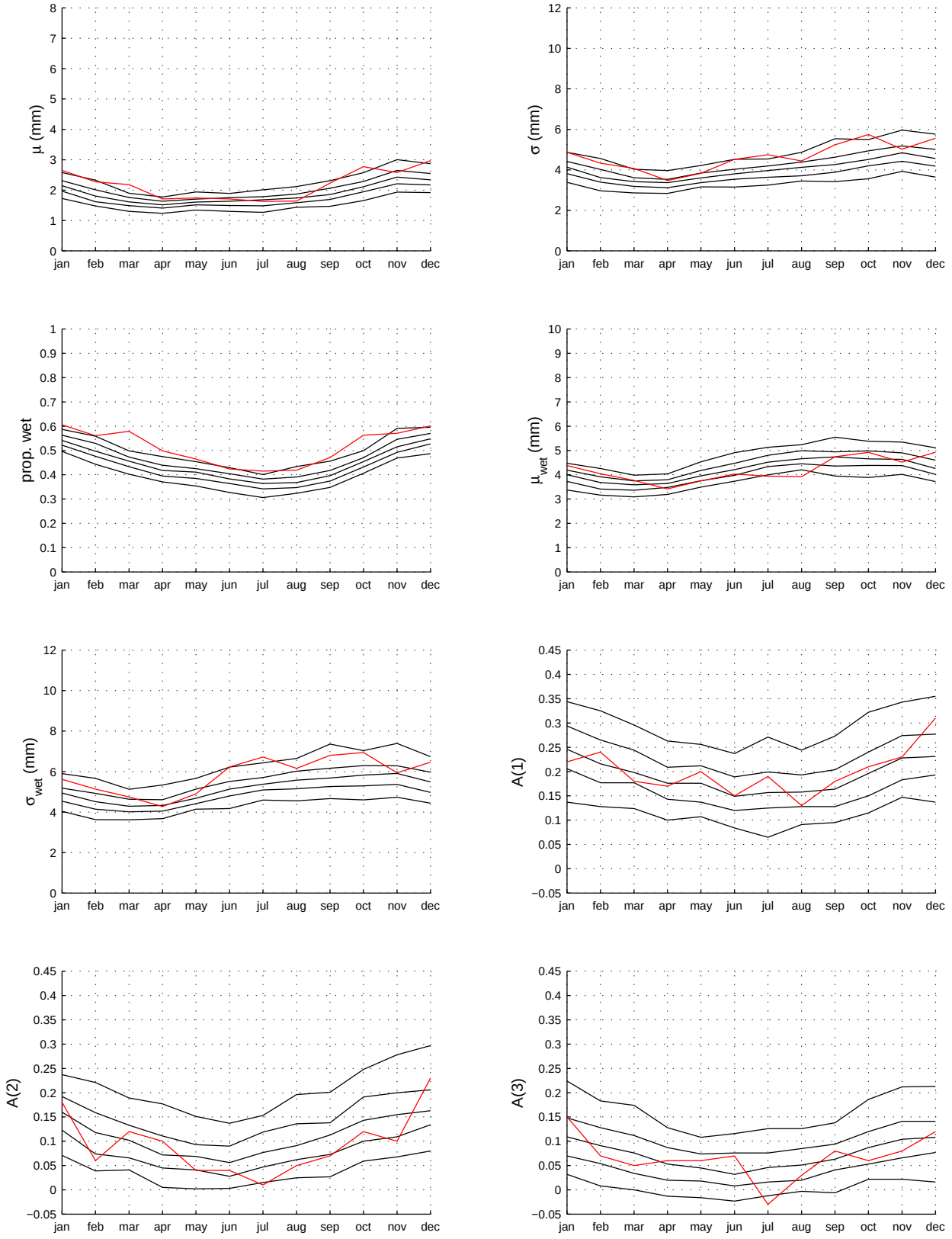

Figure A.9: *Blackwater gauge B07:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

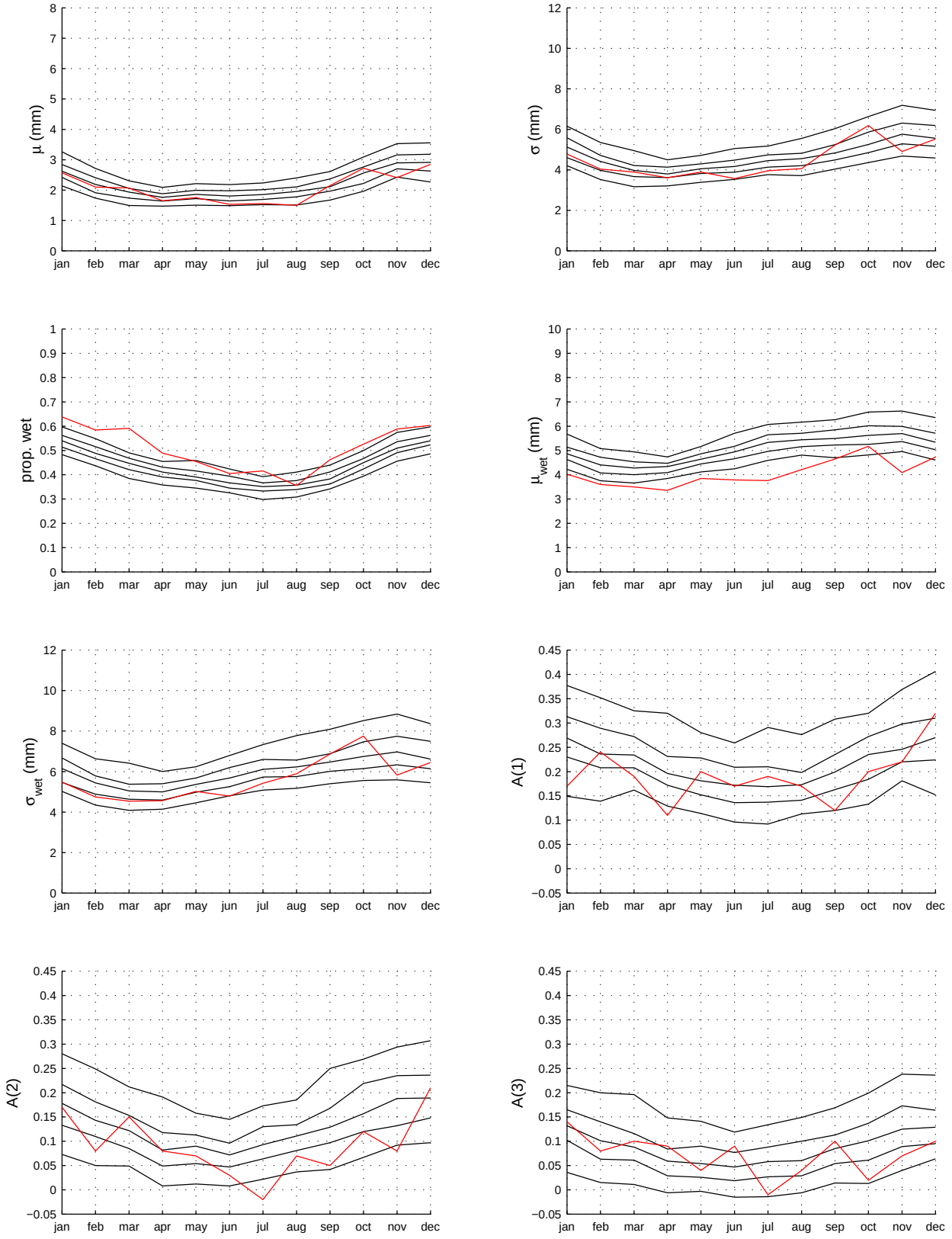

Figure A.10: *Blackwater gauge B27:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

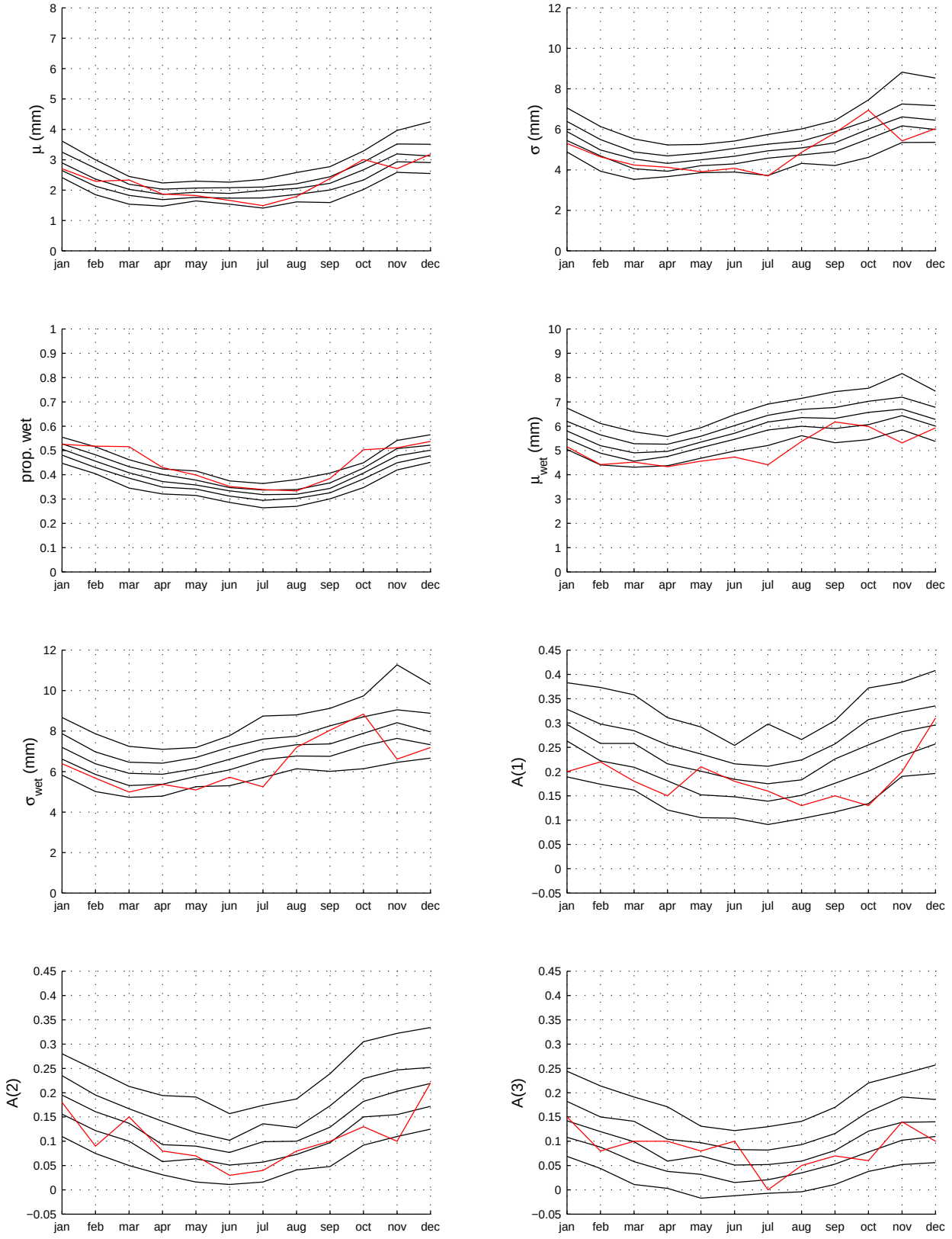

Figure A.11: *Blackwater gauge B29:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

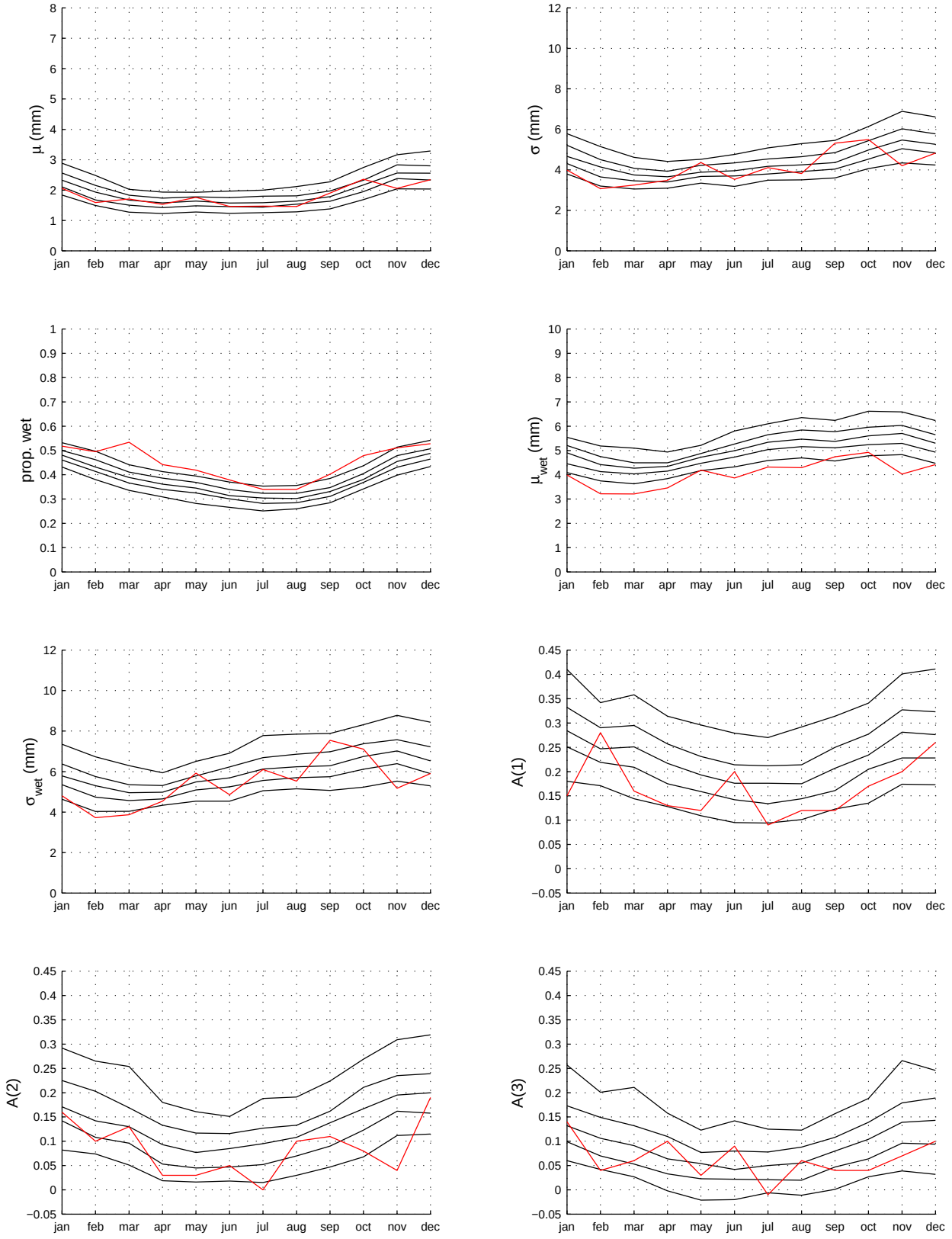

Figure A.12: *Blackwater gauge B34:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

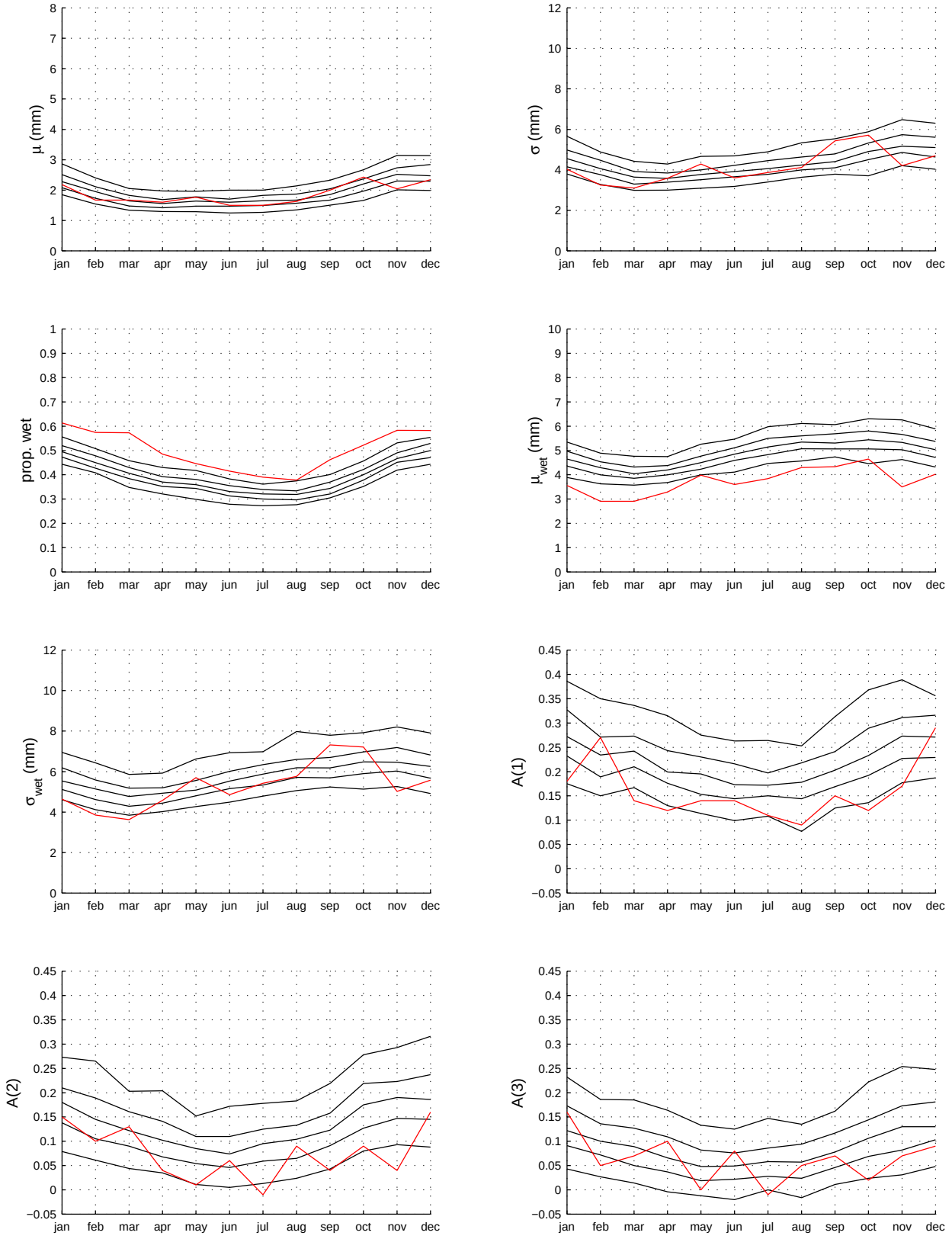

Figure A.13: *Blackwater gauge B35:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

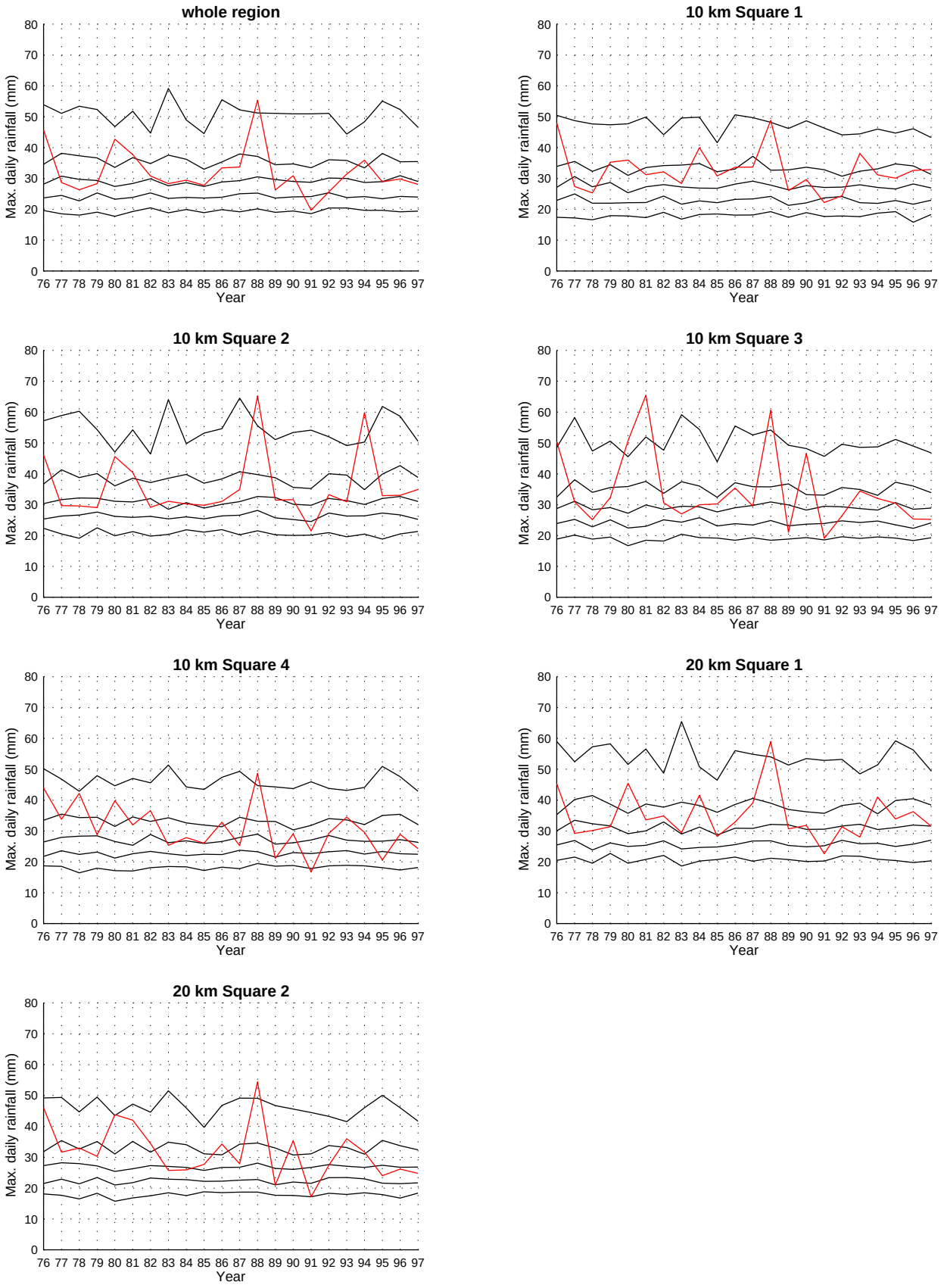

Figure A.14: *Blackwater whole region and 20km and 10km squares:* Historic annual maxima and percentiles (5, 25, 50, 75 and 95) for simulated annual maxima.

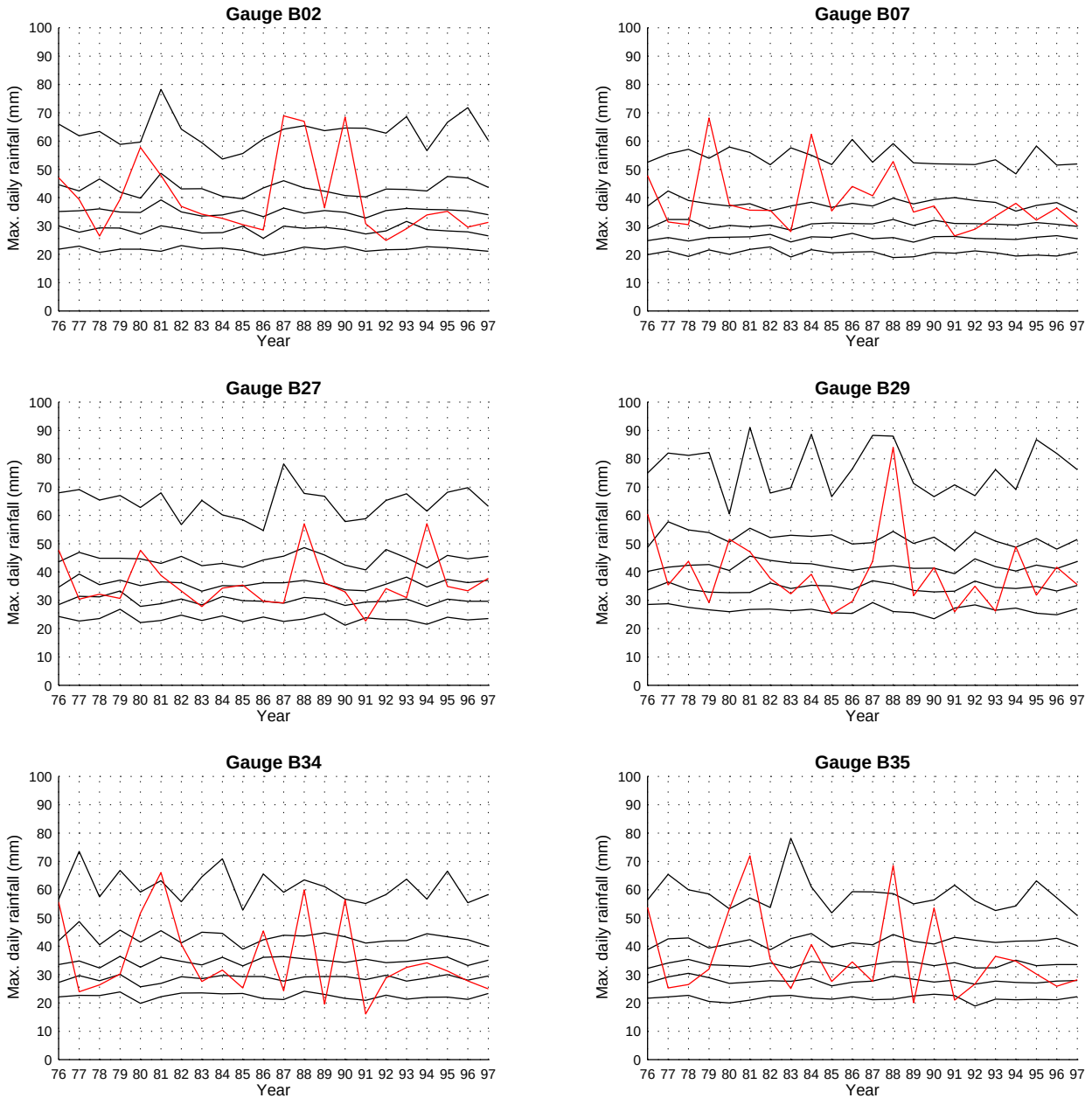

Figure A.15: *Blackwater gauges:* Historic annual maxima and percentiles (5, 25, 50, 75 and 95) for simulated annual maxima.

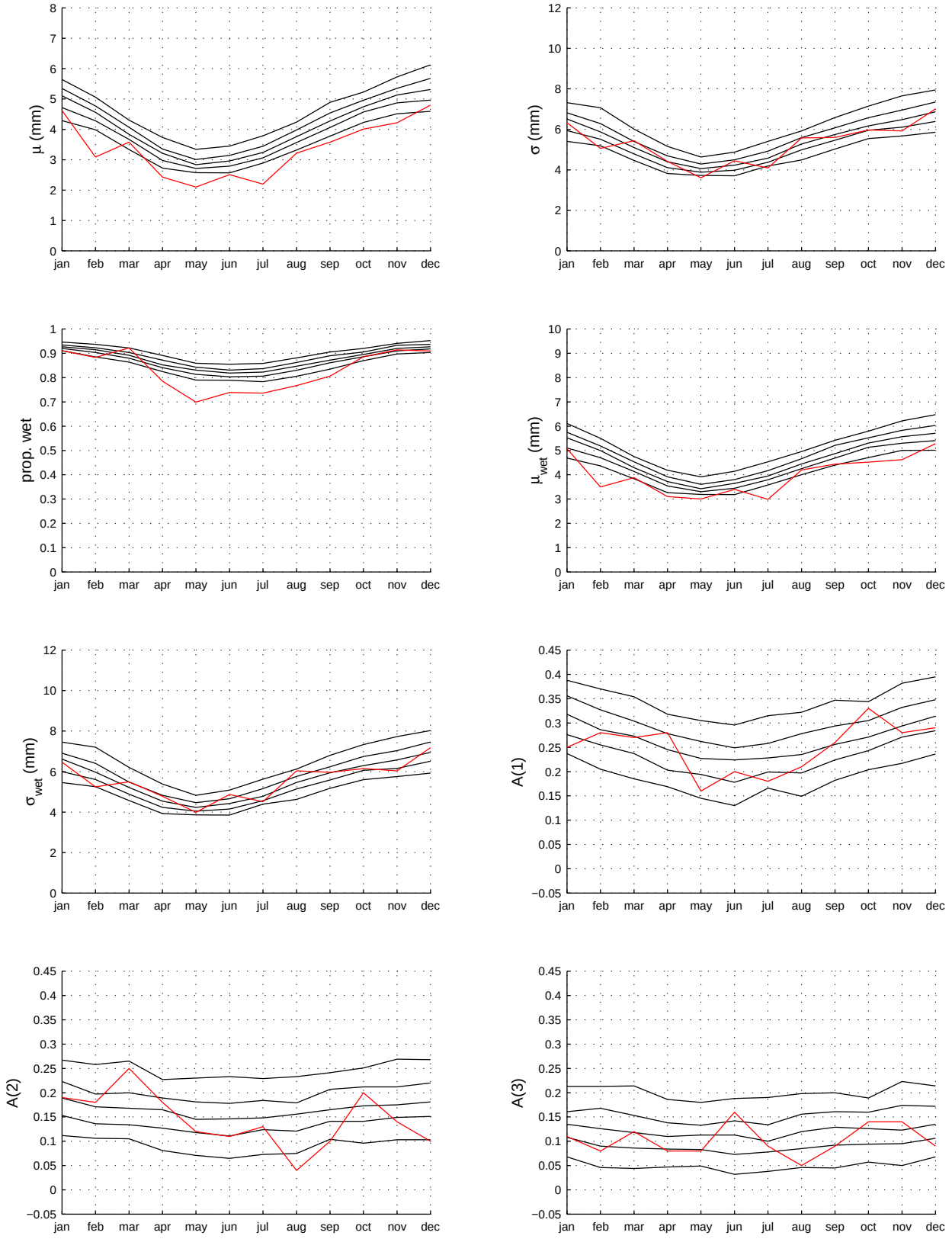

Figure A.16: *North East Lancashire whole region:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

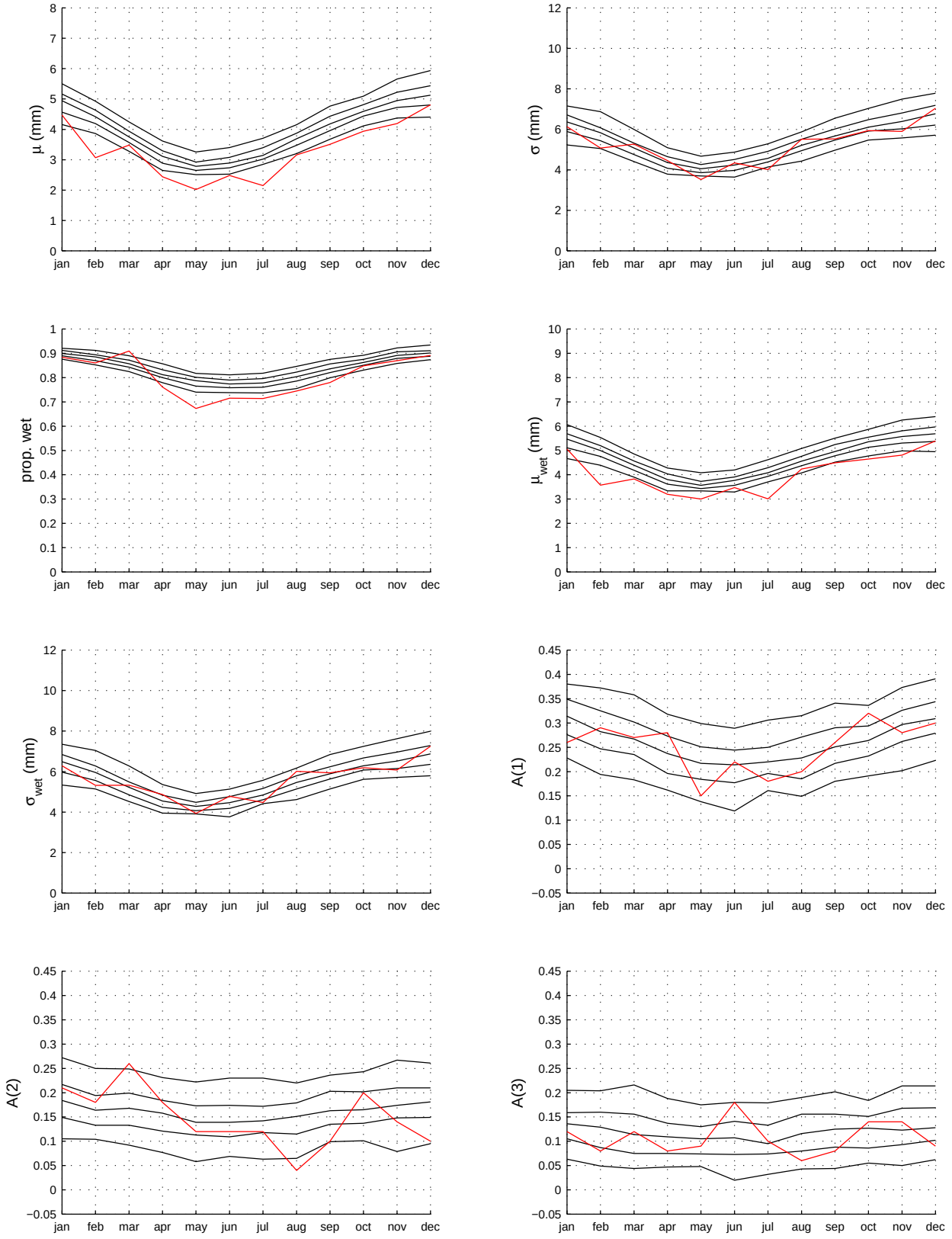

Figure A.17: *North East Lancashire 20km square 1:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

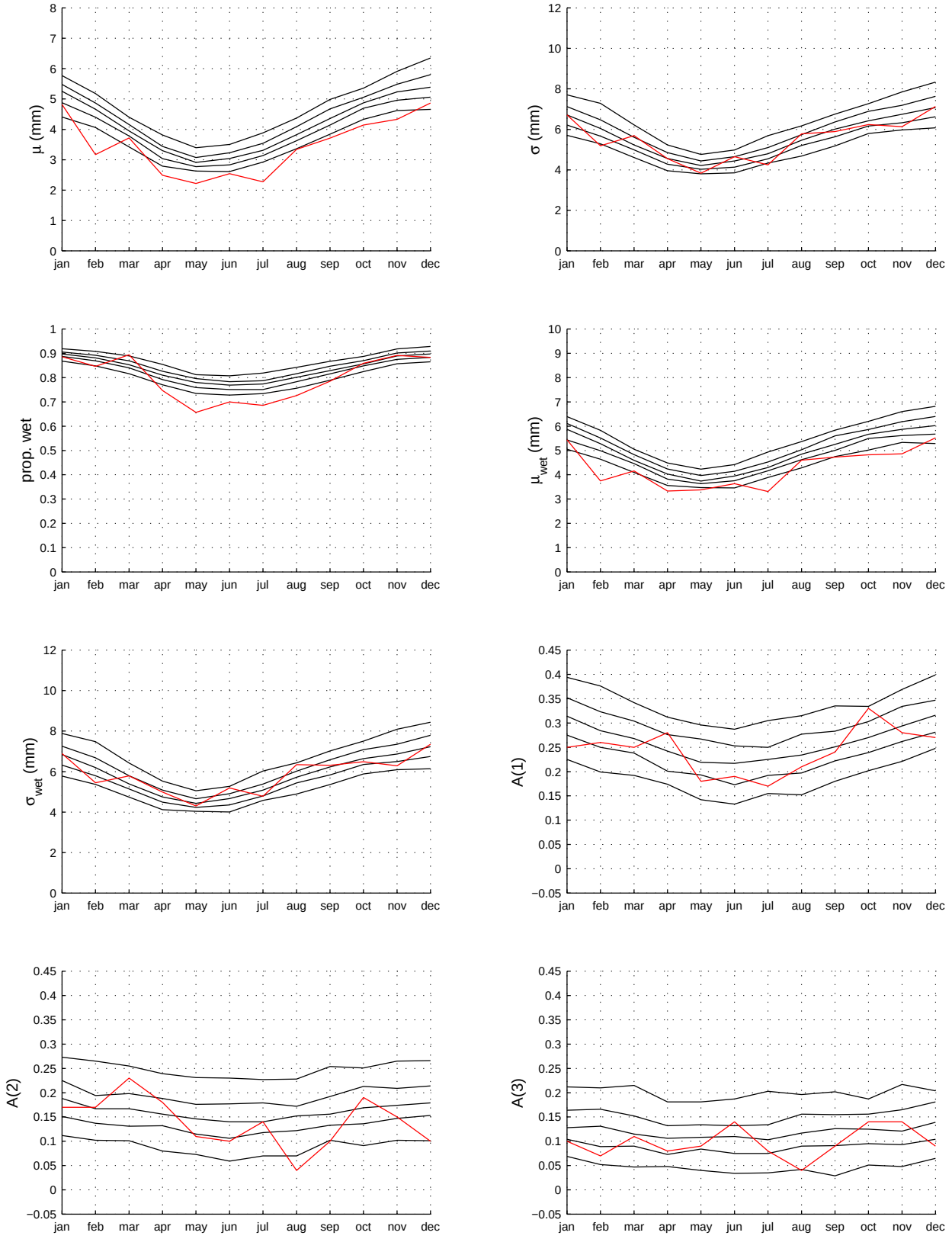

Figure A.18: *North East Lancashire 20km square 2:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

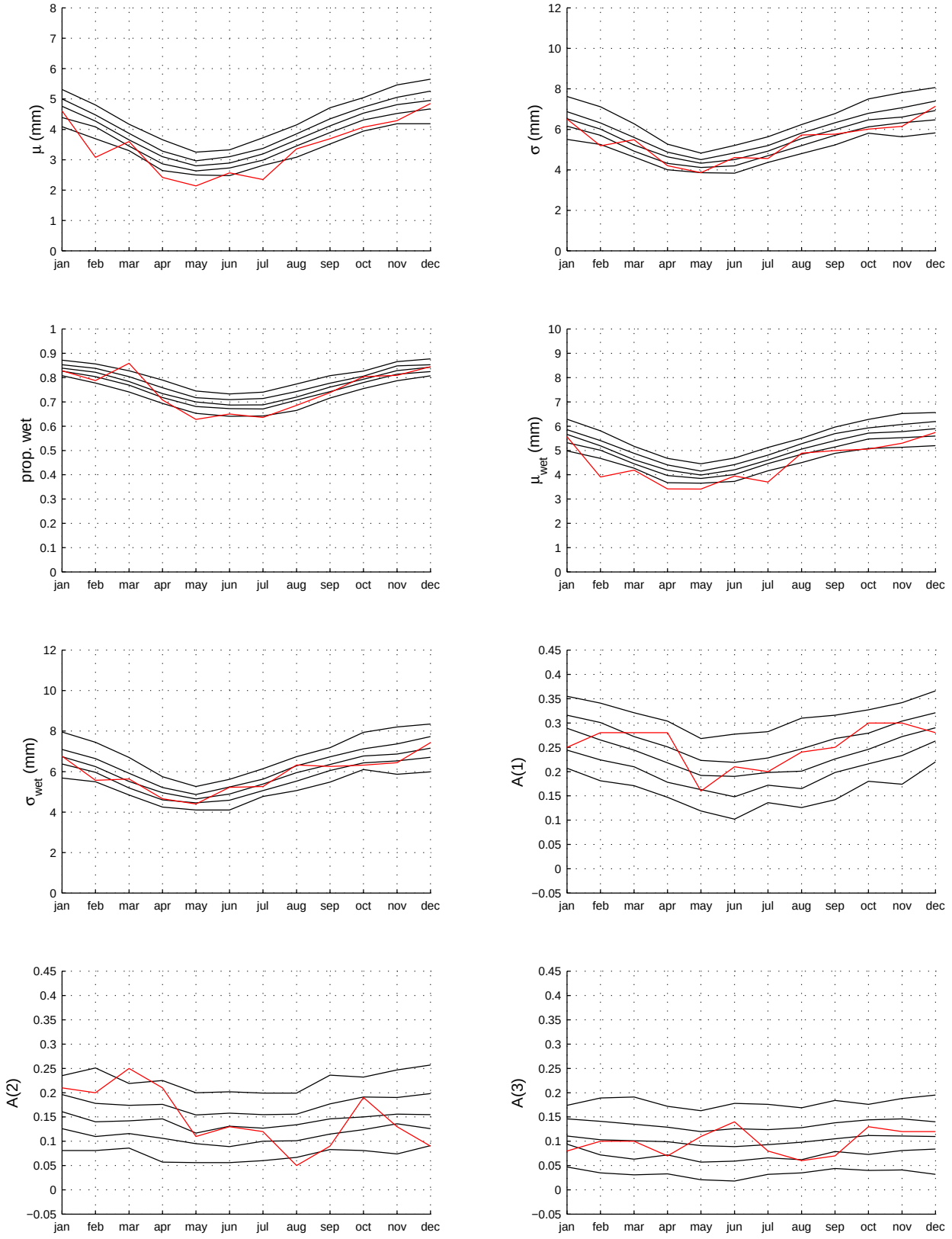

Figure A.19: *North East Lancashire 10km square 1:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

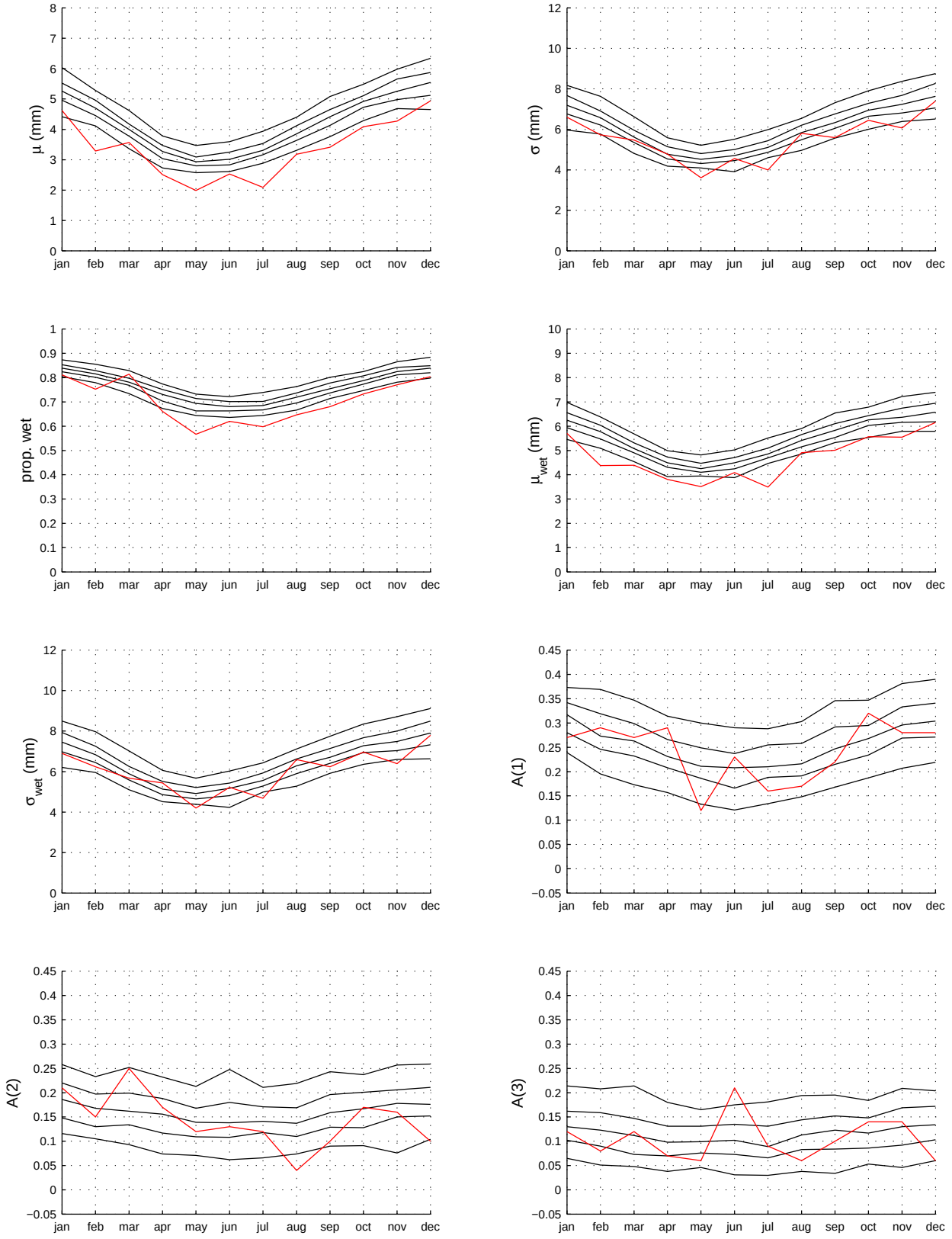

Figure A.20: *North East Lancashire 10km square 2:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

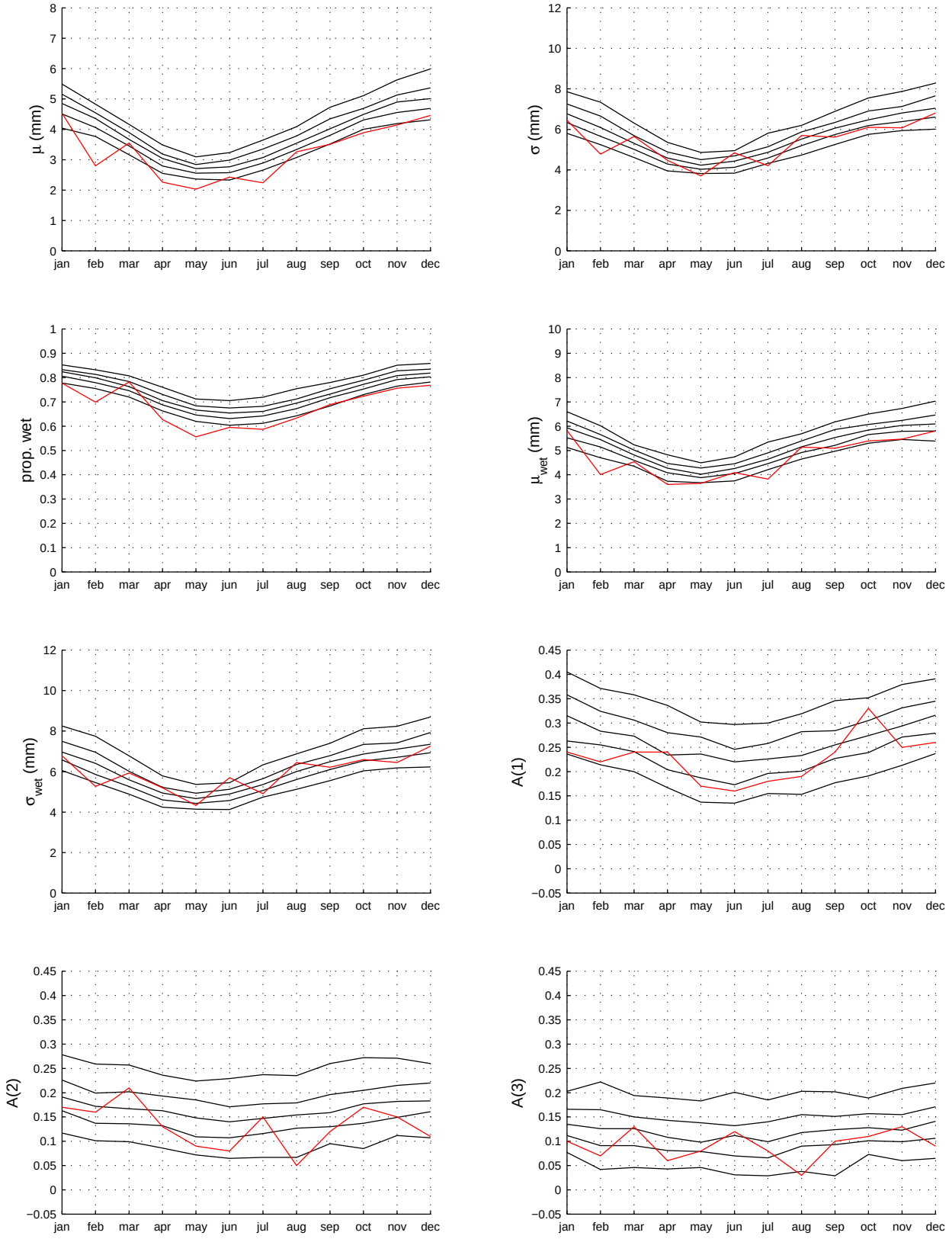

Figure A.21: *North East Lancashire 10km square 3:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

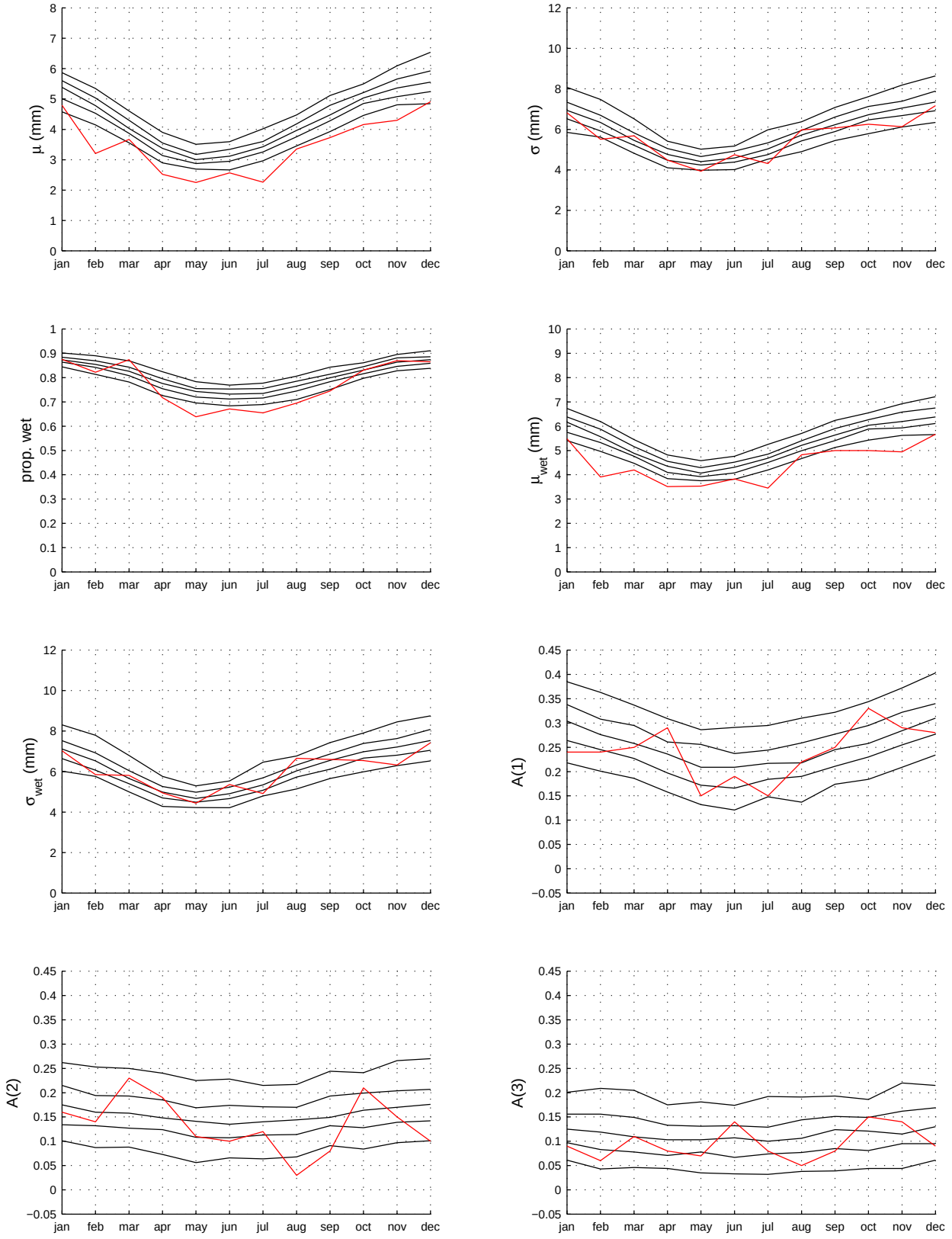

Figure A.22: *North East Lancashire 10km square 4:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

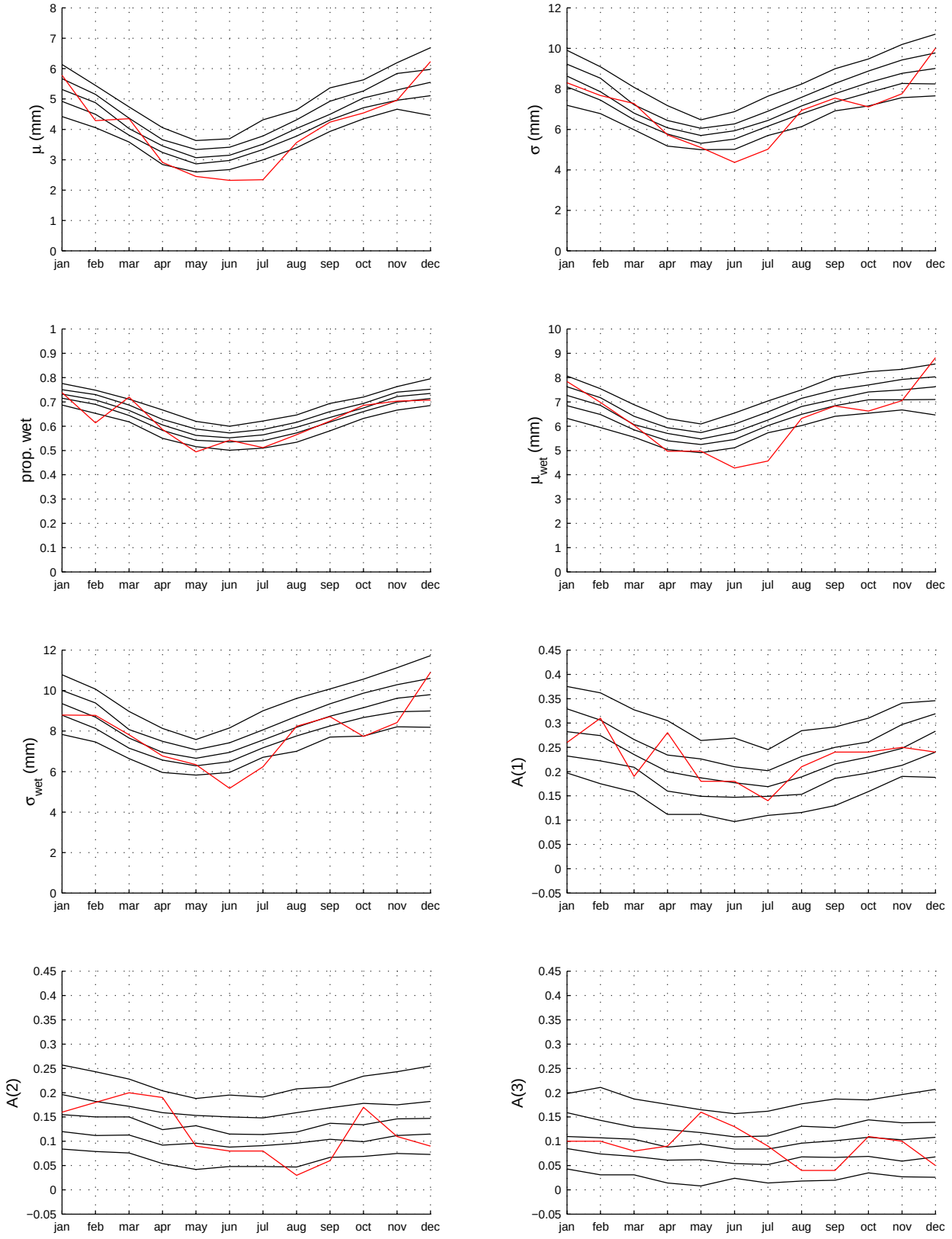

Figure A.23: *North East Lancashire gauge L06:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

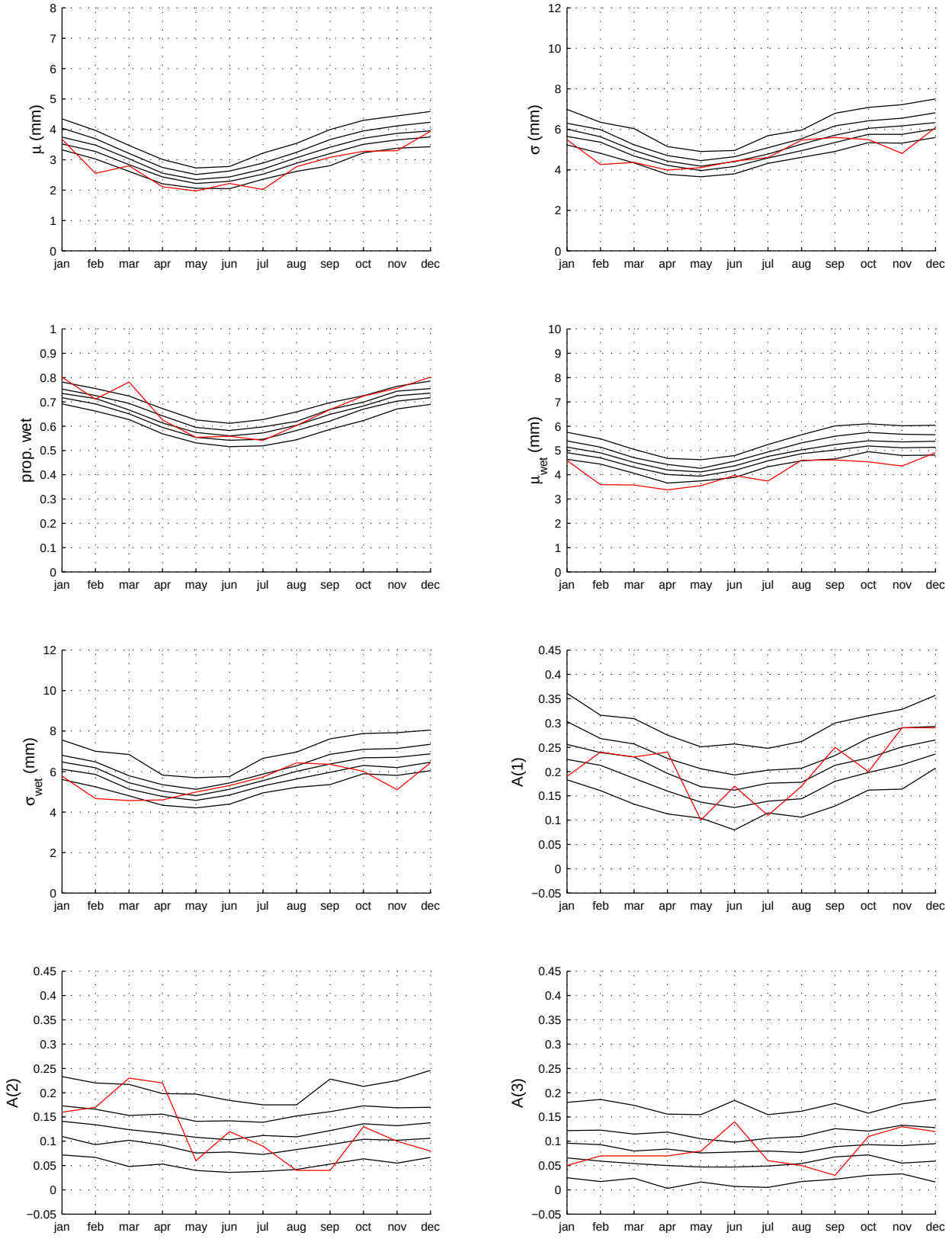

Figure A.24: *North East Lancashire gauge L10:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

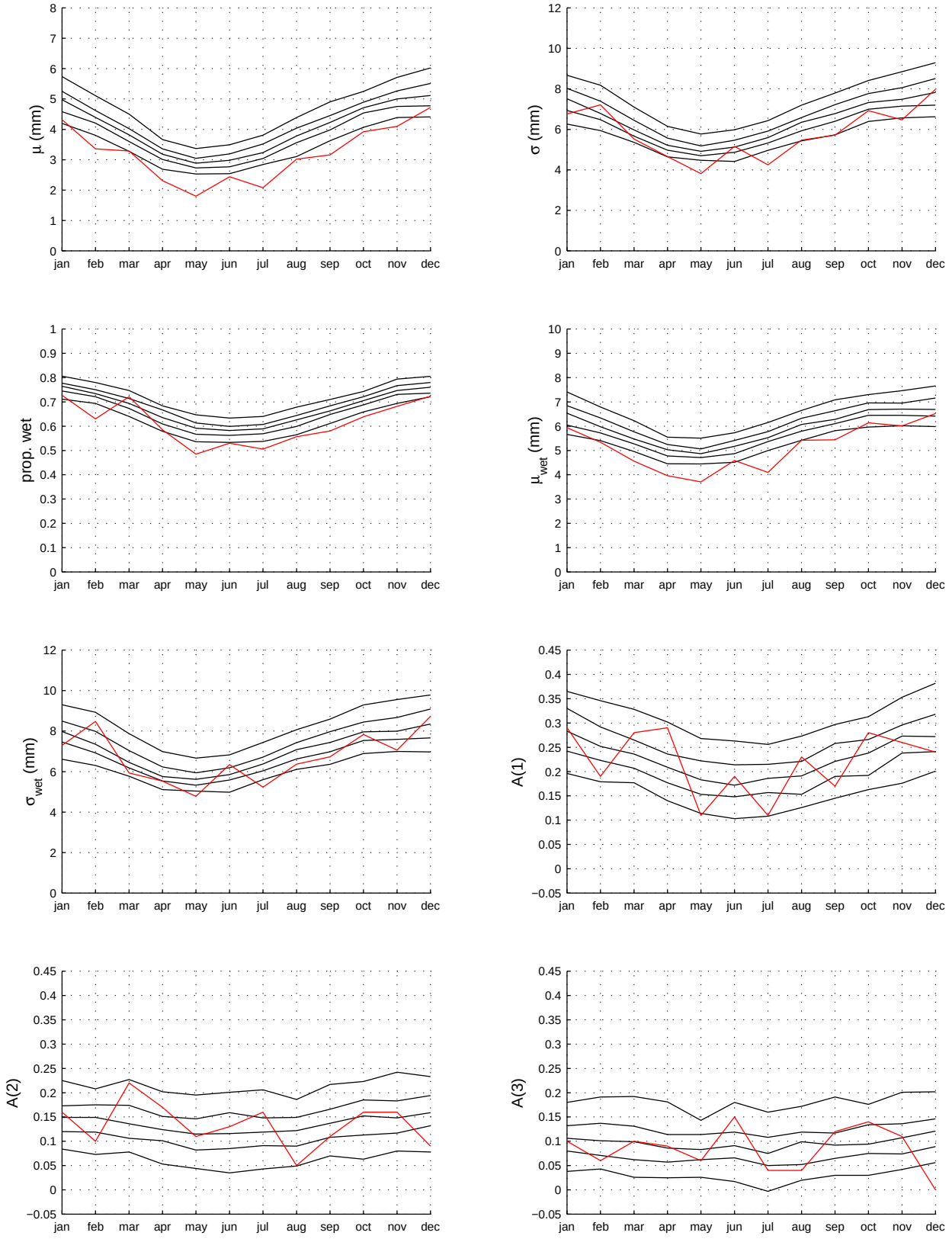

Figure A.25: *North East Lancashire gauge L17:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

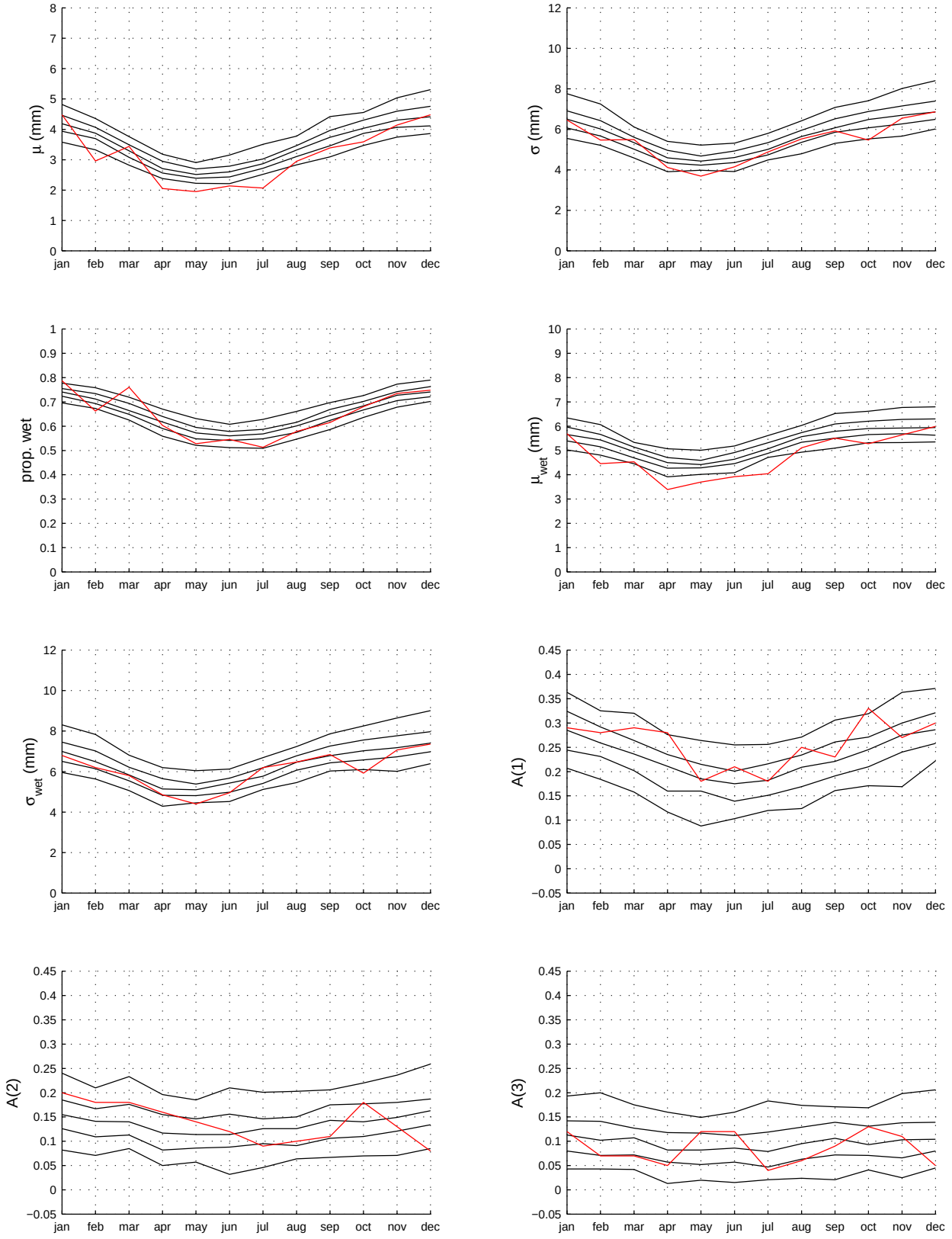

Figure A.26: *North East Lancashire gauge L19:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

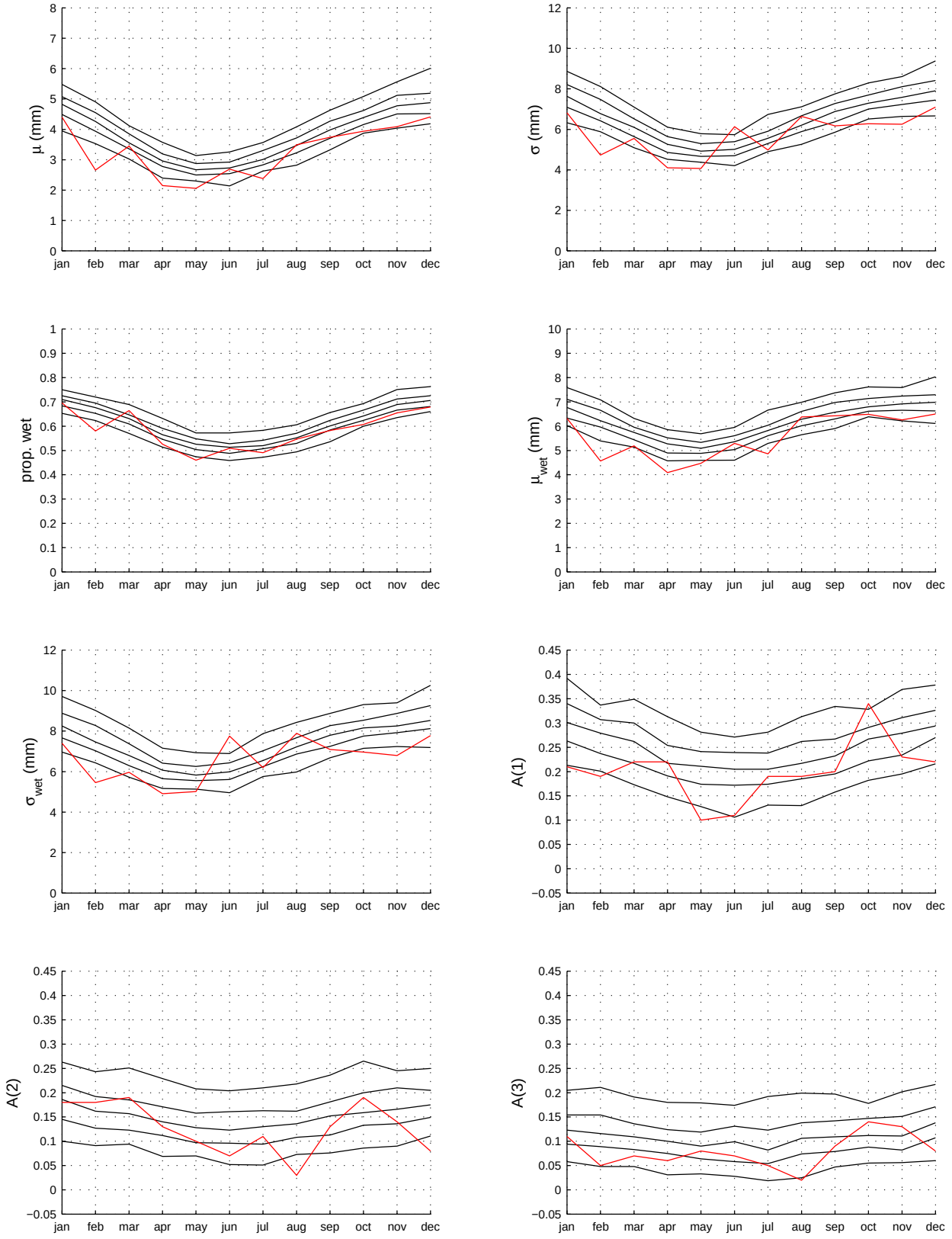

Figure A.27: *North East Lancashire gauge L35:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

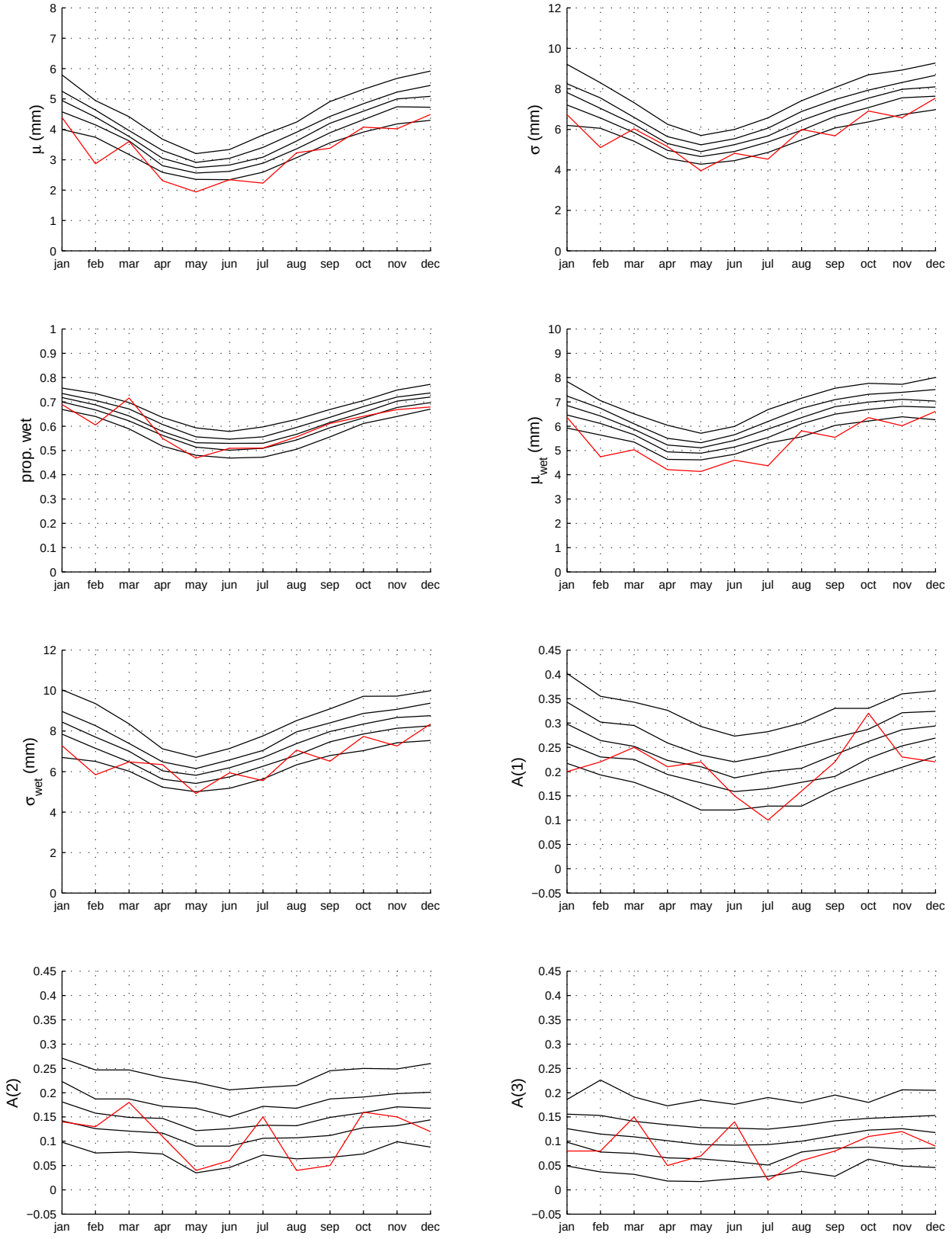

Figure A.28: *North East Lancashire gauge L41:* Historic statistics and percentiles (5, 25, 50, 75 and 95) for simulated statistics. From left to right and top to bottom are mean, standard deviation, proportion of wet days, mean of wet days, standard deviation of wet days, first second and third autocorrelations.

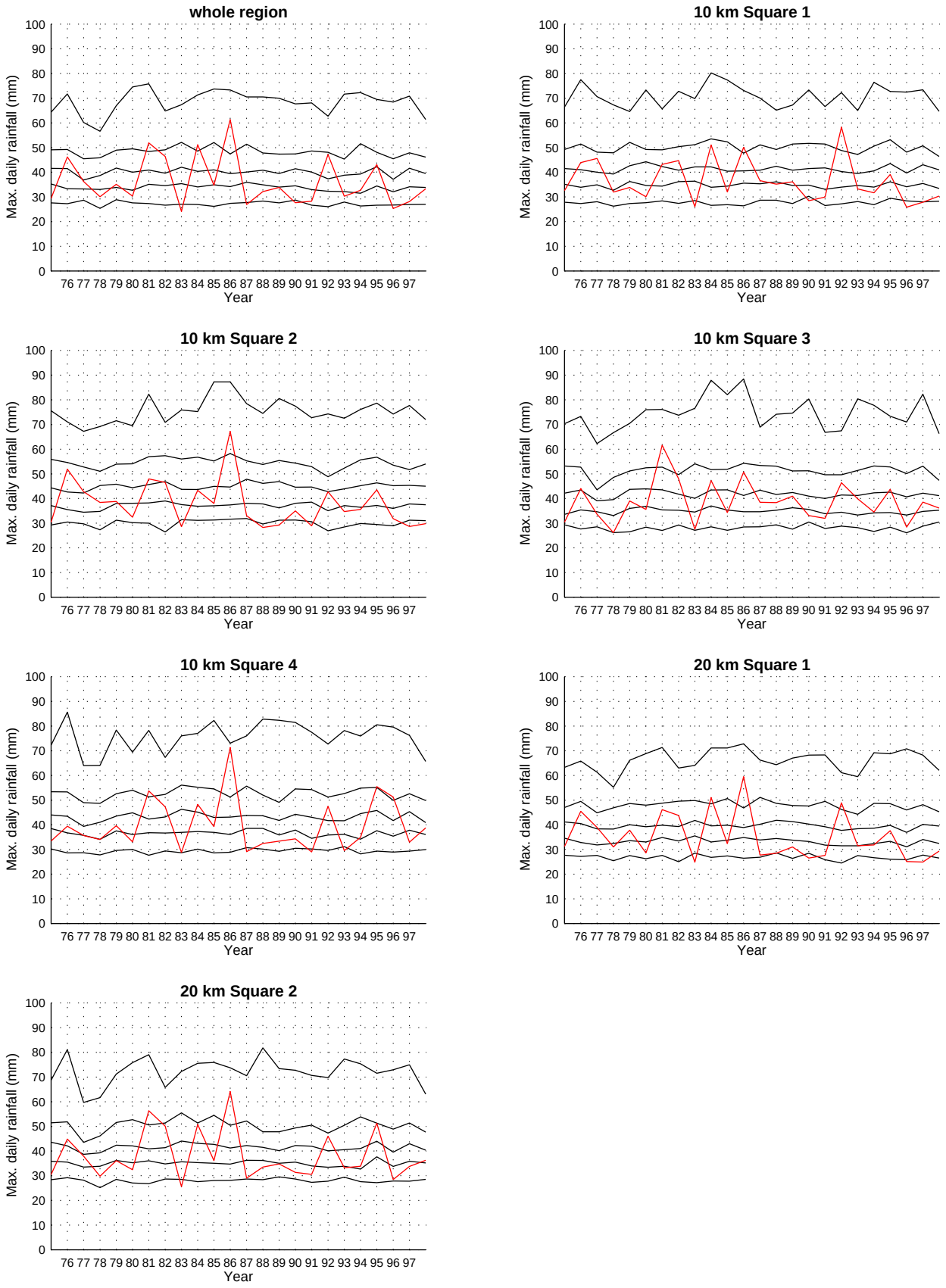

Figure A.29: *North East Lancashire whole region and 20km and 10km squares:* Historic annual maxima and percentiles (5, 25, 50, 75 and 95) for simulated annual maxima.

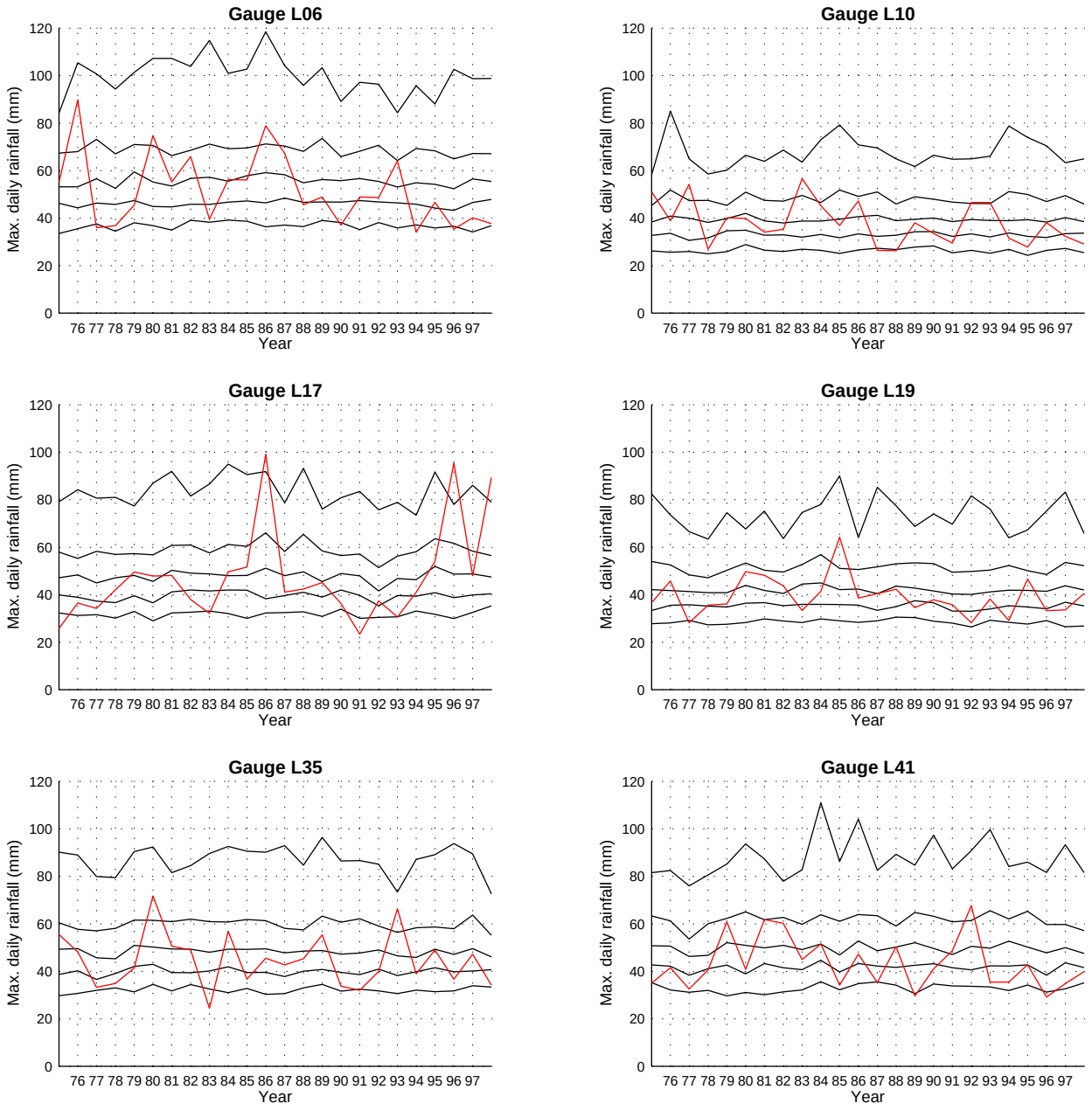

Figure A.30: *North East Lancashire gauges:* Historic annual maxima and percentiles (5, 25, 50, 75 and 95) for simulated annual maxima.

Appendix B

Results files for GLMs reported in Section 6.3

B.1 Blackwater Fitting Results

B.1.1 Blackwater logistic model fitting file

Final parameter estimates:

Main effect:	Coefficient	Std Err
Constant	-1.150700	0.0098
Fourier sine component 1 for Eastings (OS grid	0.041200	0.0084
Fourier cosine component 1 for Eastings (OS gr	0.003100	0.0076
Fourier sine component 1 for Northings (OS gri	-0.066900	0.0103
Fourier cosine component 1 for Northings (OS g	-0.031600	0.0064
Fourier sine component 2 for Eastings (OS grid	0.022300	0.0103
Fourier sine component 2 for Northings (OS gri	-0.122800	0.0083
Fourier cosine component 2 for Northings (OS g	0.044100	0.0088
Indiacator for site B38	-0.863400	0.0925
Monthly seasonal effect, cosine component	0.232500	0.0079
Monthly seasonal effect, sine component	0.038900	0.0074
Monthly half-year cycle, cosine component	0.039400	0.0062
Monthly half-year cycle, sine component	-0.014400	0.0062
Ln(1 + Previous day's value)	0.873700	0.0073
Ln(1 + Value 2 days before)	0.211400	0.0070
Ln(1 + Value 3 days before)	0.144500	0.0064
Trace indicator for Previous day's value	0.590200	0.0145

```
Trace indicator for Value 2 days before             0.209800   0.0148
Trace indicator for Value 3 days before             0.202400   0.0145
Ln(1 + Value 4 days before)                         0.094000   0.0058
Ln(1 + Value 5 days before)                         0.079400   0.0054
Trace indicator for Value 4 days before             0.126400   0.0140
Persistence indicator for preceding  2 days         0.152000   0.0203
Persistence indicator for preceding  3 days        -0.070500   0.0198

2-way interactions:                                 Coefficient  Std Err
-------------------                                 -----------  -------
Monthly seasonal effect, cosine component           0.029400   0.0094
with Ln(1 + Previous day's value)
Monthly seasonal effect, sine component             0.131300   0.0081
with Ln(1 + Previous day's value)
Monthly seasonal effect, cosine component          -0.090800   0.0159
with Persistence indicator for preceding  2 d
Fourier sine component  1 for Eastings (OS grid     0.166800   0.0182
with Fourier sine component  1 for Northings
Fourier sine component  2 for Eastings (OS grid    -0.157000   0.0136
with Fourier sine component  2 for Northings

Parameters in non-linear transformations:
-----------------------------------------
Fourier sine component  1 for Eastings (OS grid:
          Lower limit for Fourier repres:  460.0000 (prespecified)
          Upper limit for Fourier repres:  510.0000 (prespecified)
Fourier sine component  1 for Northings (OS gri:
          Lower limit for Fourier repres:  130.0000 (prespecified)
          Upper limit for Fourier repres:  170.0000 (prespecified)

Spatial dependence structure:
-----------------------------
Structure used is Beta-Binomial distribution for number of wet sites
          Shape parameter of Beta-Binomi:    0.5009


                            RESIDUAL ANALYSIS
                            =================

                 |    Observed    ||     % correct
                 |   Dry     Wet  || Observed  Expected
     ----------------------------------------------------
     Forecast dry | 110007   48147 ||   69.6      69.5
```

```
    Forecast wet |  30409   72555 ||   70.5      70.1
    ------------------------------------------------
                 OVERALL % CORRECT :   69.9      69.7

Rainday frequencies vs forecasts:
---------------------------------
Forecast decile |  1  |  2  |  3  |  4  |  5  |  6  |  7  |  8  |  9  | 10  |
-------------------------------------------------------------------------------
Observed propn. |0.000|0.170|0.224|0.356|0.505|0.598|0.676|0.738|0.801|0.859|
Expected propn. |0.000|0.187|0.248|0.344|0.447|0.552|0.649|0.748|0.848|0.927|
Number of cases |    0| 6825|81656|43575|26098|26304|28331|22578|19310| 6441|
-------------------------------------------------------------------------------

        Mean Pearson residual: -0.0063
        Standard deviation of Pearson residuals:  1.0119
        Standard error of mean Pearson residual:  0.0061


    MODEL PERFORMANCE BY MONTH
    --------------------------

           |        ||     Pearson residuals
     Month | N days ||   Mean    Std Dev   S.E. mean
    -------------------------------------------------
         1 |  21351 ||  0.0151    1.0270     0.0211
         2 |  19736 || -0.0321    1.0302     0.0219
         3 |  22072 || -0.0016    1.0049     0.0208
         4 |  21980 || -0.0086    1.0007     0.0208
         5 |  23014 ||  0.0064    1.0219     0.0203
         6 |  21960 || -0.0047    1.0043     0.0208
         7 |  22898 || -0.0109    1.0025     0.0204
         8 |  23220 || -0.0054    1.0222     0.0202
         9 |  21684 ||  0.0047    0.9906     0.0209
        10 |  21662 || -0.0142    0.9842     0.0209
        11 |  20774 || -0.0069    1.0304     0.0214
        12 |  20767 || -0.0208    1.0239     0.0214
    -------------------------------------------------

    MODEL PERFORMANCE BY SITE
    -------------------------

                                  |        ||     Pearson residuals
      Site  Name                  | N days ||   Mean    Std Dev   S.E. mean
```

B01	Site 270304 - SHINFI	1827	-0.0050	1.0097	0.0234
B02	Site 270668 - BASING	3345	-0.0430	0.9981	0.0173
B04	Site 271264 - CAMP F	4450	-0.0632	0.9952	0.0150
B06	Site 271491 - CAMBER	174	-0.0031	0.9965	0.0758
B07	Site 271868 - LONG S	7722	-0.0153	1.0087	0.0114
B08	Site 271911 - GREYWE	10819	0.0076	0.9969	0.0096
B09	Site 271922 - NORTH	4772	0.0095	0.9985	0.0145
B10	Site 271976 - ODIHAM	541	0.0926	0.9190	0.0430
B11	Site 272734 - BRACKN	3984	0.0890	0.9635	0.0158
B14	Site 272853 - WOKING	861	0.1467	0.9398	0.0341
B15	Site 279906 - VIRGIN	6198	0.0749	0.9590	0.0127
B16	Site 279940 - CHERTS	8534	-0.0325	1.0110	0.0108
B18	Site 280037 - SHEPPE	23969	-0.0488	1.0132	0.0065
B19	Site 280218 - ALTON,	5855	-0.0313	0.9996	0.0131
B20	Site 280369 - ROTHER	22753	-0.0146	1.0357	0.0066
B21	Site 280576 - ALTON	9729	-0.0633	1.0013	0.0101
B22	Site 280822 - BADSHO	4405	0.0214	0.9937	0.0151
B23	Site 280825 - FARNHA	9118	-0.1027	1.0141	0.0105
B25	Site 280838 - ALICE	8461	0.0718	0.9843	0.0109
B27	Site 280848 - BOUNDS	9376	-0.0032	1.0083	0.0103
B28	Site 280957 - SHOTTE	13023	0.0761	1.0565	0.0088
B29	Site 281185 - BORDON	5020	-0.0652	1.0180	0.0141
B31	Site 281278 - SELBOR	8859	-0.0420	1.0112	0.0106
B33	Site 282787 - MERROW	9482	0.0232	1.0140	0.0103
B34	Site 282942 - PIRBRI	9963	-0.0124	1.0224	0.0100
B35	Site 282960 - MERRIS	3021	0.1241	0.9388	0.0182
B36	Site 283424 - WISLEY	26436	0.0064	1.0144	0.0062
B37	Site 283429 - WISLEY	8770	-0.0285	1.0069	0.0107
B38	Site 283677 - ASCOT,	814	-0.0142	0.9810	0.0350
B39	Site 283710 - BAGSHO	1487	0.0186	0.9551	0.0259
B40	Site 316947 - DUNSFO	797	0.0028	1.0145	0.0354
B41	Site 324702 - ROPLEY	5468	0.0977	0.9798	0.0135
B42	Site 324738 - BISHOP	12544	-0.0312	1.0326	0.0089
B43	Site 325034 - PRESTO	8541	-0.0253	1.0208	0.0108

MODEL PERFORMANCE BY YEAR

		Pearson residuals		
Year	N days	Mean	Std Dev	S.E. mean

1908	359	-0.0541	1.0361	0.1627
1909	365	0.0397	1.0419	0.1614
1911	358	-0.1038	0.9998	0.1630
1912	366	-0.0091	1.0255	0.1612
1913	723	-0.1545	1.0629	0.1147
1914	730	-0.0998	1.0468	0.1141
1915	730	-0.1330	1.0211	0.1141
1916	732	-0.0400	1.0375	0.1140
1917	730	-0.0679	1.0471	0.1141
1918	730	-0.0499	1.0474	0.1141
1919	730	-0.0150	0.9875	0.1141
1920	732	-0.0059	1.0154	0.1140
1921	730	-0.1647	0.9525	0.1141
1922	730	-0.0182	1.0174	0.1141
1923	1088	0.0287	1.0256	0.0935
1924	366	0.0503	1.0425	0.1612
1925	723	0.0373	1.0078	0.1147
1926	730	0.0337	1.0451	0.1141
1927	730	0.0059	1.0917	0.1141
1928	732	0.0381	1.0631	0.1140
1929	730	-0.0968	1.0039	0.1141
1930	730	0.0488	1.0637	0.1141
1931	1088	-0.0064	1.0543	0.0935
1932	1098	0.0227	1.0238	0.0931
1933	1095	-0.0818	1.0461	0.0932
1934	1095	-0.0156	1.0421	0.0932
1935	1095	-0.0127	1.0274	0.0932
1936	1098	0.0121	1.0303	0.0931
1937	1095	-0.0062	1.0498	0.0932
1938	1095	0.0163	1.0033	0.0932
1939	1095	0.0250	1.0320	0.0932
1940	1098	-0.1071	1.0347	0.0931
1941	1095	-0.0277	0.9982	0.0932
1942	1095	-0.0884	1.0429	0.0932
1943	1095	-0.0995	1.0429	0.0932
1944	1098	-0.0728	1.0310	0.0931
1945	1095	-0.0565	1.0143	0.0932
1946	1095	0.0035	1.0462	0.0932
1947	1095	-0.1085	1.0170	0.0932
1948	1098	-0.0786	1.0193	0.0931
1949	1095	-0.0878	0.9899	0.0932
1950	1095	0.0661	1.0298	0.0932
1951	1095	0.0715	1.0464	0.0932

1952	1098	0.0043	1.0358	0.0931
1953	1095	-0.0656	0.9861	0.0932
1954	1095	0.0620	1.0252	0.0932
1955	1095	-0.0532	1.0392	0.0932
1956	1098	0.0125	1.0334	0.0931
1957	1095	0.0183	1.0263	0.0932
1958	1095	-0.0066	1.1017	0.0932
1959	1453	-0.0971	0.9707	0.0809
1960	1464	0.1408	1.0069	0.0806
1961	4324	-0.0465	1.0292	0.0469
1962	5020	-0.0873	1.0702	0.0435
1963	4572	0.0132	1.0642	0.0456
1964	5401	-0.0907	1.0346	0.0420
1965	5467	0.0347	1.0509	0.0417
1966	5529	0.0067	1.0192	0.0415
1967	5772	-0.0372	1.0512	0.0406
1968	5726	-0.0513	1.0726	0.0407
1969	6455	-0.0543	1.0326	0.0384
1970	6570	0.0522	1.0451	0.0380
1971	6852	-0.1265	1.0187	0.0372
1972	6676	0.0020	1.0067	0.0377
1973	4400	-0.0685	1.0233	0.0465
1974	4349	0.1076	0.9547	0.0468
1975	4625	-0.0162	1.0254	0.0453
1976	4407	-0.0882	0.9299	0.0464
1977	4121	0.1165	0.9979	0.0480
1978	4270	0.0239	0.9865	0.0472
1979	4427	0.0663	1.0012	0.0463
1980	3914	0.0267	1.0150	0.0493
1981	4867	0.0528	0.9745	0.0442
1982	5705	0.0450	1.0086	0.0408
1983	5463	0.0257	0.9527	0.0417
1984	5637	0.0243	0.9925	0.0411
1985	5482	0.0065	0.9984	0.0416
1986	5317	0.0477	1.0131	0.0423
1987	5280	0.0590	0.9915	0.0424
1988	5344	0.0390	0.9972	0.0422
1989	5808	-0.0603	0.9590	0.0405
1990	5864	-0.1114	1.0063	0.0403
1991	5813	-0.0124	0.9967	0.0404
1992	5583	0.0410	1.0113	0.0413
1993	6684	0.0003	0.9259	0.0377
1994	5574	0.0318	1.0627	0.0413

```
  1995 |   5881 || -0.0640    0.9445     0.0402
  1996 |   6084 || -0.0191    1.0155     0.0395
  1997 |   6690 || -0.0357    0.9552     0.0377
  1998 |   5770 ||  0.0552    0.9842     0.0406
  1999 |   5708 ||  0.0764    1.0004     0.0408
  2000 |   1522 ||  0.0514    0.9816     0.0790
-----------------------------------------------

NOTE: Standard errors computed for overall, monthly and yearly
means above are inflated by a factor of 3.083 to compensate for
spatial dependence.
```

B.1.2 Blackwater gamma model fitting file

```
 Final parameter estimates:

Main effect:                                    Coefficient  Std Err
------------                                    -----------  -------
Constant                                          1.251700   0.0085
Fourier sine component  1 for Northings (OS gri   0.115000   0.0063
Fourier cosine component  1 for Northings (OS g   0.052000   0.0050
Fourier sine component  2 for Eastings (OS grid   0.032400   0.0079
Fourier cosine component  2 for Eastings (OS gr   0.043500   0.0051
Fourier sine component  2 for Northings (OS gri   0.107100   0.0071
Indiacator for site B38                           0.358100   0.0876
Monthly seasonal effect, cosine component        -0.212700   0.0095
Monthly seasonal effect, sine component          -0.165600   0.0091
Ln(1 + Previous day's value)                      0.135700   0.0045
Ln(1 + Value 2 days before)                       0.059800   0.0048
Ln(1 + Value 3 days before)                       0.022500   0.0051
Ln(1 + Value 4 days before)                       0.041400   0.0048
Ln(1 + Value 5 days before)                       0.027500   0.0048
Ln(1 + Value 6 days before)                      -0.013200   0.0044
Trace indicator for Previous day's value          0.024600   0.0116
Trace indicator for Value 2 days before           0.034300   0.0117
Trace indicator for Value 3 days before          -0.074100   0.0120
Trace indicator for Value 4 days before          -0.044100   0.0116
Trace indicator for Value 5 days before          -0.034200   0.0118
Persistence indicator for preceding  3 days      -0.062000   0.0138
Persistence indicator for preceding  5 days      -0.070700   0.0147
```

```
2-way interactions:                                Coefficient  Std Err
-------------------                                -----------  -------
Fourier sine component  2 for Eastings (OS grid      0.076600   0.0111
with Fourier sine component  2 for Northings
Monthly seasonal effect, cosine component            0.073500   0.0059
with Ln(1 + Previous day's value)
Monthly seasonal effect, sine component              0.030300   0.0059
with Ln(1 + Previous day's value)
Monthly seasonal effect, cosine component            0.081500   0.0061
with Ln(1 + Value 2 days before)
Monthly seasonal effect, sine component             -0.017600   0.0064
with Ln(1 + Value 2 days before)
Monthly seasonal effect, cosine component            0.045700   0.0062
with Ln(1 + Value 3 days before)
Monthly seasonal effect, sine component              0.029800   0.0066
with Ln(1 + Value 3 days before)
Monthly seasonal effect, cosine component            0.051200   0.0159
with Trace indicator for Previous day's value
Persistence indicator for preceding  3 days         -0.071600   0.0149
with Monthly seasonal effect, sine component
Monthly seasonal effect, cosine component           -0.042900   0.0161
with Persistence indicator for preceding  5 d
Monthly seasonal effect, cosine component            0.033400   0.0063
with Ln(1 + Value 4 days before)
Monthly seasonal effect, sine component              0.014100   0.0058
with Ln(1 + Value 6 days before)

Parameters in non-linear transformations:
-----------------------------------------
Fourier sine component  1 for Northings (OS gri:
          Lower limit for Fourier repres:  130.0000 (prespecified)
          Upper limit for Fourier repres:  170.0000 (prespecified)
Fourier sine component  2 for Eastings (OS grid:
          Lower limit for Fourier repres:  460.0000 (prespecified)
          Upper limit for Fourier repres:  510.0000 (prespecified)

Spatial dependence structure:
-----------------------------
Structure used is Constant correlation between all pairs of sites
          Inter-site correlation        :    0.8122


                          RESIDUAL ANALYSIS
```

```
================

Mean of observations:   4.457
Standard deviation of observations:   5.911

Mean error (observed - predicted):  -0.040
Root mean squared error:   5.801
Proportion of variance explained by model:  0.037

Mean Pearson residual:  -0.005
Standard deviation of Pearson residuals:   1.276
Expected std dev of Pearson residuals:   1.276

Mean Anscombe residual:  0.8453 (expected:  0.8210)
Std Dev of Anscombe residuals:  0.3783 (expected:  0.4258)


MODEL PERFORMANCE BY MONTH
--------------------------
```

			Pearson residuals		
Month	N days		Mean	Std Dev	S.E. mean
1	12503		0.018	1.182	0.035
2	9886		-0.027	1.184	0.040
3	10558		-0.013	1.220	0.038
4	9967		0.011	1.162	0.040
5	9823		-0.019	1.289	0.040
6	8620		0.021	1.387	0.043
7	8194		-0.048	1.439	0.044
8	8843		-0.035	1.333	0.042
9	9085		0.021	1.418	0.041
10	10229		0.012	1.309	0.039
11	11355		0.006	1.256	0.037
12	11639		-0.017	1.197	0.037

```
MODEL PERFORMANCE BY SITE
-------------------------
```

				Pearson residuals		
Site	Name		N days	Mean	Std Dev	S.E. mean

B01	Site 270304 - SHINFI	858	-0.098	1.211	0.044
B02	Site 270668 - BASING	1159	0.001	1.241	0.037
B04	Site 271264 - CAMP F	1748	0.007	1.337	0.031
B06	Site 271491 - CAMBER	73	-0.254	0.814	0.149
B07	Site 271868 - LONG S	3972	0.164	1.471	0.020
B08	Site 271911 - GREYWE	5087	-0.007	1.283	0.018
B09	Site 271922 - NORTH	2187	-0.089	1.227	0.027
B10	Site 271976 - ODIHAM	324	0.067	1.399	0.071
B11	Site 272734 - BRACKN	2245	-0.046	1.352	0.027
B14	Site 272853 - WOKING	490	-0.117	1.092	0.058
B15	Site 279906 - VIRGIN	3376	-0.063	1.312	0.022
B16	Site 279940 - CHERTS	3652	0.019	1.336	0.021
B18	Site 280037 - SHEPPE	10292	0.014	1.258	0.013
B19	Site 280218 - ALTON,	2855	0.001	1.260	0.024
B20	Site 280369 - ROTHER	10369	0.036	1.222	0.013
B21	Site 280576 - ALTON	4324	0.030	1.251	0.019
B22	Site 280822 - BADSHO	2311	0.003	1.271	0.027
B23	Site 280825 - FARNHA	3800	0.080	1.301	0.021
B25	Site 280838 - ALICE	4319	-0.020	1.256	0.019
B27	Site 280848 - BOUNDS	4383	-0.050	1.233	0.019
B28	Site 280957 - SHOTTE	6137	-0.098	1.120	0.016
B29	Site 281185 - BORDON	2142	-0.024	1.156	0.028
B31	Site 281278 - SELBOR	4009	0.009	1.330	0.020
B33	Site 282787 - MERROW	4561	0.041	1.424	0.019
B34	Site 282942 - PIRBRI	4303	-0.038	1.237	0.019
B35	Site 282960 - MERRIS	1623	-0.023	1.355	0.032
B36	Site 283424 - WISLEY	12570	-0.013	1.342	0.011
B37	Site 283429 - WISLEY	3902	0.023	1.387	0.020
B38	Site 283677 - ASCOT,	232	0.000	0.937	0.084
B39	Site 283710 - BAGSHO	870	0.076	1.440	0.043
B40	Site 316947 - DUNSFO	336	-0.007	1.032	0.070
B41	Site 324702 - ROPLEY	2948	-0.093	1.192	0.024
B42	Site 324738 - BISHOP	5258	-0.059	1.101	0.018
B43	Site 325034 - PRESTO	3987	-0.018	1.211	0.020

MODEL PERFORMANCE BY YEAR

		Pearson residuals		
Year	N days	Mean	Std Dev	S.E. mean
1908	153	-0.040	1.217	0.319

1909	185	-0.038	1.216	0.290
1911	141	0.039	1.235	0.333
1912	177	-0.003	1.149	0.297
1913	263	0.157	1.276	0.244
1914	302	0.133	1.357	0.227
1915	293	0.337	1.586	0.231
1916	340	0.092	1.241	0.214
1917	309	0.124	1.488	0.225
1918	322	0.114	1.408	0.220
1919	331	0.035	1.314	0.217
1920	335	-0.083	1.178	0.216
1921	224	-0.257	0.981	0.264
1922	330	-0.036	1.136	0.217
1923	529	-0.026	1.272	0.172
1924	196	0.155	1.892	0.282
1925	359	0.036	1.332	0.208
1926	362	-0.055	1.097	0.208
1927	379	0.191	1.393	0.203
1928	376	0.031	1.212	0.204
1929	289	0.050	1.236	0.232
1930	363	-0.088	1.048	0.207
1931	499	-0.045	1.120	0.177
1932	517	-0.076	1.195	0.174
1933	446	0.007	1.503	0.187
1934	494	-0.065	1.051	0.178
1935	515	0.082	1.263	0.174
1936	538	-0.040	1.074	0.170
1937	540	0.104	1.247	0.170
1938	494	-0.215	1.066	0.178
1939	544	-0.036	1.158	0.169
1940	444	0.092	1.343	0.187
1941	512	0.072	1.348	0.175
1942	445	0.067	1.190	0.187
1943	424	-0.015	1.089	0.192
1944	438	-0.053	1.229	0.189
1945	450	-0.022	1.362	0.186
1946	533	0.075	1.191	0.171
1947	425	0.127	1.869	0.192
1948	463	0.096	1.153	0.184
1949	432	-0.035	1.241	0.190
1950	563	-0.040	1.223	0.166
1951	614	0.108	1.321	0.159
1952	516	0.007	1.167	0.174

1953	438	-0.083	1.123	0.189
1954	577	-0.019	1.251	0.164
1955	453	-0.037	1.355	0.186
1956	508	-0.089	1.309	0.175
1957	513	-0.114	1.072	0.174
1958	522	0.050	1.257	0.173
1959	565	0.085	1.345	0.166
1960	848	-0.049	1.152	0.136
1961	1851	-0.018	1.118	0.092
1962	2016	0.027	1.300	0.088
1963	2186	-0.104	1.050	0.084
1964	2135	0.019	1.374	0.085
1965	2687	-0.034	1.117	0.076
1966	2684	0.064	1.250	0.076
1967	2668	0.162	1.396	0.076
1968	2593	0.213	1.561	0.078
1969	2753	0.010	1.404	0.075
1970	3353	-0.104	1.040	0.068
1971	2629	0.162	1.611	0.077
1972	3187	-0.111	1.023	0.070
1973	1688	-0.034	1.505	0.096
1974	2467	0.082	1.341	0.080
1975	2090	0.051	1.500	0.086
1976	1680	-0.038	1.159	0.096
1977	2203	-0.108	1.238	0.084
1978	2046	-0.002	1.297	0.087
1979	2342	0.066	1.406	0.082
1980	1867	-0.107	1.235	0.091
1981	2500	0.010	1.284	0.079
1982	2910	-0.031	1.184	0.073
1983	2578	-0.047	1.348	0.078
1984	2753	0.063	1.459	0.075
1985	2538	-0.094	1.079	0.078
1986	2743	-0.037	1.159	0.075
1987	2623	-0.046	1.382	0.077
1988	2568	-0.130	1.051	0.078
1989	2394	0.019	1.404	0.081
1990	2181	-0.074	1.091	0.085
1991	2559	-0.071	1.270	0.078
1992	2761	-0.047	1.223	0.075
1993	3073	0.039	1.306	0.071
1994	2798	-0.025	1.100	0.075
1995	2403	0.012	1.253	0.081

```
  1996 |   2620 ||  -0.043     1.378      0.077
  1997 |   2846 ||  -0.095     1.055      0.074
  1998 |   3036 ||   0.030     1.263      0.072
  1999 |   3011 ||  -0.005     1.312      0.072
  2000 |    854 ||   0.097     1.478      0.135
-----------------------------------------------

NOTE: Standard errors computed for overall, monthly and yearly
means above are inflated by a factor of 3.094 to compensate for
spatial dependence.
```

B.2 North East Lancashire Fitting results

B.2.1 North East Lancashire logistic model fitting file

```
 Final parameter estimates:

Main effect:                                    Coefficient  Std Err
------------                                    -----------  -------
Constant                                         -0.964300   0.0177
Altitude (hundreds of metres                      0.088400   0.0078
Legendre polynomial  1 for Eastings (OS grid, k   0.033000   0.0090
Legendre polynomial  1 for Northings (OS grid,    0.049600   0.0120
Legendre polynomial  2 for Eastings (OS grid, k   0.084000   0.0129
Legendre polynomial  2 for Northings (OS grid,   -0.100900   0.0124
Legendre polynomial  3 for Eastings (OS grid, k  -0.004800   0.0163
Legendre polynomial  3 for Northings (OS grid,   -0.180700   0.0174
Monthly seasonal effect, cosine component         0.270700   0.0090
Monthly seasonal effect, sine component           0.069000   0.0081
Monthly half-year cycle, cosine component         0.012500   0.0062
Monthly half-year cycle, sine component           0.012200   0.0062
Ln(1 + Previous day's value)                      0.972300   0.0074
Ln(1 + Value 2 days before)                       0.141700   0.0068
Ln(1 + Value 3 days before)                       0.110500   0.0062
Ln(1 + Value 4 days before)                       0.094600   0.0060
Ln(1 + Value 5 days before)                       0.041800   0.0051
Trace indicator for Previous day's value          0.421800   0.0136
Trace indicator for Value 2 days before           0.153500   0.0139
Trace indicator for Value 3 days before           0.171000   0.0137
Trace indicator for Value 4 days before           0.132500   0.0135
```

```
Trace indicator for Value 5 days before              0.103900   0.0131
Persistence indicator for preceding  2 days          0.362500   0.0207
Persistence indicator for preceding  3 days         -0.095300   0.0257
Persistence indicator for preceding  4 days          0.048300   0.0219

2-way interactions:                                 Coefficient  Std Err
-------------------                                 -----------  -------
Monthly seasonal effect, cosine component           -0.041500   0.0254
with Persistence indicator for preceding  2 d
Monthly seasonal effect, sine component             -0.059700   0.0227
with Persistence indicator for preceding  2 d
Monthly seasonal effect, cosine component            0.057400   0.0241
with Persistence indicator for preceding  3 d
Monthly seasonal effect, sine component              0.100500   0.0238
with Persistence indicator for preceding  3 d
Monthly seasonal effect, cosine component           -0.029000   0.0105
with Ln(1 + Previous day's value)
Legendre polynomial  1 for Eastings (OS grid, k      0.110400   0.0261
with Legendre polynomial  1 for Northings (OS
Legendre polynomial  2 for Eastings (OS grid, k      0.018700   0.0307
with Legendre polynomial  2 for Northings (OS
Legendre polynomial  3 for Eastings (OS grid, k     -0.448200   0.0589
with Legendre polynomial  3 for Northings (OS
Monthly seasonal effect, cosine component           -0.093000   0.0191
with Trace indicator for Previous day's value
Monthly seasonal effect, sine component             -0.044600   0.0173
with Trace indicator for Previous day's value
Monthly seasonal effect, cosine component           -0.072900   0.0181
with Trace indicator for Value 2 days before
Monthly seasonal effect, sine component             -0.094300   0.0177
with Trace indicator for Value 2 days before

Parameters in non-linear transformations:
-----------------------------------------
Legendre polynomial  1 for Eastings (OS grid, k:
          Lower limit for polynomial rep:  370.0000 (prespecified)
          Upper limit for polynomial rep:  410.0000 (prespecified)
Legendre polynomial  1 for Northings (OS grid, :
          Lower limit for polynomial rep:  420.0000 (prespecified)
          Upper limit for polynomial rep:  480.0000 (prespecified)

Spatial dependence structure:
-----------------------------
```

```
Structure used is Beta-Binomial distribution for number of wet sites
          Shape parameter of Beta-Binomi:    0.6900


                            RESIDUAL ANALYSIS
                            =================

                |    Observed    ||     % correct
                |   Dry     Wet  || Observed  Expected
    ----------------------------------------------------
    Forecast dry |  76085   38850 ||   66.2      64.7
    Forecast wet |  37680  140832 ||   78.9      77.9
    ----------------------------------------------------
                OVERALL % CORRECT :   73.9      72.8

 Rainday frequencies vs forecasts:
 --------------------------------
 Forecast decile |  1  |  2  |  3  |  4  |  5  |  6  |  7  |  8  |  9  | 10  |
 ------------------------------------------------------------------------------
 Observed propn. |0.000|0.000|0.228|0.332|0.456|0.601|0.680|0.779|0.853|0.907|
 Expected propn. |0.000|0.000|0.266|0.348|0.445|0.550|0.654|0.749|0.853|0.938|
 Number of cases |    0|    0|32909|49189|32837|21219|31701|39521|43440|42631|
 ------------------------------------------------------------------------------

        Mean Pearson residual: -0.0108
        Standard deviation of Pearson residuals:  1.0293
        Standard error of mean Pearson residual:  0.0069


    MODEL PERFORMANCE BY MONTH
    --------------------------

           |         ||     Pearson residuals
     Month | N days  ||   Mean    Std Dev   S.E. mean
    ---------------------------------------------------
         1 |  24307  || -0.0234    1.0792     0.0233
         2 |  22217  || -0.0239    1.0456     0.0244
         3 |  24736  ||  0.0144    1.0208     0.0231
         4 |  24616  || -0.0025    1.0067     0.0232
         5 |  26152  || -0.0438    1.0124     0.0225
         6 |  24894  ||  0.0134    1.0265     0.0231
         7 |  24994  || -0.0123    1.0267     0.0230
         8 |  25012  || -0.0216    1.0548     0.0230
```

```
     9 |  24366 || -0.0135    1.0336     0.0233
    10 |  24707 || -0.0135    0.9806     0.0231
    11 |  23610 ||  0.0134    1.0481     0.0237
    12 |  23836 || -0.0157    1.0149     0.0236
-------------------------------------------------

MODEL PERFORMANCE BY SITE
-------------------------
```

Site	Name	N days	Pearson residuals Mean	Std Dev	S.E. mean
L01	Site 57427 - SCAR HO	12803	-0.0175	1.0356	0.0088
L02	Site 61246 - KETTLEW	4751	-0.0267	0.9805	0.0145
L03	Site 61562 - ARNCLIF	2837	0.0224	0.9474	0.0188
L04	Site 62060 - GRIMWIT	3970	0.0834	0.8340	0.0159
L05	Site 62254 - LOWER B	13969	0.0062	0.9881	0.0085
L06	Site 62381 - CHELKER	13801	0.0003	0.9851	0.0085
L07	Site 73420 - MALHAM	3147	0.0556	0.9059	0.0178
L09	Site 74022 - SALTERF	11002	-0.0750	1.0891	0.0095
L10	Site 74091 - EARBY S	10888	-0.0295	1.0499	0.0096
L11	Site 74118 - ELSLACK	5152	0.0822	0.9303	0.0139
L12	Site 74130 - BANK NE	3755	-0.0439	0.9537	0.0163
L13	Site 74328 - EMBSAY	4514	0.0287	0.9970	0.0149
L14	Site 74739 - SILSDEN	10320	-0.0473	1.0296	0.0098
L15	Site 74852 - WATERSH	11547	-0.0198	1.0626	0.0093
L16	Site 74921 - LOWER L	10503	-0.0111	1.0378	0.0098
L17	Site 75150 - MARLEY	9430	-0.0059	1.0177	0.0103
L18	Site 77064 - GORPLEY	10951	-0.0625	1.1036	0.0096
L19	Site 77137 - EASTWOO	13439	-0.0134	1.0694	0.0086
L20	Site 77468 - MYTHOLM	13260	-0.0108	1.0617	0.0087
L21	Site 560626 - WEIR,	4288	0.0506	0.9675	0.0153
L22	Site 571894 - STAINF	2246	-0.0717	1.0641	0.0211
L23	Site 571895 - STAINF	2681	0.0679	0.8748	0.0193
L24	Site 572147 - LONG P	5917	-0.0050	0.9256	0.0130
L25	Site 572437 - GREENB	8947	-0.0243	1.0222	0.0106
L26	Site 572468 - HORTON	5288	0.0028	0.9701	0.0138
L27	Site 572866 - RIMING	1929	0.0204	0.9555	0.0228
L28	Site 573106 - CHATBU	9877	-0.0404	1.0964	0.0101
L29	Site 573174 - BARROW	2882	-0.0386	1.0333	0.0186
L30	Site 573339 - SLAIDB	5929	-0.0244	1.0763	0.0130
L31	Site 573427 - CROASD	2903	-0.0382	1.1242	0.0186
L32	Site 574488 - BUTTOC	2634	-0.0295	1.1079	0.0195

```
   L33  Site 574719 - COLNE  |  12806 || -0.0205    1.0776     0.0088
   L34  Site 574762 - BARROW |  10379 || -0.0113    1.0597     0.0098
   L35  Site 574785 - COLDWE |   9973 ||  0.0190    1.0209     0.0100
   L36  Site 574889 - BURNLE |   6451 ||  0.0231    1.0434     0.0125
   L37  Site 575332 - CHURN  |   2325 ||  0.0135    1.0722     0.0207
   L38  Site 575384 - GREAT  |   2803 ||  0.0396    0.9986     0.0189
   L39  Site 575974 - PICKUP |  10403 || -0.0189    1.0309     0.0098
   L41  Site 582626 - AUSTWI |  12747 ||  0.0044    0.9837     0.0089
-------------------------------------------------------------------------

MODEL PERFORMANCE BY YEAR
-------------------------

      |         ||     Pearson residuals
 Year | N days  ||   Mean    Std Dev   S.E. mean
-----------------------------------------------
 1961 |   6114  || -0.1399    1.1614     0.0465
 1962 |   6559  || -0.1311    1.1276     0.0449
 1963 |   6063  || -0.0429    1.0683     0.0467
 1964 |   6365  || -0.0935    1.1181     0.0456
 1965 |   6502  || -0.0725    1.1430     0.0451
 1966 |   6537  || -0.0425    1.1379     0.0450
 1967 |   6859  || -0.0209    1.0936     0.0439
 1968 |   6889  || -0.1429    1.0798     0.0438
 1969 |   6528  || -0.0561    1.1360     0.0450
 1970 |   6924  || -0.0536    1.1276     0.0437
 1971 |   6840  || -0.1723    1.1355     0.0440
 1972 |   6976  || -0.0568    1.0224     0.0435
 1973 |   6247  || -0.0746    1.0945     0.0460
 1974 |   6512  ||  0.0221    0.9897     0.0451
 1975 |   6123  || -0.0438    1.0054     0.0465
 1976 |   6023  || -0.0347    0.9993     0.0469
 1977 |   6156  || -0.0232    1.0106     0.0464
 1978 |   5163  ||  0.0138    0.9562     0.0506
 1979 |   5082  ||  0.0602    1.0413     0.0510
 1980 |   5485  ||  0.0222    0.9907     0.0491
 1981 |   5961  ||  0.0271    0.9968     0.0471
 1982 |   6823  ||  0.0023    1.0328     0.0440
 1983 |   6836  ||  0.0406    0.9413     0.0440
 1984 |   7609  || -0.0685    1.0405     0.0417
 1985 |   8220  ||  0.0922    0.9683     0.0401
 1986 |   7835  ||  0.0664    0.9487     0.0411
 1987 |   7771  ||  0.0461    1.0143     0.0413
```

```
  1988 |   8869 ||  0.0480    0.9485     0.0386
  1989 |   8089 || -0.0588    1.0118     0.0404
  1990 |   8531 ||  0.0463    0.9427     0.0394
  1991 |   9985 || -0.0585    1.0148     0.0364
  1992 |   9490 || -0.0007    1.0323     0.0373
  1993 |   9516 ||  0.0468    0.9724     0.0373
  1994 |   9832 ||  0.0596    0.9525     0.0367
  1995 |   9430 || -0.0444    1.0205     0.0375
  1996 |  10379 ||  0.0258    0.9672     0.0357
  1997 |  10159 ||  0.0044    0.9762     0.0361
  1998 |  10246 ||  0.1037    0.9793     0.0359
  1999 |   9180 ||  0.0843    1.0043     0.0380
  2000 |   2739 || -0.0240    1.0336     0.0695
-------------------------------------------------

NOTE: Standard errors computed for overall, monthly and yearly
means above are inflated by a factor of 3.637 to compensate for
spatial dependence.
```

B.2.2 North East Lancashire gamma model fitting file

```
Final parameter estimates:

Main effect:                                        Coefficient  Std Err
------------                                        -----------  -------
Constant                                              1.268300   0.0130
Legendre polynomial  1 for Eastings (OS grid, k      -0.089500   0.0061
Legendre polynomial  1 for Northings (OS grid,        0.035600   0.0080
Legendre polynomial  2 for Eastings (OS grid, k      -0.097700   0.0090
Legendre polynomial  2 for Northings (OS grid,        0.129200   0.0084
Legendre polynomial  3 for Eastings (OS grid, k      -0.124000   0.0112
Legendre polynomial  3 for Northings (OS grid,        0.121700   0.0117
Altitude (hundreds of metres                          0.033500   0.0053
Monthly seasonal effect, cosine component            -0.105800   0.0076
Monthly seasonal effect, sine component              -0.108200   0.0065
Monthly half-year cycle, sine component               0.022100   0.0043
Monthly half-year cycle, cosine component            -0.014000   0.0042
Ln(1 + Previous day's value)                          0.153300   0.0039
Ln(1 + Value 2 days before)                           0.057900   0.0044
Ln(1 + Value 3 days before)                           0.052200   0.0038
Ln(1 + Value 4 days before)                           0.033200   0.0037
```

```
Ln(1 + Value 5 days before)                          0.014200   0.0037
Trace indicator for Previous day's value            -0.051600   0.0100
Trace indicator for Value 2 days before             -0.059500   0.0102
Trace indicator for Value 4 days before             -0.048800   0.0093
Trace indicator for Value 5 days before             -0.033700   0.0093
Trace indicator for Value 6 days before             -0.034000   0.0087
Persistence indicator for preceding  2 days          0.072900   0.0129
Persistence indicator for preceding  3 days         -0.065200   0.0129
Persistence indicator for preceding  5 days         -0.057500   0.0107

2-way interactions:                                 Coefficient  Std Err
-------------------                                 -----------  -------
Legendre polynomial  1 for Eastings (OS grid, k      -0.041800   0.0179
with Legendre polynomial  1 for Northings (OS
Legendre polynomial  2 for Eastings (OS grid, k      -0.066100   0.0211
with Legendre polynomial  2 for Northings (OS
Legendre polynomial  3 for Eastings (OS grid, k       0.336900   0.0401
with Legendre polynomial  3 for Northings (OS
Monthly seasonal effect, cosine component             0.096500   0.0048
with Ln(1 + Previous day's value)
Monthly seasonal effect, sine component               0.026700   0.0042
with Ln(1 + Previous day's value)
Monthly seasonal effect, cosine component             0.071700   0.0101
with Persistence indicator for preceding  2 d

Parameters in non-linear transformations:
-----------------------------------------
Legendre polynomial  1 for Eastings (OS grid, k:
          Lower limit for polynomial rep:  370.0000 (prespecified)
          Upper limit for polynomial rep:  410.0000 (prespecified)
Legendre polynomial  1 for Northings (OS grid, :
          Lower limit for polynomial rep:  420.0000 (prespecified)
          Upper limit for polynomial rep:  480.0000 (prespecified)

Spatial dependence structure:
-----------------------------
Structure used is Constant correlation between all pairs of sites
          Inter-site correlation        :    0.7817


                              RESIDUAL ANALYSIS
                              =================
```

```
Mean of observations:   5.392
Standard deviation of observations:   6.824

Mean error (observed - predicted):  -0.042
Root mean squared error:   6.655
Proportion of variance explained by model:  0.049

Mean Pearson residual:  -0.005
Standard deviation of Pearson residuals:   1.250
Expected std dev of Pearson residuals:   1.250

Mean Anscombe residual:  0.8490 (expected:  0.8281)
Std Dev of Anscombe residuals:  0.3750 (expected:  0.4171)


MODEL PERFORMANCE BY MONTH
--------------------------
```

		Pearson residuals		
Month	N days	Mean	Std Dev	S.E. mean
1	16751	-0.010	1.200	0.037
2	14105	-0.031	1.327	0.040
3	15444	0.001	1.213	0.039
4	13840	-0.009	1.272	0.041
5	12950	0.012	1.244	0.042
6	13161	0.014	1.300	0.042
7	12370	-0.066	1.237	0.043
8	13622	0.040	1.335	0.041
9	14062	0.010	1.266	0.040
10	15257	-0.018	1.175	0.039
11	16502	-0.037	1.163	0.037
12	16272	0.029	1.278	0.038

```
MODEL PERFORMANCE BY SITE
-------------------------
```

			Pearson residuals		
Site	Name	N days	Mean	Std Dev	S.E. mean
L01	Site 57427 - SCAR HO	8378	0.000	1.344	0.014
L02	Site 61246 - KETTLEW	2776	-0.026	1.231	0.024

```
   L03  Site 61562 - ARNCLIF |   1768 ||   0.029     1.272      0.030
   L04  Site 62060 - GRIMWIT |   2865 ||   0.105     1.492      0.023
   L05  Site 62254 - LOWER B |   9303 ||   0.015     1.315      0.013
   L06  Site 62381 - CHELKER |   8832 ||  -0.055     1.246      0.013
   L07  Site 73420 - MALHAM  |   2176 ||   0.026     1.152      0.027
   L09  Site 74022 - SALTERF |   6075 ||   0.069     1.302      0.016
   L10  Site 74091 - EARBY S |   6065 ||  -0.009     1.245      0.016
   L11  Site 74118 - ELSLACK |   3285 ||  -0.068     1.202      0.022
   L12  Site 74130 - BANK NE |   1821 ||  -0.079     1.159      0.029
   L13  Site 74328 - EMBSAY  |   2786 ||  -0.065     1.219      0.024
   L14  Site 74739 - SILSDEN |   5800 ||  -0.025     1.227      0.016
   L15  Site 74852 - WATERSH |   7464 ||  -0.012     1.195      0.014
   L16  Site 74921 - LOWER L |   6407 ||   0.055     1.309      0.016
   L17  Site 75150 - MARLEY  |   4600 ||   0.003     1.317      0.018
   L18  Site 77064 - GORPLEY |   6486 ||   0.026     1.254      0.016
   L19  Site 77137 - EASTWOO |   7549 ||   0.004     1.195      0.014
   L20  Site 77468 - MYTHOLM |   7324 ||  -0.013     1.218      0.015
   L21  Site 560626 - WEIR,  |   2813 ||  -0.057     1.145      0.024
   L22  Site 571894 - STAINF |   1344 ||  -0.102     1.013      0.034
   L23  Site 571895 - STAINF |   1721 ||  -0.036     1.133      0.030
   L24  Site 572147 - LONG P |   3389 ||  -0.004     1.305      0.021
   L25  Site 572437 - GREENB |   4969 ||  -0.003     1.241      0.018
   L26  Site 572468 - HORTON |   2855 ||   0.056     1.284      0.023
   L27  Site 572866 - RIMING |   1002 ||  -0.044     1.227      0.039
   L28  Site 573106 - CHATBU |   5652 ||   0.085     1.280      0.017
   L29  Site 573174 - BARROW |   1434 ||  -0.012     1.219      0.033
   L30  Site 573339 - SLAIDB |   3785 ||   0.005     1.501      0.020
   L31  Site 573427 - CROASD |   1845 ||   0.024     1.284      0.029
   L32  Site 574488 - BUTTOC |   1627 ||  -0.028     1.249      0.031
   L33  Site 574719 - COLNE  |   7109 ||   0.011     1.236      0.015
   L34  Site 574762 - BARROW |   5750 ||  -0.011     1.212      0.016
   L35  Site 574785 - COLDWE |   6182 ||  -0.080     1.112      0.016
   L36  Site 574889 - BURNLE |   3459 ||  -0.056     1.074      0.021
   L37  Site 575332 - CHURN  |   1372 ||  -0.160     1.076      0.034
   L38  Site 575384 - GREAT  |   1847 ||  -0.113     1.119      0.029
   L39  Site 575974 - PICKUP |   6468 ||   0.002     1.222      0.016
   L41  Site 582626 - AUSTWI |   7953 ||  -0.010     1.259      0.014
-------------------------------------------------------------------------------

MODEL PERFORMANCE BY YEAR
-------------------------

       |        ||     Pearson residuals
```

Year	N days	Mean	Std Dev	S.E. mean
1961	3329	0.123	1.386	0.083
1962	3459	0.068	1.387	0.082
1963	3556	-0.035	1.111	0.080
1964	3446	-0.032	1.274	0.082
1965	3804	0.072	1.199	0.078
1966	4070	0.103	1.364	0.075
1967	4348	0.081	1.299	0.073
1968	3644	0.248	1.575	0.079
1969	3796	-0.049	1.196	0.078
1970	4157	0.013	1.207	0.074
1971	3276	0.067	1.404	0.084
1972	4054	0.008	1.210	0.075
1973	3400	-0.008	1.287	0.082
1974	4007	-0.010	1.054	0.076
1975	3277	-0.073	1.235	0.084
1976	3357	-0.095	1.130	0.083
1977	3595	0.007	1.147	0.080
1978	2974	-0.084	1.117	0.088
1979	3254	0.012	1.275	0.084
1980	3358	0.055	1.292	0.083
1981	3638	0.039	1.452	0.079
1982	4017	0.002	1.324	0.076
1983	4225	0.044	1.317	0.074
1984	4176	0.026	1.298	0.074
1985	5337	-0.101	1.220	0.066
1986	5055	0.001	1.371	0.067
1987	4859	-0.054	1.131	0.069
1988	5669	0.034	1.263	0.064
1989	4261	-0.026	1.244	0.073
1990	5220	-0.031	1.100	0.066
1991	5355	0.001	1.317	0.066
1992	5728	-0.066	1.108	0.063
1993	5868	-0.051	1.222	0.063
1994	6465	-0.007	1.092	0.060
1995	5064	-0.165	1.073	0.067
1996	6162	-0.117	1.189	0.061
1997	6048	-0.059	1.123	0.062
1998	7177	0.002	1.224	0.057
1999	6114	0.013	1.425	0.061
2000	1737	0.104	1.445	0.115

```
NOTE: Standard errors computed for overall, monthly and yearly
means above are inflated by a factor of 3.837 to compensate for
spatial dependence.
```